**********
**********

## Proceedings of

# National Journal of Parallel and Soft Computing

## Volume 03, Issue 02


**Organized by**

**University of Computer Studies, Yangon**

**Ministry of Science and Technology, Myanmar**

**December, 2022**

# EDITORIAL BOARD

# REVIEWER BOARD

# National Journal of Parallel and Soft Computing Volume 03, Issue 02

## December,2022

## CONTENTS

**********

## Distributed and Parallel Computing

## Image and Signal Processing

## Natural Language and Speech Processing

## Networking and Security

# Distributed and Parallel Computing

# An Improved Red-Black Ant Colony System Algorithm based on Traffic for Supermarket Distribution Center Route Planning

Htet Htet Win, Tin Zar Thaw
*University of Computer Studies, Yangon*
*htetlayucsy@gmail.com, tinzarthaw@ucsy.edu.mm*

## Abstract

*With the rapid growth of retailing during the modernization of Yangon, there is an increasing demand to improve the service quality of supermarkets. The supermarket shuttle service can have a direct impact on extending supermarket access, increasing shared transports, and improving customers' satisfaction. The routing software that helps supermarket delivery drivers to reach the customer's location or branch on time by providing the shortest route is the best one. Using machine learning methods not only provide efficient routes but also help business owners save time and fuel. The proposed system presented Route Planning mechanism using the improved Red-Black Ant Colony System Algorithm based on traffic information to apply supermarket delivery business.*

*This system is implemented using ASP.Net programming language on Microsoft Visual Studio 2013 IDE with Microsoft SQL Server 2008 R2 Database Engine*

**Keywords:** Red-Black Ant Colony Algorithm, Shortest Route, Route Planning Mechanism

## 1. Introduction

The chain supermarket has become a major part of Yangon's retail industry, and the optimization of chain supermarkets' distribution route is an important issue that needs to be considered for the distribution center, because for a chain supermarket it affects the logistics cost and the competition in the market directly. Yangon is the largest city in Myanmar and the industrial and commercial city of the country. The central area of the city contains the commercial district of banks, universities, trading corporations, and offices, as well as shops, brokerage houses, and bazaars. In Yangon, many million trips are using road transport every day and appearing traffic in rush hour. The optimization of chain supermarkets' distribution route is extremely depending on the traffic.

There are many kinds of algorithms for the optimization design of distribution routes. These algorithms are the ant colony system algorithm, the red-black ant colony system algorithm, the simulated annealing algorithm, the taboo search algorithm, the neural network algorithm, and so on. In this system, the red-black ant colony system algorithm is used to find the shortest route from the citymart distribution center: Ayer-Won branch to the citymart supermarket branches in Yangon based on traffic information. Then, the presented system executes the potential ways of arriving at all objective area.

This paper is organized as follows: Section 2 discusses related work and Section 3 explains background theory. Supermarket and traffic information is discussed in Section 4 and Implementation of the proposed system and experimental result are described in Section 5 and Section 6. Finally, Section 7 presents conclusion.

## 2. Related Work

The author presented the conveyance administration issue by A* Shortest Path Algorithm as [1]. Providers need to find the way plan to where their items are being conveyed, particularly concerning the geographic place of the articles, its environmental factors and the most limited way of the objective spot from current spot. This framework takes care of the providers dealing with issue in finding their most brief way by utilizing A* Shortest Path Algorithm.

The author showed the framework that is executed a system giving the transport data and course arranging administration to take care of the transport transportation issues in [2]. An electronic transport registry administration and

course arranging utilizing Dijkstra's briefest way calculation is created. This framework gives transport data, prevent data from the transport lines and Yangon transport catalog and show the most limited ways for the source and objective.

The authors proposed Red-Black Ant Colony System that is to create improved results for the arrangement of bigger mobile sales rep issues for [3]. This approach would be an extremely encouraging one in light of its consensus and due to its viability in finding generally excellent arrangements effectively to different fields of troublesome issues. The fields of interest might be the troublesome combinatorial issues, for example, load adjusting issue in broadcast communications organizations, numerous fuel financial burden dispatch issue, double requirement fulfillment issue, arbitrary number age, information mining, work booking issue, design acknowledgment and significantly more.

The aim of [9] was to tackle the issue of uneven quest for the briefest distance however disregarding the vacationer experience during the time spent the travel industry course arranging, a superior subterranean insect settlement enhancement calculation is proposed for the travel industry course arranging. Relevant data of grand spots essentially impact individuals' decision of the travel industry objective, so the pheromone update system is joined with the context-oriented data like climate and solace level of the beautiful spot during the time spent looking through the worldwide ideal course, so the pheromone update keeps an eye on the way appropriate for vacationers. The trial results show that the upgraded the travel industry course has incredibly further developed the vacationer experience, the course distance is abbreviated by 20.5% and the assembly speed is expanded by 21.2% contrasted and the fundamental calculation, which demonstrates that the better calculation is prominently viable.

The objective of the author was to dissect the ongoing dispersion circumstance of chain general stores both at home and abroad and concentrating on the quantum-enlivened developmental calculation (QEA) [6]. This author set up the numerical model of chain general stores' conveyance course and tackled the streamlined circulation course all through QEA. The specialists take Hongqi Chain Supermarket in Chengdu as an illustration to play out the trial and contrast QEA and the hereditary calculation (GA) in the fields of the combination, the ideal arrangement, the hunt capacity, etc. The investigation results show that the conveyance course streamlined by QEA acts better compared to that by GA, and QEA has more grounded worldwide quest capacity for both a limited scale chain grocery store and an enormous scope chain general store. Besides, the achievement pace of QEA in looking through courses is higher than that of GA.

## 3. Background Theory

Traveling Salesman problem (TSP) consists of finding the shortest route in complete weighted graph G with n nodes and n (n-1) edges, so all other nodes in this tour are visited exactly once. The most famous pragmatic use of Traveling Salesman Problem (TSP) are: normal appropriation of products or assets, finding the briefest way of the client want course, arranging transport lines and so on [5,7].

The most popular practical application of TSP are: regular distribution of goods or resources, finding the shortest path of the user desire route, planning bus lines etc., but also in the areas that have nothing to do with travel routes. In ant colony system (ACS), a number of artificial ants are utilized which are at first positioned arbitrarily in the urban communities [8]. Every ant fabricates a visit, that is to say, a practical answer for the TSP [4]. Be that as it may, when the component of the issue builds, the time has come consuming to create the outcome and the outcome isn't frequently ideal. In this way, a change was expected for tackling enormous voyaging issues. The reason for the change or improvement is to tackle the issue in a brief time frame and simultaneously, the outcome is required to have been streamlined. This issue roused for another calculation by further developing the ant colony system.

### 3.1. Red-Black Ant Colony Algorithm

Red-Black Ant Colony System (RB-ACS) is the improved ACS and it utilizes the fundamental idea of ant colony system, it has a few significant changes [3]. The progressions that are made to the ACS are as per the following:

**Separate nearby ways:** In ACS, just a single gathering of ants is utilized to look and the ants might utilize the way of different ants. Thus, rather than utilizing one gathering, RB-ACS utilizes two gatherings of ants groups, specifically the black gathering and the other is red gathering.

**Different boundary values:** In RB-ACS, the two gatherings have separate qualities. In this way, RB-ACS involves separate qualities for the boundaries in neighborhood refreshing.

**Worldwide refreshing:** When every one of the ants has made their visits, then worldwide refreshing is applied. In ACS, unquestionably the best ant is permitted to store pheromone on its way. In the RB-ACS, two best ants from each gathering are permitted to store pheromone.

**Benefits of RB-ACS:** It intrinsic parallelism for two gatherings. Positive Feedback represents fast revelation of good arrangements. Efficient for Traveling Problem and comparative issues can be utilized in powerful applications.

## 4. Supermarket and Traffic Information

This system managed the stocks distribution of Ayer -Won Distribution center by using Red-Black Ant colony system algorithm. The 34 citymart supermarket branches information and their traffic information are collected from Google Maps [11]. Table 1 shows the sample collected information for the proposed Supermarket Distribution Center Route Planning system.

**Table 1. The sample information of the System**

| Source | Destination | Path | Weekday | | Weekend | |
|---|---|---|---|---|---|---|
| | | | Traffic | Non-Traffic | Traffic | Non-Traffic |
| Marketplace by Citymart (6.5 mile) | Citymart Supermarket North-Okkalpa | Pyay Rd. →Kabar Aye Pagoda Rd. →Swae Taw MyRd. →Thudhamma Rd. | 18 min (9.9 km) | 15 min (9.9 km) | 16 min (9.9 km) | 14 min (9.9 km) |
| | | Pyay Rd →Oakkala Rd. →Radio Station Rd. →Thudhamma Rd. | 20 min (9.8 km) | 18 min (9.8 km) | 17 min (9.8 km) | 16 min (9.8 km) |
| | | Pyay Rd. → Kabar Aye Pagoda Rd. → Gan Da Mar Rd. →Thudhamma Rd. | 22 min (12 km) | 20 min (12 km) | 21min (12 km) | 19min (12 km) |

Traffic in Yangon can be a real problem. The rush hour times are being gathered as the information of the system from trip advisor websites [10], the commercial district of banks, universities, trading corporations, and offices. The rush hour is being assumed from 6:30-9:30 am and in the evening 15:30-18:30 pm.

***Calculating Weekdays and Traffic***

```
IF (Day is Saturday or Sunday)
    Then Weekdays=off;
Else Weekdays=on;
IF (Start time is between 6:30-9:30 am and
between15:30-18:30)
    Then Traffic=on;
Else Traffic=off;
```

***The proposed rules based on traffic information***

```
If (Weekdays==on and Traffic== on) then
    Takes delivery times based on traffic for Weekday
If (Weekdays==on and Traffic== off) then
    Takes delivery times based on non_traffic for
    Weekday
If (Weekdays==off and Traffic== on) then
    Takes delivery times based on traffic for Weekend
If (Weekdays==off and Traffic== off) then
    Takes delivery times based on non_traffic for
    Weekend
```

**Figure 1. The proposed traffic rules**

The delivering time with traffic are gathering from Google Maps that has a feature called "Popular times" that shows how long one can expect to wait at places like restaurants and supermarkets for weekdays and weekends. By knowing how busy somewhere is, and how long it takes than expected, the user can save time by planning to go on another day or at a different time. Figure 1 represents the traffic rules for Yangon, Myanmar and these rules are added the RB-ACS Algorithm.

## 5. Implementation of the Proposed System

Two groups of artificial ants are used to cooperate with other ants. Exchange information via pheromone deposited on graph edges. Pheromone is a chemical which is deposited on the ground by ants while walking. Ants prefer the paths where more pheromone is deposited.

Figure 2 represents the improved Red-Black Ant Colony System Algorithm and we add the calculation of traffic and choose the paths with the proposed traffic time rules in the step 2. So, the improved Red-Black Ant Colony System Algorithm presents Route Planning mechanism based on traffic information to apply supermarket delivery business.

**Step 1**: Place the ants of each group (Red group and Black group) randomly in each supermarket
**Step 2**: For each supermarket do
For each ant of the two groups do
•Calculate traffic
•Choose the paths with the proposed traffic rules
•Choose the next supermarket according to the state transition rule
If (every supermarket is visited) then
Place the ants in their starting supermarkets respectively
Update the visited edges using local updating rule.
**Step 3**: For each group of ants do
•Compute the best paths of each group
Update the paths
**Step 4**: If (End_condition = True) then
Print shortest path
Else Go to Step 2

**Figure 2. The improved Red-Black Ant Colony System Algorithm**

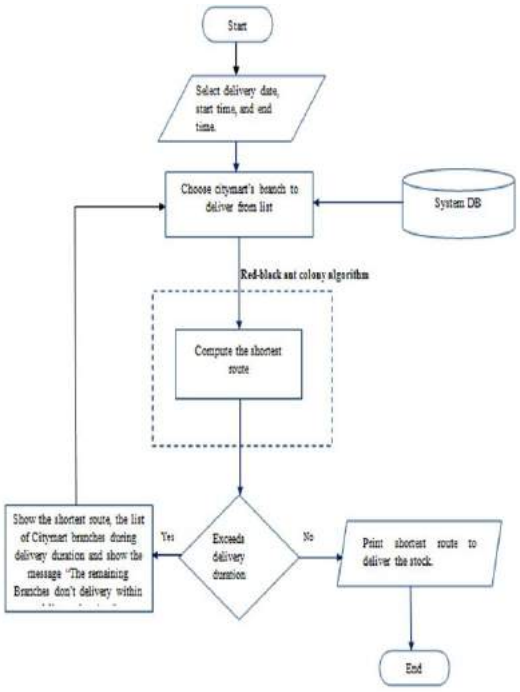



**Figure 3. The System Flow**

The Figure 3 shows the system flow of the proposed system. The system displays the Delivery Date, Start time, End time, and the List of Citymart Branches which retrieve from the system database. The user imports the above data to the system for computing the shortest delivery route. Then, the system will process the user desired data by depending on the start time and end time. After executing the process, the system will show the shortest delivery route of the system. The presented system will show the most limited course for all client characterized places. Assuming that the complete cross time surpasses the client characterized trip term, the system will print the message to the client "It isn't sufficient opportunity to send in one day".

Then, the system displays the list of supermarkets to send the stock in one day. If the duration is enough, the system will also display the shortest route of user desired citymart branch to deliver the stock. This system presents delivery date and the list of citymart Supermarket branches in the Yangon Region.

## 6. Experimental Results of the System

The proposed algorithm is compared with the existing ACS algorithm based on the randomly selected 35 paths from the distribution center.

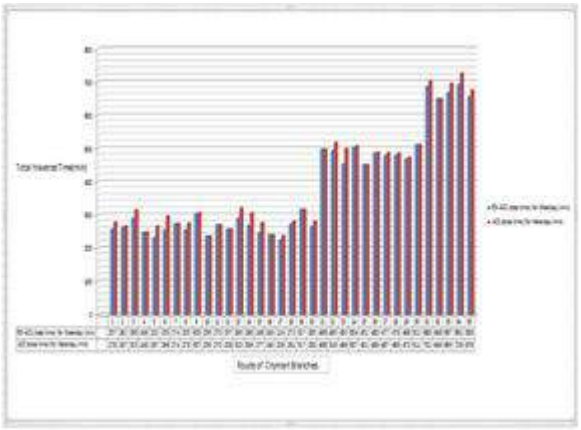


**Figure 4. Comparison of ACS and RB-ACS Based on Total Traverse Time**

Figure 4 shows the results of two algorithms based on total traverse time. When the total traverse time of these two algorithms are compared, the result of the proposed RB-ACS algorithm is more optimize than the result of the ACS to save time because of considering traffic information. The results of the proposed algorithm are more effective than ACS. Although some results of these two algorithms are same, most results are optimized than the ACS algorithm.

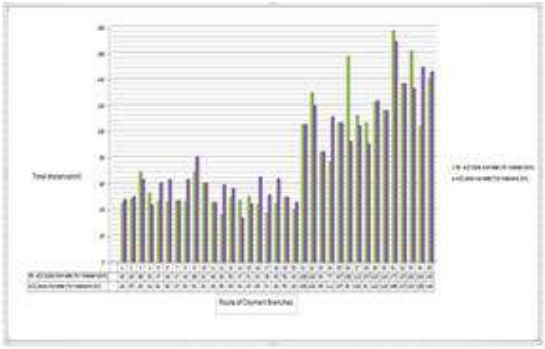


**Figure 5. Comparison of ACS and RB-ACS Based on Total Distance (km)**

Figure 5 shows the comparison results of the proposed system and the ACS algorithm based on total distance (km). According to the comparison results, 20 out of 35 routs of the proposed algorithm are same or less distance than the ACS algorithm. This is because the traverse time does not direct ratio to the distance. The system is calculated the shortest paths based on traffic information. So, the shortest route effects on the shortest time directly. Although the total distance is longer, the traverse time can be shorter. In spite of the total distance is shorter, the traverse time can be longer. So, the results can change based on the traffic. Generally, the result distance of ACS is longer than the proposed algorithm.

## 7. Conclusion

The improved Red-Black Ant Colony System Algorithm based on traffic is used to find the best route of Supermarket Distribution Center to save time. The proposed system is compared with the existing ACS based on the randomly selected 35 routes. According to the above experiments, the proposed system provides a significant improvement for obtaining the shortest path's traverse time for the user-desired routes based on traffic during the user-desired start and end time by comparing the Ant Colony System. The proposed system is simple and effective in finding very good solutions efficiently to various fields of difficult problems based on traffic.

## References

[1] Aye Lai Lai Soe, "Improving Delivery Service Applying Shortest Path Algorithm For Large Road Network", University of Computer Studies, Yangon, M.C.Sc 2010.

[2] Kyaw Zayar Oo, "Web-Based Bus Directory Service and Route Planning Using Dijkstra's Shortest Path Algorithm", University of Computer Studies, Yangon, M.C.Sc, 2009.

[3] Md. Rakib Hassan, Md. Kamrul Hasan and M.M.A Hashem, "An Improved ACS Algorithm for the Solutions of Larger TSP Problems", Department of Computer Science & Engineering Khulna University of Engineering & Technology, Khulna-9203, Bangladesh, 2009.

[4] Ivan Brezina Jr. Zuzana Čičková, "Solving the Travelling Salesman Problem Using the Ant Colony Optimization", Management Information Systems, 2011.

[5] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy-Kan and D.B. Shmoys, "The Travelling Salesman Problem," New York: Wiley, 1985.

[6] Handawi, "A Study on the Optimization of Chain Supermarkets' Distribution Route Based on the Quantum-Inspired Evolutionary Algorithm", Mathematical Problems in Engineering Volume 2017, Article ID 7964545, 11 pages.

[7] J.L. Bentley, "Fast algorithms for geometric traveling salesman problems," ORSA Journal on Computing, Vol. 4, pp. 387–411, 1992.

[8] M. Dorigo and L.M. Gambardella, "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem," IEEE Transactions on Evolutionary Computation, Vol.1, No.1, pp. 53-66,1997.

[9] S. Liang, T. Jiao, W. Du and S. Qu, "An improved ant colony optimization algorithm based on context for tourism route planning", September 16, 2021, doi: 10.1371/journal.pone.02573 17. eCollection 2021.

[10] https://www.tripadvisor.com/ShowTopic-g294191-i9454-k10208867 What_are_the_peak_times_one_should_avoid_going_downtown Yangon_Rangoon_Yangon_Region.html

[11] https://www.cnbc.com/2018/08/04/google-maps-popular-times-shows-you-wait-times.html

# Data Cache Consistency by Broadcasting with Catalog Binding in E-Commerce System

Myo Myo San, Zaw Tun
*University of Computer Studies, Yangon*
*myomyosan@gmail.com,zawtun78@gmail.com*

## Abstract

*Caching is a simple and effective way to provide faster access to information on the Internet, and it has been used effectively to improve the E- commerce servers. However, there is a growing concern by today's Web service providers to ensure that the data cached at a remote client is up to date with the current value in the server. A unique aspect of cache consistency for E-commerce applications is that most of the applications on the Web can tolerate some degree of inconsistency between the clients and the server. This system presents Broadcasting with Catalog Binding (BWC) Cache Consistency which exploits the tolerance exhibited by E-commerce applications, and which is suitable for E-commerce transactions on the Internet. This system is implemented using ASP.Net programming language with Microsoft SQL Server Database 2017 Express.*

**Keywords**: Caching, E-commerce, cache consistency, BWC

## 1. Introduction

Recent growth in Web-based commerce has prompted progressively dynamic examination in the plan of elite execution electronic trade Web locales. Today there is a developing interest in giving storing backing to E-trade applications on the Web - as reserving is a straightforward and powerful method for giving quicker admittance to data on the Internet.

Reserving alludes to putting away Web objects got to by clients at places through which the client's solicitation passes. Consequently, assuming the solicitation is for an item that is put away in the reserve, then the article is provided from the store as opposed to the server. A focal issue of reserving is store consistency, which alludes to the component by which the server guarantees that the information stored at a far off client is in the know regarding the ongoing worth in the server.

The principal objective is to foster a reserve consistency calculation for Web-based electronic trade (E-business) applications, and to assess for Web responsibilities. Specifically, this framework center around fostering a casual store consistency plot for E-business kinds of uses, as a significant number of the Web applications can endure some level of irregularity.

## 2. Related Work

The possibility of ADCC started with the memory coherency convention utilized in SGI Origin multiprocessor frameworks [5]. ADCC is recognized from Origin memory coherency convention in a few significant viewpoints. ADCC is programming based and utilizes a two-level catalog, while the convention in Origin frameworks is equipment based and utilizes a solitary level registry in particular. Another significant contrast is that the coherency convention in multiprocessor frameworks keeps a predictable perspective on memory for each processor on every memory activity. ADCC has a coarser granularity of atomicity, which requires a grouping of tasks to be executed overall. In this manner, ADCC should deal with stops and cuts short at the exchange (a succession of tasks) level.

ADCC, with the calculations proposed for information transporting DBMS during the last ten years. The calculations can be grouped into two classifications as per their approach for invalid access counteraction: evasion based and discovery based [4]. The calculations in general, with the exception of ADCC, totally depend on a unified server for simultaneousness control. CBL is broadly acknowledged as the main calculation because of its great presentation and low cut short rate [3]. As a general rule, it has preferable

execution over Caching Two-Phase Locking (C2PL) [7], No-Wait Locking (NWL) [7], Cache Locks [8] and Notify Locks [8]. Hopeful Two-Phase Locking (O2PL) [2] and Adaptive Optimistic Concurrency Control (AOCC) [1] have comparable or higher throughput. Be that as it may, the significant downside of these two hopeful methodologies is the conceded consistency check, which prompts high cut short rates. The cut short rate is a basic issue for clients in the profoundly intuitive conditions that are normal for page servers. Offbeat Avoidance-based Cache Consistency (AACC) [6] can bring down the cut short rate while keeping up with high throughput. In any case, both AOCC and AACC were proposed for versatile locking, which switches locking between the page and the article level.

## 3. Background Theory

**Concurrency** means that various clients approach the data set simultaneously. The undertaking of a simultaneousness control system is to guarantee the consistency of the information base while permitting a bunch of exchanges to simultaneously execute [7].

**Consistency** means that every client sees a consistence perspective on the information, including noticeable changes made by the client's own exchanges and exchange of different clients.

### 3.1. Consistency Problems

Consistency problems causes by concurrent processing include-

- Lost or buried Updates
- Inconsistent Analysis (Non repeatable Read)
- Uncommitted Dependency (Dirty Read)
- Phantom Reads

#### 3.1.1. Lost or buried Updates

This issue happens when at least two exchanges are perused and update on similar information thing at the offer data set. Every exchange knows nothing about different exchanges.

In the event that a subsequent exchange read a thing for update after the main exchange has understood it, yet before the principal exchange has committed. Whichever of the exchange commit first, that update will be lost [3].

#### 3.1.2. Inconsistent Analysis (Non repeatable Read)

A transaction, on the off chance that it peruses similar information thing at least a few times, ought to constantly peruse a similar worth [2, 3].

Non repeatable read emerges when a subsequent exchange gets to similar information thing a few times and peruses various information each time in light of the fact that the another exchange has been refreshed this thing while the subsequent exchange is perusing. Conflicting investigation includes various read (at least two) of a similar thing and each time the data is changed by another exchange; consequently, this term is non repeatable read.

#### 3.1.3. Uncommitted Dependency (Dirty Read)

A transaction, assuming that it recover or refresh an information thing that has been update by one more exchange however not yet dedicated by that other exchange. Filthy read resembles to conflicting examination, the thing read by the one exchange was committed by the other exchange that rolled out the improvement [4].

#### 3.1.4. Phantom Reads

A transaction re-executes an inquiry, tracking down a bunch of information not equivalent to a past one-albeit the hunt condition is unaltered. Ghost peruses may causes when inset or erase activity is performed against a column that has a place with the scope of lines being by an exchange.

### 3.2. Cache Consistency

In the client-server system, the information duplicate from the server (Global data set) is put away in the client's data set (neighborhood/reserve data set) - storing.

When the client's adjustment exchange is effectively dedicated in nearby data set (Local Cache) and worldwide data set, information in the data set will be conflicting with the information on the opposite side of clients, on the grounds

that other client doesn't have the foggiest idea about the alteration.

To stay away from the circumstance, after the adjustment is effectively dedicated, the changed data should be criticism to one more client to keep the information consistency (Cache Consistency) [5].

### 3.3. Update Propagation Techniques

There are five possible strategies for information spread. The primary qualification among the various calculations depends on whether the server monitors the information stored by clients, which is known as the *data binding information [8]*.

***On Demand Strategy (ODM)*:** This arrangement alludes to the clients mentioning the information from the server on an on request premise. The server needs to does no accounting to monitor the client store status. Whenever there is a solicitation, the client gives the server the limiting data. The server utilizes the limiting data to channel the bits of the server logs that should be shipped off the clients.

***Broadcasting with No Catalog Binding (BNC)*:** This system utilizes broadcasting methods. The server pushes updates or information alterations to all clients upon the commit of an update exchange. This strategy keeps up with no limiting data in regards to the client's reserve status and consequently doesn't monitor which pages are stored at which client; it pushes the updates to all clients in the framework, whether or not a client has stored the page or not. This framework isn't versatile considering countless clients. The benefit of this strategy is that the server keeps away from a portion of the above, for example, look-into log tasks and the calculation expected to decide the objective of updates. Endless supply of an update from the server, the client verifies whether the update influences its nearby activity; provided that this is true, it cuts off.

***Broadcasting with Catalog Binding (BWC)*:** In this plan, the server monitors the situation with client stores. Upon an update, the server settles on the clients that should be told, and spreads refreshes in view of the limiting data. This strategy decreases the quantity of updates to be proliferated at the expense of keeping up with restricting data for every one of the clients.

***Periodic broadcasting with catalog binding (PWC)*** and ***Periodic broadcasting with no catalog binding (PNC)*** integrate the possibility of intermittent update broadcasting. In these strategies, the server gathers the progressions not seen by the client at some customary time frame and starts the engendering of the updates to the client. In the event that the server keeps up with accounting data about the client's reserve status, the server sends just a part of the updates to the client (PWC). Then again, on the off chance that the server keeps up with no limiting data, then, at that point, every one of the updates are propagated to all the clients (PNC).

## 4. The Proposed System

This system proposed to develop a data consistent online shopping system.

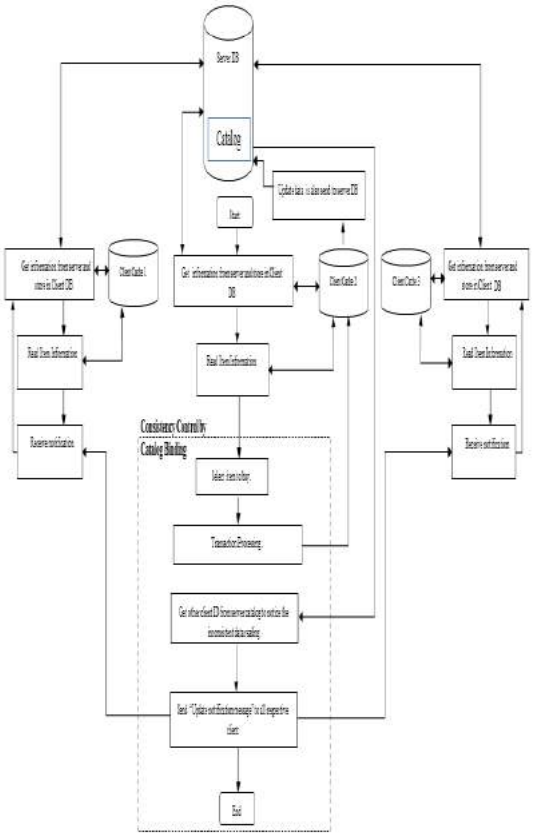



**Figure 1. The System Overflow**

In this system, the online customer receives a catalog of all the items s/he may need and adds the item s/he is interested in to a shopping cart.

When s/he wants to check out, the client submits an order for all the items placed in the shopping cart. However, not all orders are accepted by the server; there are cases in which a requested item has been sold out by the time the customer checks out. Moreover, the customer usually spends a considerable amount of time making their decision to purchase an item and, hence, most of the transactions are lengthy. The System flow is shown in figure 1.

Since it is of some concern to the customer transaction, there is a need for the server to send notifications regarding the availability of the item, and any price changes, to the client before s/he checks out.

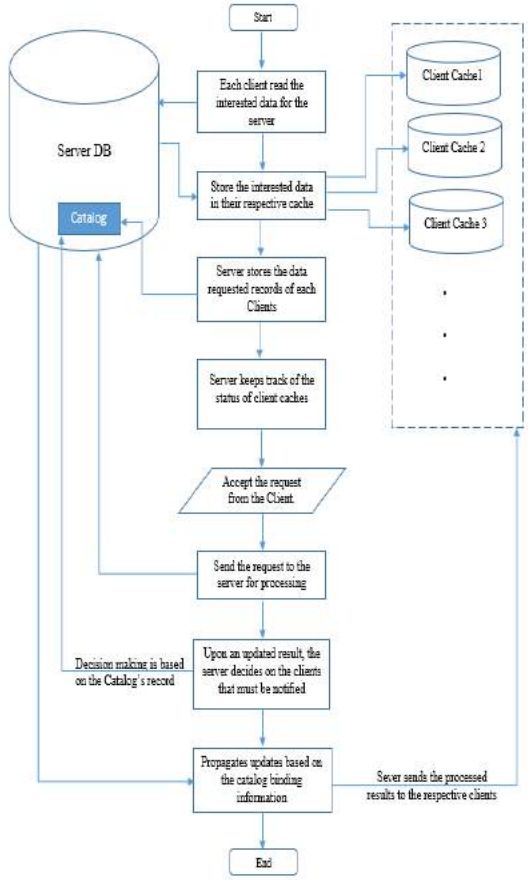



**Figure 2. Data Update Propagation Flow**

This proposed system decides on the clients that must be notified, and propagates updates based on the binding information by the "Broadcasting with Catalog Binding (BWC) approach". This minimizes the number of aborted transactions and helps better satisfy the demands of customers. The data broadcasting steps are detail explained in figure 2. The server can also provide hints about the depletion rate of an item from its inventory so that customers can make an immediate decision to reserve or purchase an item.

### 4.1. Broadcasting with Catalog Binding Algorithm

```
BEGIN
Let CD = Client Data Store, SD = Server Data;
At the start of the System, CD ← SD;
SD ← Number of Clients and Clients' ID;
If (type of transaction = = "Read")
{
   Check_notification( );
   If (notification-status ! = alert)
  {
   The data is in the latest and consistence data;
    Read Commit( );
   }
   Else
   {
    received  notification ( ) ;
    get_Update_from_server( );
    The data is in latest and consistence data;
    Read Commit();
   }
}
Else If (type of transaction = = "Write")
{
  Check_Concurrency( )
   If (status ! = concurrent access)
  {
  CD ← Update data in client's local database;
         Send_Update_to_Server( )
    Server_checks_catalog ()
    Server_send_notification( )
   }
 Else If (status = = the concurrent writing on the
same data item)
   {
    CD ← Update data in client's local database;
    Send_Update_to_Server( )
    Server_checks_catalog ()
    Server_send_notification( )
 }
}
END
```

The system checks the clients' timestamp from the catalog and then the earliest timestamp will be granted the write operation.

### 4.2. Notification for Consistency Control

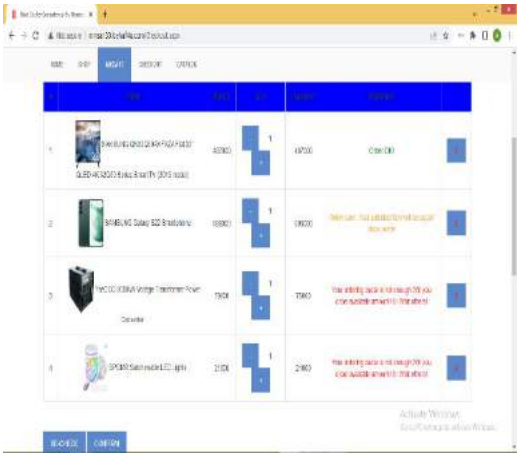


**Figure 3. MYCART Page of System (Notification for consistency control)**

"MYCART" page is support to temporarily add the user desire item (serve as customer cache) before confirming the item to buy. In this page, registered user's added item information are shown as figure 3. Although the user added item amount is shown in "QTY" field, the user can increase or decrease the desire amount of item. The "REMARK" field is the main notification section of the consistency control and concurrency control. In this field, there are one of three possible notification message. "Order OK!" message is used to show the notification that is the user selected/added item is ready to busy and there is no concurrency case. "Order Soon. Your selected item will be out of stock soon" message is to warn the user who is early selected user and the user selected item amount will be no enough soon. Because of the later user also added this item. In this situation, the early user add amount of the item and later user add amount of the same item are not totally enough to fulfill their requirements. So, the later users will also be received the notification message as "Your ordering quota is not enough. Will you order available amount. Or Wait Others!". But all of the messages are only send to the respective user. Not to send no overlap item added user.

## 5. Conclusion

Store consistency control framework can be directed the concurrent admittance to the data set and information consistency issue in multi-client climate. In this way, the obstruction among applications doesn't cause a deficiency of consistency. This proposed framework means to give the consistency control of disseminated simultaneous exchange for online business framework by utilizing information broadcasting with inventory restricting. This control works by trading advise data during the commit convention. Assuming the notice of information update is gotten, rehash the significant altered information on data set by the notice and update the information on the client side. By utilizing this methodology, a few advantages are: staying away from the stops; keeping up with information consistency and safeguarding the showing up of simultaneousness issues enjoyed lost update issue and conflicting issue. Also, correspondence cost is lower and the extra expense for the framework can be diminished. The proposed technique which is differentiated based largely on how they implement updates. The proposed E-commerce system can impact a large number of clients, client cache size, and epsilon value for Web workloads.

### 5.1. Benefit of the System

Deadlock describes a situation where two or more threads are blocked forever, waiting for each other due to a lock. At some point intensive or inappropriate exclusive locking can lead to a "deadlock" situation between two locks, where none of the locks can be released because they try to acquire resources mutually from each other. A deadlock can occur when two or more users are waiting for data locked by each other. Deadlocks prevent some threads from continuing to work.

The main advantage of this system is deadlock free, thus saving the expense that deadlock detection usually required in locking approach. Processes can be run concurrently without affecting other processes and without failing. Fetching objects at the client side and working there locally thus reduced the processing time and network latency. "Broadcasting with catalog binding", inform each client sides about the task according to the user ID, Item ID, data processed time (timestamp) and current updated data. So, all objects remain in a consistent state when they are accessed by multiple transactions.

### 5.2. Limitations and Further Extensions

This system does not contain any payment method with bank. It implements the online sale

system for approving consistency and concurrency control at the distributed database by catalog binding. Since it is the distributed database system, it depends on the server database and client database. If the server database is crashed, this system can't effort to get the original data. So, this system can be extended the data recovery service to be perfect data reliability.

## References

[1] Adya, A., Gruber, R., Liskov, B. and Maheshwari, U. "Efficient optimistic concurrency control using loosely synchronized clocks," in Proceedings of the ACM SIGMOD Conference on Management of Data. San Jose, CA, pp. 23– 34.

[2] Carey, M.J., Franklin, M.J., Livny M. and Shekita, E. J. "Data Caching Tradeoffs in Client-Server DBMS Architectures," in Proceedings of the ACM SIGMOD, pp. 357-366.

[3] Franklin, M.J. "Client Data Caching: A Foundation for High Performance Object Database Systems," Kluwer Academic Publishers, Boston, MA.

[4] Franklin, M.J., Carey, M.J. and Livny, M. "Transactional client-server cache consistency: alternatives and performance," ACM Transactions on Database Systems, vol. 22(3), pp. 315-363.

[5] Laudon, J. and Lenoski, D. "The SGI Origin: A ccNUMA highly scalable server," in Proceedings of the 24th Annual International Symposium on Computer Architecture, vol. 25(2), pp. 241-251.

[6] Ozsu, M.T., Voruganti, K. and Unrau, R. "An asynchronous avoidance-based Cache Consistency Algorithm for Client Caching DBMSs," in Proceedings of the Conference on Very Large Data Bases (VLDB). New York, NY, pp. 440-451.

[7] Wang, Y. and Rowe, L.A. "Cache consistency and concurrency control in a client/server DBMS architecture," in Proceedings of the ACM SIGMOD Conference on Management of Data. Denver, CO, pp. 367–377.

[8] Wilkinson, K. and Neiman, M.-A. "Maintaining consistency of client-cached data," in Proceedings of the Conference on Very Large Data Bases (VLDB). Brisbane, Australia, pp. 122-133.

# Peer-to-Peer File Sharing System in Campus Using Multi-Agent

Thida Win, Aye Mya Hlaing
*University of Computer Studies, Yangon*
*thidawin@ucsy.edu.mm, ayemyahlaing@ucsy.edu.mm*

## Abstract

*Peer-to-Peer (P2P) file sharing is a technique of network file sharing. Clients can direct access and share the files to other clients in P2P network. The P2P file sharing system does not require any central server. P2P is more effective than client-server method because the computers have shared responsibilities to communicate with one another. Every computer on a P2P network can operate a server as well as a client. The integration of P2P network architecture and multi-agent technology are able to be an optimal solution for the real environmental problems. Multi-agent innovation has demonstrated better performance results about the further development, adequacy and accuracy in dynamic and distributed environments. This system depends on the multi-agent architecture and presents how to manage and share the files with the concept of load balancing mechanism on the P2P local area network. The proposed system is developed by utilizing the Microsoft.Net platform and SQL Server.*

## 1. Introduction

Nowadays, Peer-to-Peer (P2P) network is uncommonly notable in file sharing. P2P network is used to download and share files like recordings, music, videos, images, digital books, games, programs, and so on. P2P frameworks are described by decentralized control, enormous scope and outrageous dynamism of their workspace. This system is implemented on P2P local area network and designed for file sharing on campus. Because of the way that there is no central server, there are no expenses charged by the server that is facilitating the application. Each peer is responsible for storing and sending the requested information.

Developing intelligent and autonomous agents is a central goal of Artificial Intelligence (AI). Agents are well known research objects in all fields. Agents derived from the AI field. Agents are extremely delicate to the environment and give a few elements like to follow up for the benefit of others, independent, intelligent, reactive, proactive, receptive, capacity, collaborate and negotiate. A Multi-Agent System is made out of independent various agents. Agents can relate, team up and arrange to deal with problems that are past the singular limits each agent.

The proposed system depends on the multi-agent framework. It is a distributed P2P file sharing system and composed of independent different multiple agents. In this system, each peer has four agents: Manager Agent, File Send Agent, Download Agent and Load Balancing Agent. File Send Agent can send multiple files to multiple peers. The system can save memory storage space and keep away from copy files due to Download Agent. Agents can reduce their workload and allocate their tasks to be available peers within the same LAN by the Load Balancing Agent.

## 2. Related Works

Ozalp Babaoglu, Hein Meling and Alberto Montresor developed “Anthill: A Framework for the Development of Agent-Based Peer-to-Peer Systems” [20]. Anthill is a system to help the plan, execution and assessment of P2P applications. It comprises of a powerful organization of friend hubs; social orders of versatile agents go through this organization, connecting with hubs and helping out different agents to tackle complex issues. It can be utilized to build various classes of P2P administrations and display strength, transformation and self-association properties. It can be also utilized for the acknowledgment of distributed applications by giving a JXTA based network foundation.

M. Grivas and S. J. Turner proposed “Agent Technology in Load Balancing for Network Applications” [15], the agent advancement in

load adjusting to arrange the network applications. This paper consolidates gathering and executions that using programming experts to accomplish load changing in the Web programs. Conveyed enrolling described the issues and one more field of assessment for load changing. Many kinds of assessment proposed different techniques to stack changing in the scattered system yet researcher uses agent's base response for load changes because of agents have new properties for handling the problem in this workspace.

Budditha Hettige, Asoka Karunananda, Kathriarachchi and Weerasinghe described "ITray, Multi-agent solution for LAN based file sharing" [7]. ITray is designed as a multi-agent system using the MaSMT, a Java-based framework. Agents communicate with each other via Java socket.

By utilizing load balancing component, agent reduces the own load and use tasks scheduling to allocate the task to available resources in the network. Utilizing these assets can be shared more effetely in the dynamic and distributed environment. The core system consists of an XML-based ontology. The MaSMT is fully tested under actual environment and it demonstrates the way that resources can share perfectly within the network. Furthermore, research can be further developed by sharing handling power within the network as per the requirements of the network.

## 3. Background Theory

There is no generally acknowledged meaning of the idea of the agent. Be that as it may, the accompanying four properties are broadly acknowledged to characterize agents: independence, reactivity, social capacity and favorable to liveliness. However, the following four properties are widely accepted to characterize agents: autonomy, reactivity, social capacity and pro-activeness.

Agents are independent computational substances (independence), which communicate with their current circumstance (reactivity) and different agents (social capacity) to accomplish their own objectives (supportive of movement). Agents are equipped for connection with different agents by correspondence, exchange and coordination. Agent is whatever can be seen as seeing its current circumstance through sensors and following up on that environment through actuators.


**Figure 1. An agent in its environment**

Figure 1. gives an abstract view of an agent. The agent takes sensory input from the environment and produces as output actions that affect it. The interaction is usually an ongoing, non-termination one.

### 3.1. The Nature of Environments

Task environment is the environment in which agent performs its task. Its detail incorporates Performance Measure, Environment, Actuators and Sensor (PEAS). In planning an agent, the initial step should constantly be to determine the undertaking environment as completely as could really be expected. Task environment shifts along a few critical aspects. They can be fully or partially observable, deterministic or stochastic, episodic or sequential, static or dynamic, discrete or continuous and single-agent or multi-agent [4].

### 3.2. Intelligent Agent (IA)

Developing intelligent and autonomous agents is a central goal of Artificial Intelligence. An intelligent agent (IA) is an independent element which act coordinating its action towards achieving goals, upon an environment utilizing perception through sensors and resulting actuators. They may moreover learn or use data to achieve their goals.

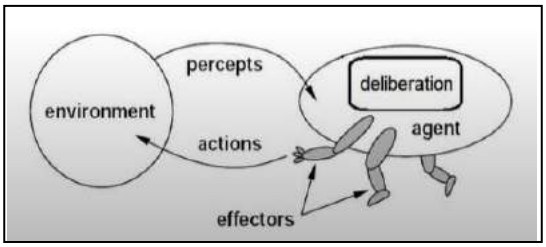



**Figure 2. An Intelligent Agent (IA)**

A simple intelligent agent (IA) is shown in Figure 2. Intelligent agents have properties such as reactivity, pro-activity and social capacity.

Reactivity: Agents can see their current circumstance, and answer in an ideal style to changes that happen in it to fulfill their plan targets.

Pro-activity: Agents can show objective coordinated conduct by stepping up to fulfill their plan targets.

Social capacity: Agents are equipped for connecting with different specialists to fulfill their plan targets.

Simple reflex agents, model-based reflex agents, goal-based agents and utility-based agents are four basic types of intelligent agents.

### 3.3. Multi-Agent System (MAS)

Multi-agent system (MAS) is an inexactly coupled organization of programming agents that collaborates to take care of issues that are past the singular limits or information on every issue solver. Multi-agent framework comprises of different specialists, which collaborate with each other by trading messages through PC network foundation.

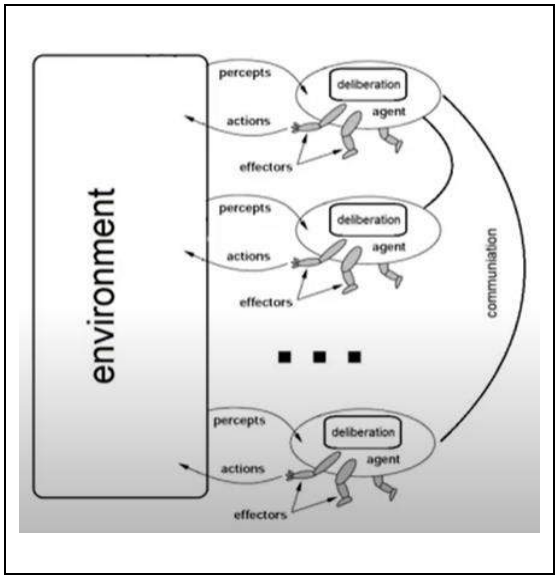



**Figure 3. A Multi-Agent System (MAS)**

A multi-agent system (MAS) is shown in Figure 3. The attributes of Multi-Agent System are that every agent has fragmented data or capacities for taking care of the issue, there is no framework worldwide control, information is decentralized and calculation is non concurrent.

### 3.4. Peer-to-Peer (P2P) File Sharing

Network file sharing is the strategy associated with duplicating data, information, documents starting with one PC, then onto the next utilizing a live network association. Peer-to-peer is a methodology for network file sharing that permits clients to straightforwardly admittance to different clients to share documents. Files can be shared directly without the need of a central server. P2P technique is more powerful than client-server strategy in light of the fact that the PCs have shared liabilities to communicate with one another. P2P file sharing system utilize no focal servers with the exception of rather permit all PCs on the network to work both as a client and a server.

## 4. Implementation of the System

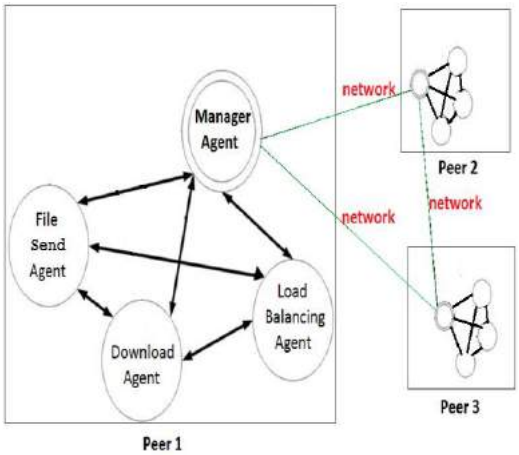



**Figure 4. Agent Architecture of the Proposed System**

Figure 4. shows the agent architecture of the proposed system. The system comprised of multiple agents: Manager agent, File Send Agent, Download Agent and Load Balancing Agent.

Manager Agent: Manager Agent can contact with other Manager Agent in different computer devices inside a similar network. Manager Agent can invoke its ordinary agents to do the client's task by passing the message. Manager Agent manages the file inventory of the which can direct the incoming process to an appropriate agent.

File Send Agent: Manager Agent summons File Send Agent by passing the directive to send records to the other peer or companions in a similar organization. That message contains the

IP address of the beneficiary peer and record area which need to ship off the other machine by the client. As indicated by this message, File Send Agent gets the record area and send the document through an organization.

Download Agent: Manager Agent invokes Download Agent by passing the message when the client enters a URL (Uniform Resource Identifier) to the framework. That message contains the URL of the file to be downloaded which is entered by the client. Download Agent checks the URL whether or not that document is existed or not in the system and downloads in the event that that record isn't existed. Subsequently, this structure can diminish memory wastage and can make an effort not to duplicate the copying record.

Load Balancing Agent: When the client enters a web URL, the framework sends that URL to any remaining peers in LAN. Other peer's Manager Agent gets the URL and ensures that the file is accessible. Assuming that file is accessible, send that file by File Send Agent to the mentioned peer. In any case, the framework keeps a queue to store URL in a first in first out based mechanism. Manager Agent invokes Download Agent through communicating something specific to download that URL.

When the download queue is become increasing, Manager Agent invokes Load Balancing Agent to get shared file and to assign download tasks with other peers. Load Balancing Agent distribute the URL to any remaining accessible peers in the network to download that URL and send back downloaded files to the mentioned peer. Other accessible peers' Manager Agent gets the URL and invokes their Load Balancing agent to really look at the responsibility of the framework.

At the point when the framework is free, that file is downloaded by Download Agent. Then Manager Agent invokes File Send Agent and sends that file to the mentioned peer which sends the URL. Subsequent to downloading the record or totally distribution undertaking to the next agent, the URL is consequently eliminated from download queue.

## 5. System Performance and Evaluation

By using the Download Agent instead of duplicate file downloading, the system can save the memory storage. The system can reduce the huge workload and allocate the download tasks to be available peers within the same LAN by using the Load Balancing agent.

As the performance of load balancing, the downloading of the text files is tested and results are shown in the below tables and figure. Downloading single file without load balancing is shown in table 4.1 and downloading multiple files is shown in table 4.2.

**Table 1. Downloading Single File without Load Balancing**

| No. | Downloaded File | Size | Download Time Without Load Balancing (in second 's') |
|---|---|---|---|
| 1 | Test1.txt | 15 KB | 3 sec |
| 2 | Test2.txt | 17 KB | 3 sec |
| 3 | Test3.txt | 19 KB | 4 sec |
| 4 | Test4.txt | 14 KB | 2 sec |
| 5 | Test5.txt | 18 KB | 3 sec |
| 6 | Test6.txt | 15 KB | 3 sec |
| 7 | Test7.txt | 16 KB | 3 sec |
| 8 | Test8.txt | 14 KB | 2 sec |
| 9 | Total | 128 KB | 23 sec |

**Table 2. Downloading Multiple Files**

| No. | Download Multiple File | Without Load Balancing | With Load Balancing |
|---|---|---|---|
| 1 | TestingFile1.txt<br>TestingFile2.txt<br>TestingFile3.txt | 31 sec | 10 sec |
| 2 | TestingFileA.txt<br>TestingFileB.txt<br>TestingFileC.txt<br>TestingFileD.txt<br>TestingFileE.txt | 35 sec | 12 sec |

## 6. Conclusion

In this paper, peer-to-peer (P2P) file sharing application has well known a hot conversation on P2P file sharing between all businesses. The

proposed system is being revolved around comfort and cost adequacy by utilizing P2P file sharing method on the LAN and capacities of multi-agent technology. This framework is executed considering P2P file sharing system and introduced file sharing component and load balancing component. Utilizing load balancing component agent diminishes the own load and use tasks, planning to designate the task to accessible assets in the network.

## References

[1] Agostino Poggi and Michele Tomaiuolo, "Integrating Peer-to-Peer and Multi-agent Technologies for the Realization of Content Sharing Applications", Italy.

[2] Alberto Grosso, "An Agent Programming Framework Based on the C# Language and the CLI", University of Genova, DIST, Italy, 2003

[3] "An Introduction to Multi-Agent Systems" by Michael Wooldridge.

[4] "Artificial Intelligence A Modern Approach Second Edition" by Stuart J.Russel and Peter Norvig

[5] "Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems" by Stefano V. Albrecht and Peter Stone

[6] Bertolini, D., Busetta, P., Nori, M., Perini, "Peer-to-peer multi-agent systems technology for knowledge management applications". An agent-oriented analysis. In: WOA 2002, Milano, Italy, pp. 1–6 (2002)

[7]Budditha Hettige, Asoka Karunananda, Kathriarachchi and Weerasinghe, "ITray: Multi-agent solution for LAN based file sharing" Article in International Journal of Computer Applications, June 2017

[8] Chongjie Zhang, Victor Lesser and Prashant Shenoy, "A multi-agent learning approach to resource sharing across computing clusters", UMass Computer Science UM-CS-2008-035,2008.

[9] "Computation Aspects of Cooperative Game Theory" by Georgios Chalkiadakis, Edith Elkind, Michael Wooldridge

[10] "Intelligent Agents: Theory and Practice" by Michael Wooldridge and Nicholas R. Jennings

# University Data Recovery System using Seed Block Algorithm

Myat Phoo Nge, Yu Wai Hlaing
*University of Computer Studies, Yangon*
*myatphoonge5@gmail.com, myatphoonge@ucsy.edu.mm*

## Abstract

*Since organizations such as universities have become very dependent on digital data processing, a breakdown may disrupt the business' regular routine and stop its operation for a certain amount of time. To prevent data loss and minimize disruptions there must be well-designed file backup and recovery procedures. The recovery process can rebuild the system when it goes down. The data recovery services are required to maintain this data very efficiently. This system uses a smart remote data backup algorithm, Seed Block Algorithm (SBA). There are two main ideas of this algorithm. First, it will help the users to collect information from any remote location in the non-presence of or loss of network connectivity. The second one is to recover the files if by mistake file gets deleted or if the primary data gets destroyed due to any reason.*

**Keywords**: system breakdown, data recovery, data backup, Seed Block Algorithm

## 1. Introduction

In information technology, a **backup and recovery**, or the process of backing up, alludes to the duplicating and chronicling of PC information so reestablishing the first after an information misfortune event might be utilized. Reinforcements have two purposes. The basic role is to recuperate information after its misfortune, be it by information erasure or defilement. Information misfortune can be a typical encounter of PC clients; a 2008 review viewed that as 66% of respondents had lost documents on their home PC. The optional reason for reinforcements is to recuperate information from a prior time, as per a client characterized information maintenance strategy, regularly designed inside a reinforcement application for how long duplicates of information are required. However, reinforcements address a basic type of calamity recuperation, and ought to be essential for any catastrophe recuperation plan; reinforcements without anyone else ought not to be viewed as a total debacle recuperation plan.

Since a reinforcement framework contains no less than one duplicate of all information considered worth saving, the information stockpiling prerequisites can be critical. Putting together this extra room and dealing with the reinforcement cycle can be a convoluted embrace. An information vault model might be utilized to give design to the capacity. These days, there are a wide range of kinds of information stockpiling gadgets that are valuable for making reinforcements. There are likewise a wide range of manners by which these gadgets can be set up to give geographic overt repetitiveness, information security, and movability.

Before information is shipped off their capacity areas, they are chosen, extricated, and controlled. A wide range of methods have been created to enhance the reinforcement system. These incorporate improvements for managing open records and live information sources as well as pressure, encryption, and de-duplication, among others. Each reinforcement plan ought to incorporate run-throughs that approve the unwavering quality of the information being upheld. It is essential to perceive the restrictions and human elements engaged with any reinforcement plot [6].

## 2. Related Work

Exchange Handling Framework (TPS) is the information lifeline for a specific business association since it is the wellspring of information for data frameworks like MIS (The board Data Framework) and DSS (Choice Emotionally supportive networks). TPS can be utilized for the different associations, where information plays the most noteworthy need. In

the improvement of Software engineering, exchange handling framework are advancing in numerous areas, for example, carrier reservation, banking, inn, the travel industry, industry, shopping center and medical services. This framework is quick and dependable enough to deal with the hourly bank exchanges [2].

Exchange The board Framework (TMS) gave admittance to their transmission framework to advance culmination in discount power market [3]. Their TMS is imagined mechanizing and incorporate the market interface and to facilitate security processes. Their framework upheld the administration of transmission and subordinate assistance is quick and dependable enough to deal with the hourly market exchanges. Exchanges are without a doubt the unit of recuperation [1]. Business associations for the most part use recuperation strategies for information security and dependability.

## 3. Background Theory

To plan and carry out a recuperation part, one should know definitively which sorts of disappointments are to be thought of, how frequently they will happen, how long is normal for recuperation, and so on. One must likewise make suspicions about the unwavering quality of the fundamental equipment and capacity media, and about conditions between various disappointment modes. [2]. In any case, the rundown of expected disappointments won't ever be finished consequently:

- For each set of failures that one can think of, there is at least one that was forgotten.
- Some failures are extremely rare. The expense of overt repetitiveness expected to adapt to them might be high to the point that it could be a reasonable plan choice to reject these disappointments from thought. Assuming one of them happens, nonetheless, the framework cannot recuperate from the circumstance naturally, and the information base will be ruined.

The main types of failures are described as follows:

**Transaction Failure**: The transaction of failure has proactively been referenced in the past area. Considering multiple factors, the exchange program doesn't arrive at its typical commit and must be reset back to its start, either at its own solicitation or for the DBMS. Inside one application, the proportion of exchanges that cut short themselves is fairly consistent, contingent just upon how much mistaken input information, the nature of consistency checking performed by the exchange program, and so forth • The proportion of exchanges being cut short by the DBMS, particularly those brought about by halts, relies generally upon the level of parallelism, the granularity of locking utilized by the DBMS, the legitimate pattern (there might be problem area information, or information that are regularly referred to by numerous simultaneous exchanges), and the level of impedance between con-current exercises (which is, thus, very application subordinate). For our order, it is adequate to say that exchange disappointments happen 10-100 times each moment, and that recuperation from these disappointments should occur inside the time expected by the exchange for its ordinary execution.

**System Failure**: The system failures that we are thinking about can be brought about by a bug in the DBMS code, a working framework issue, or an equipment disappointment. In every one of these cases handling is ended in an unconsavaged way, and we accept that the items in principal memory are lost. Since data set related optional (nonvolatile) capacity stays unaffected, we expect that recuperation occurs in the very measure of time that would have been expected for the execution of every single hindered exchange. On the off chance that one exchange is executed inside the request for 10 milliseconds to 1 second, the recuperation ought to take something like a couple of moments. A framework disappointment is expected to happen a few times each week, contingent upon the security of both the DBMS and its functional climate.

**Media Failure**: Besides these pretty much typical disappointments, we need to expect the deficiency of some or all the optional stockpiling holding the data set.

Such a circumstance must be overwhelmed by full overt repetitiveness, or at least, by a duplicate of the information base and a review trail covering what has occurred from that point forward. Attractive capacity gadgets are typically truly solid, and recuperation from a media

disappointment isn't probably going to happen more frequently than a few times per year. Contingent upon the size of an information base, the media utilized for putting away the duplicate, and the age of the duplicate, recuperation of this kind will assume the request for 60 minutes [6].

## 4. Proposed System

Since organizations are extremely reliant upon information. Exchange Handling Framework is utilized to deal with information exactness and information misfortune avoidance. There are three primary capabilities: information misfortune avoidance, information locking and stop location and goal. Since the greater part of the business associations - worried on the precision of information, it keeps the information exactness - forestall information misfortune during exchanges. Information consistency is chiefly dependent on secure exchanges to show unwavering quality to its clients. This exchange handling framework depends on seed block recuperation data and forestalls information misfortune during exchange period. During the exchanges, because of different mistakes, the exchange is intruded on and the association with different separates.

In every university, there has a lot of data about students, teachers, staff, and other school related activities. As technology advances with the times, we need many changes in everyday lives. Education is very important for all ages. Therefore, all the information about university should be backup and recovery planning. In our "University Data Recovery System", Excel, Power Point and Microsoft Word Files which are mostly used in university have been recovered in a short time using Seed Block Algorithm. When the system breaks down, it stops the regular routines of the business and stops its operation for a certain amount of time. To recover incomplete transactions, to prevent data loss, and to minimize disruption, the well-designed backup and recovery procedure is put into use. This system is using a very efficient algorithm for data backup called Seed Block Algorithm (SBA) [4]. The process flow of the proposed system is shown in figure 1.

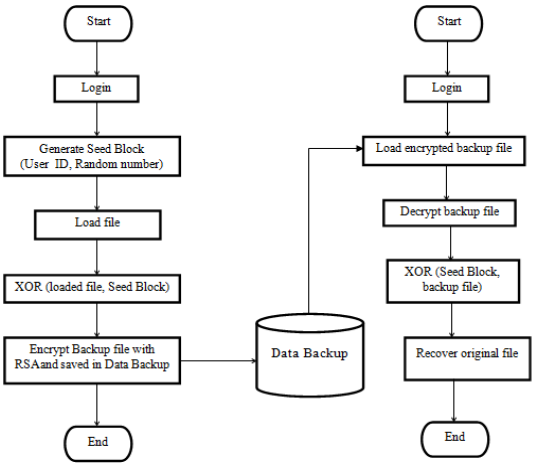



**Figure 1. The System Flow**

### 4.1. Proposed Data Recovery Technique - Seed Block Algorithm (SBA)

1. BEGIN
2. Set a random number in the main storage and unique client id for every client
3. Whenever the client id is being registered in the main storage, then client id and random number is getting EXORed( ) with each other to generate seed block for the particular client.
4. Whenever client creates the file in storage/ backup first time, it is stored at the main storage.
5. When it is stored in main storage (blob), the main file of client is being EXORed with the Seed Block of the particular client.
6. It is also encrypted using public key RSA
7. That output file is stored at the backup storage (blob) in the form of file' (pronounced as File dash).
8. During Retrieval, check if data present in main storage If present then EXOR with seed block and retrieve data If not present, retrieve data from backup storage.

9. During Retrieval from backup storage, the private key of the user will used to decrypt file'

10. The user will get the original file by EXORing on decrypted file' with the seed block of the corresponding client to produce the original file and return the resulted file in case of crash.

11. END

### 4.2. Example Operation of System

Step-1 Start

Step-2 Assume there are five clients in the system
Let client id for client1 = cid01
client id for client2 = cid02
client id for client3 = cid03
client id for client4 = cid04
client id for client5 = cid05
random number = 12345

Step-3 clientid for client2 XOR random number
cid02 ⊕ 12345
= 09910510002 ⊕ 12345
= 9910502347
(Seed block for client2)

Step-4 Client2 creates file in storage/ backup E.g., 'myat' = 6d796174)
Change text to hexadecimal with converter

Step-5 Client2's file XOR with Client2's seed block
6d796174 ⊕ 9910502347
= 997d294230 (result stores in main storage)

Step-6 Encrypt by public key RSA (Using Online RSA Encryption, Decryption And key Generator Tool)

Step-7 Encrypted file stored at backup server.

Step-8 When the wanted file is retrieved, result in Step-5 XOR with Client2's seed block.
997d294230⊕9910502347
= 6d796174
(Result show as hexadecimal, change to text again)

Step-9 If not present, we retrieve data from back up storage. Using private key with RSA calculation, Client2 will get decrypted file.

Step-10 To get original file, decrypted file XOR with Client2's seed block.

Step-11 Stop

ASCII code for c, i, d
c=099
i=105
d=100

**RSA Calculation**

Let p = 3
q = 5
r = p * q = 3*5 = 15
Encryption key, e = 11 (Public key, greater than p & q)
Decryption key, d =? (Private key)
d * e = 1 modulo (p-1) * (q-1)
d * 11 = 1 modulo 2 * 4
d * 11 = 1 modulo 8
if d = 1 => 1 * 11 = 1 modulo 8
Divide 8 into 11, and the answer is 1 with remainder 3(1≠ 3)
if d = 2 => 2 * 11 = 1 modulo 8
Divide 8 into 22, and the answer is 2 with remainder 6(1≠ 6)
if d = 3 => 3 * 11 = 1 modulo 8
Divide 8 into 33, and the answer is 4 with remainder 1(1 = 1)
Therefore, d = 3.
Let Plaintext, P = 13 (integer)
Cipher text, $C = P^e$ modulo r
$= 13^{11}$ modulo 15
=1,792,160,394,037modulo 15
= 7
$P = C^d$ modulo r
$= 7^3$ modulo 15
= 343 modulo 15
= 13

### 4.3. RSA Encryption Algorithm

[5] Rivest-Shamir-Adleman is the most used public key encryption algorithm. It can be used both for encryption and for digital signatures. The security of RSA is generally considered equivalent to factoring, although this has not been proved.

RSA computation occurs with integers modulo n = p * q, for two large secret primes p, q. To encrypt a message m, it is exponentiated

with a small public exponent e. For decryption, the recipient of the cipher text

$$c = m^e \pmod{n}$$

computes the multiplicative reverse

$$d = e^{-1} \pmod{(p-1) * (q-1)}$$

(We expect that e is chosen appropriately for it to exist) and gets cd = m e * d = m (mod n). The confidential key comprises of n, p, q, e, d (where p and q can be overlooked); the public key contains just n and e. The issue for the aggressor is that processing the opposite d of e is thought to be no more straightforward than factorizing n.
The key size ought to be more noteworthy than 1024 pieces for a healthy degree of safety. Keys of size, say, 2048 pieces ought to permit security for quite a long time.

### 4.4. Benefits of Remote Backup Services

[3] The following issues must be covered in Remote Backup services:

1. Data Integrity
2. Data security
3. Data Confidentiality
4. Genuine Characteristic
5. Cost efficiency

Data Integrity

Server's entire design along with all total states informs us concerning Information trustworthiness of server. At the hour of transmission and gathering, information which oppose to any sort of progress in it. Such type of data is verifies using Data Integrity. Validity of Data on Remote server is also checked by Data integrity.

Data security

The Remote servers have essential need to give all out security to information of client. Also, either purposefully or non-deliberately, just specific client ought to approach that information or no other users.

Data Confidentiality

In certain times, the framework needs to be maintained client's information documents to be mystery to such an extent that if no. of clients at the same time getting to the capacity/ reinforcement, when different clients getting to documents on the capacity/reinforcement ought to unfit to see specific information record that is has a place with just that specific client. This is otherwise called Information Secrecy trademark.

Genuine Characteristic

Reliability is the significant trait of the Far-off capacity/reinforcement. Remote backup server should possess trustworthiness that because every user is having their private data on server.

Cost efficiency

The cost of processing of data recovery should be proficient so that enormous number of organizations alongside clients can take advantage of back-up and recuperation administration. There are many huge quantities of techniques that have zeroed in on these issues. The front said issues happen at the hour of recuperation moreover in back-up of domain of storage/ backup computing is discussed.

## 5. Conclusion

Nowadays large amount of data is stored in the storage/ backup and becoming very important to all the organization. Today, it is an internet age and almost all organizations are based on transaction processing. So, data backup and recovery are very important for data reliability and data availability. Seed Block Algorithm (SBA) is robust in assisting the users to collect information from any remote location in the loss of network connection and if file deletion occurs due to any reason, this system can also recover files. The SBA also focuses on the security issue for the back-up files stored at remote server, without using any of the existing encryption techniques. The SBA will take minimum time for process recovery so that the issues correspond to time can be solved. The proposed system consists of the storage/ backup recovery section which will proceed the failure file request transaction to become a successfully completed transaction. So, this system is reliable for the data loss recovery during file request transaction processing because of the original database and data backup are parallel stored.

## References

[1] Amman, P., ET. al., "Recovery from Malicious Transactions", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, 2015.

[2] Ghazi Alkhatib and Ronny S. Labban, "*Transaction Management in Distributed Database Systems*: the Case of Oracle's Two-Phase Commit", Senior Lecturer of MIS, Qatar College of Technology, Doha, Qatar and Computer & Communications Engineer; Consolidated Contractors International Company Athens, Greece; Alkhatib@qu.edu.sa and r.s.labban@ieee.org.

[3] Mahantesh N. Birje, Praveen S. Challagidad, "Remote backup and recovery review: concepts, technology, challenges and security", International Journal of Cloud Computing, InderScience Publishers, vol. 6, issue 1, 2017.

[4] Ruchira. H. Titare, Prof. Pravin Kulurkar, "Remote Data Back-up and Privacy Preserving Data Distribution: A Review", International Journal of Computer Science and Mobile Applications, Vol. 2, Issue. 11, November 2014

[5] Shireen Nisha, Mohammed Farik, "RSA Public Key Cryptography Algorithm – A Review", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 6, ISSUE 07, JULY 2017

[6] Tripathy, S. and B. Panda, "Post-Intrusion Recovery Using Data Dependency Approach", Proceedings of the 2021 IEEE Workshop on Information Assurance and Security, pp. 156-160.

# Cache Coherency Control by Using MOESI Protocol in Online Doctor Appointment

Thura Zaw, Si Si Mar Win
*University of Computer Studies, Yangon*
*medcthurazaw@gmail.com, sisimarwin@ucsy.edu.mm*

## Abstract

*The rapidly growth of internet and the World Wide Web have made the possible for millions of users to gain access to geographically distributed web content. However, raising the population of online user and the unbalanced use of content accesses, popular objects may cause server and network overload and increase the latency for content access significantly. Caching is a common technique to reduce content access latencies. In caching system, data items are retrieved from the server machines, cached and processed at the client machines, and then write back to the server. This system controls the cache consistency by using MOESI protocol in online doctor appointment. This appointment system allows patients to get an appointment with doctor from any time at anywhere using mobile devices and internet services. It can control the concurrent processing and saves the waiting time not only for the users but also makes the appointment process efficient.*

**Keywords**: cache consistency, MOESI, caching

## 1. Introduction

Throughout the 20th century to the present day, the way that view and use of technologies have changed dramatically. Since a highly technological age, where every part of our daily lives is related to technology in some way, technology plays a role in creating resources that provide us with information at our fingertips. Continuous advancements in technology have given rise to many new methods of electronic communication, such as websites, email, voicemail, and video conferencing. These advanced communication technology tools help remove time and distance barriers to effective communication. Also, information technology has made significant contributions in the medical industry. With the increased use of electronic health record, telehealth services, and mobile technologies like smartphones and tablets, physicians and patients are both seeing the benefits that these new medical technologies are bringing. Not only do patients have access to quick and accurate medical information using their devices, but they can also use applications to book or keep track of doctor's appointments, get real time reminders and much more. Health clinics are getting a boost in their everyday job by reducing time spent in the filing, record maintenance, and other routine tasks. Doctors can communicate directly with their patients and achieve greater procedural efficiency. Most of the practices are now are offering patients the option of making appointments online. Different options of online appointment technology are available on the market, and practices are able to choose the system which suits the practice's requirements.

Appointment booking systems are an essential feature in patient scheduling. At the same time, many patients cancel the appointment just few hours before the appointment time, causing a gap in the doctor's schedule. The idea of this study is, to argue for and plan a system, which introduces a new and creative idea of what the current market requires. It is also easy to use and gives patients the possibility to directly book an appointment with just one click.

## 2. Related Work

Data centric model using home-based lazy release consistency technique for developing distributed application was presented by [1]. In their system, consistency control and vector timestamp synchronization were implemented using train ticket sales system in Myanmar railway transportation as case study. For data modification, their system used vector timestamps to define the user who can modify the

data. Client centric concurrency control system was presented in [2]. Their work applied monotonic write consistency control and vector clock time synchronization algorithm. To ensure consistency control for the monotonic-write, write operations were propagated in to the all clients at several sites with the correct order.

The simulation of various concurrency control methods were performed by Sonal Kanungo and Morena Rustom [3]. They compared the simulation results of 1. Lock-Based Protocols, 2. Timestamp-Based Protocols, 3. Validation – Based Protocols and 4. Multiversion Schemes based on number of transactions, committed transactions and rollback transactions. They suggested according to their simulations, Locking Protocols are good for update intensive applications but they are not free from deadlocks and causes the locking overhead. Although Time stamp protocols provide better concurrency than locking protocols, they suffer the numerous rollbacks and may also causes the storage overheads for keeping timestamps. Optimistic protocols may abort more transactions if not prevented in frequent-update systems.

## 3. Background Theory

Distributed Concurrency control is the activities of managing concurrent access to a database in a distributed system. It is the main task of transaction processing in the distributed database systems. Database transactions in the distributed systems should be coordinated so that the result remains the same as if they are executed sequentially. It is called concurrency control in database system.

There are various concurrency control mechanisms for managing concurrent access in a distributed system. In the distributed database system, concurrency control is used to resolve conflict with the concurrent access or update of data [4].

### 3.1. Consistency Problems

Consistency problems causes by concurrent processing include-

- Lost or buried Updates
- Inconsistent Analysis (Non repeatable Read)
- Uncommitted Dependency (Dirty Read)
- Phantom Reads

#### 3.1.1. Lost or buried Updates

This problem occurs when you peruse two or more exchanges and update them with similar information from the quote dataset. All exchanges know nothing about other exchanges. Subsequent exchanges assume that the main exchange understands and then reads what to update before the main exchange commits. Updates are lost whichever changes is committed first [5].

#### 3.1.2. Inconsistent Analysis (Non repeatable Read)

If the exchange reads similar information at least once or twice, it should continue to read similar values. Non-repeatable reads occur each time a different message is perused, because subsequent exchanges make an information item similar multiple times, while another exchange refreshes the item, and subsequent exchanges peruse it. Conflicting exams consist of re-reading a large number of similar terms over and over (at least twice).

#### 3.1.3. Uncommitted Dependency (Dirty Read)

A transaction, on the off chance that it recover or refresh the information that has been refreshed by another exchange but has not been dedicated by another exchange. Messy read resembles to conflict checking, where what one exchange reads is committed by another exchange that rolled out improvements.

#### 3.1.4. Phantom Reads

A transaction reruns the query and tracks a set of information that isn't equivalent to the past, but the hunt conditions are unchanged. Ghost browsing can occur when an insert or erase activity is performed on a column that has a range of rows placed by an exchange [6].

### 3.2. Cache Consistency

In a client-server system, a copy of the data from the server (global database) is stored in the client's database (local/cache database) (cache).

When a client's update transaction is successfully committed in both the local database (local cache) and the global database, the data in the database is inconsistent with the data on the other side of the user because other users are unaware of the changes.

To avoid this, after a change has been successfully committed, the changed information should be fed back to another user to keep the data consistent (cache-coherency) [7].

## 4. Overview of the Proposed System

In a caching system, the client fetches objects from the server machine, operates on them locally, and sends the changes back to the server. This architecture leverages the processing power of client machines to improve system performance. It reduces the loads on the server by performing as many computations as possible on the client computer.

Consistency control is an important part of running transactions concurrently on a database. Therefore, client-side caching introduces inconsistencies into the system. Some problems arise when two clients read the same file at the same time and modify it at the same time. One is that when the third process reads the file from the server, it gets the original version, not one of the two newer versions. The problem might be that the effects of file changes are not globally visible. Another problem: when two files are written back to the server, the last one written overwrites the other. When a cache entry (file or block) changes, the new value is kept in the cache, but also sent to the server. As a result, when another process reads the file, it will get the latest value [8].

**States of MOESI:** In computing, MOESI is a fully cache coherency protocol that encompasses all of the possible states commonly used in other protocols.

- Modified(M)
- Owned (O)
- Exclusive(E)
- Shared(S)
- Invalid(I)

**Modified (M):** Cache line only exist in the current cache and is *dirty*. It has been modified from the value in main memory (M-state). The cache must write the data back to main memory sometime in the future before allowing the main memory state to be read again (no longer valid). Write back changes the row to shared state (S).

**Owned (O):** The cache is one of several that has a valid copy of the cache line but has exclusive rights to modify the cache line. Other caches can read cache lines, but not write them. When this cache modifies data in a cache line, it must broadcast those changes to all other caches that share the line. The introduction of owned state allows dirty sharing of data. This means that modified cache blocks can be moved between different caches without updating main memory. After invalidating all shared copies, the cache line can be changed to a modified state, or it can be changed to a shared state by writing the changes back to main memory. Owned cache lines must respond to snooping requests with data.

**Exclusive (E):** The cache line is only in the current cache, but it's *clean* and matches the main memory. It can change to shared state at any time in response to a read request. Alternatively, it can be changed to the modified state on write.

**Shared(S):** Indicates that this cache line may be stored in other caches on the machine and is clean. It also matches the main memory. The line can be discarded (changed to an invalid state) at any time.

**Invalid (I):** Indicates that this cache line is invalid (unused) [9].

### 4.1. Implementation of System

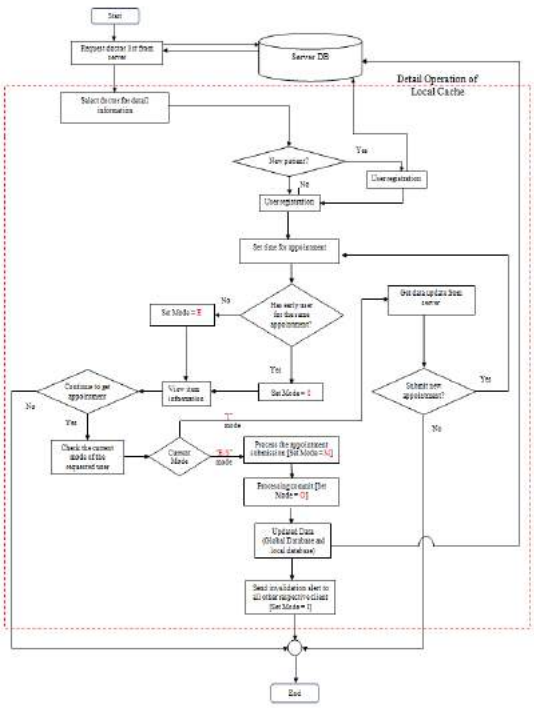



**Figure 1. Overview of the Proposed System**

At the start of the system, the client requests a doctor information from server and then that is stored in the client's cache. When patient reads that information, the system checks the read doctor information is owned by other clients. If it has more than one patients reads, the system set the shared mode in its client's directory. If it has only one reader patient, the system set the excc mode in its client's directory. So, the client knows that the doctor information for appointment is concurrently reading or not.

When the patient requests the write (booking) grant to the doctor information related document, the system checks the requested document is owned by other clients. If it has more than one owner, the system set the "S" mode in its client's directory. Then the system checks the late owners and broadcast these patients not to get write grant (booking). The early write request patient gets the write grant (appointment) and change the "M" mode in its client's directory. After commit the appointment (write transaction), the client's committed data update is also update in the server database. Then the server checks the clients list in its directory for the consistency data replication to other client's database.

If the document has only one owner patient, the system set the "M" mode in its patient's directory. Then the committed appointment update is stored in server but no need to replicate other client patients because the document is not in share mode. So, the system can control the client cache consistency and server consistency by the use of the two tier directory (Client directory and Server directory). All of the patients can know which patient can own which document and which patient granted which process mode by the user of patient (client) directory and server directory. This is the property of MOESI protocol and the architecture of the proposed system is depicted in Figure 1.

As the proposed system, the multiple patients can read the doctor list and their information from the server database. They can book the doctors based on their specialties and the date/time available. For the simultaneously booking processes of multiple clients, this system applies MOESI protocol. The sample of the implemented forms are shown as figure 2 and figure 3.

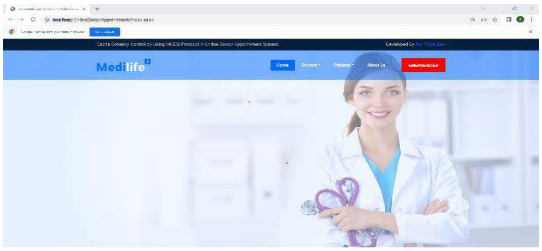



**Figure 2. Online Doctor Appointment System' Home Page**


**Figure 3. Booking Doctor Form**

## 4.2. Consistency Controlling

For the same page, some clients and server's directories are not consistent. Therefore, these clients' directories may have stale existence flags set. This directories mismatch will only cause problems when those clients want to refresh the page. When the server receives a speculative update request for a shared page, the server compares its directory with the client's directory. When the server detects that the client's directory is stale, it allows guessing, but at the same time informs the client of the difference. If there are some new shared clients, speculative clients are not allowed to commit until the new shared clients have invalidated their copies.

On commit, the client sends a log to the server, changes the status of the updated page from busy to modified, and then activates other client requests to wait for that client to commit. When the server receives a commit message, it deletes the entry in the CLT associated with the committing transaction and modifies the updated page's busy/exclusive status to exclusive (using the busy client as the commit client), moves the logs on to a persistent storage space and activates transactions of other clients that are waiting for commit [10].

### 4.3. Sequence Diagram of System

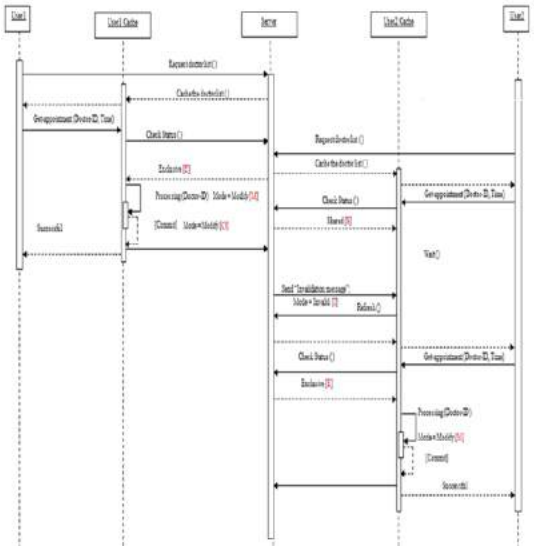


**Figure 4. Sequence Diagram**

Appointment management systems are used by all the primary healthcare clinics and also hospitals for managing access to service providers. Appointment managers can have a huge impact on the success of the organization, therefore they deserve proper attention. Ensuring patients are scheduled effectively and efficiently is critical to maintaining and maximizing clinics revenue. While excellent clinical care remains the main patient expectation, health care consumers are now also seeking supporting systems like appointment managers that fulfill their needs. The consistency controlling process of proposed doctor appointment system is briefly explained in Figure 4.

## 5. Conclusion

The proposed system can control the data consistency by the use of MOESI protocol; because of the user can notify the data inconsistency by the “Invalidation” phase. By keeping clients awareness of the global state of cached data, clients can actively participate in maintaining cache coherency. Client-side directory consistency is designed to reduce the message overhead of maintaining client-side directory consistency. This system provides data sharing capabilities – on doctor appointment time schedule - can collaborate with each other effectively. From the application perspective, this system will handle booking management and booking issues based on the user's selection/request. Patients can choose a doctor based on their medical needs and can view details and reviews to learn more. Therefore, it provides an efficient solution that allows users to view the different booking slots available and select their preferred date and time.

## 6. Limitations and Future Work

In this system, system clock is used for the timestamps such as Read, Write for each transactions. To decide the write permission of concurrent clients’ processes, the timestamp of commit transaction is only considered. Although the quantitative performance of MOESI over MESI would be better, this system implements the online doctor appointment process with MOESI protocol only. Therefore, the comparisons results of MOESI with other concurrency protocols cannot be described in this paper. To analyze the performance results of various concurrency protocols, the proposed system will be implemented by using other protocols based on multiple factors in future.

## References

[1] Z.Z. Moe, and T.T. Naing, Implementation of Home-based lazy release consistency system for a distributed application, Parallel and Soft Computing (PSC), University of Computer Studies, Yangon, Myanmar, 2010.

[2] A.N. Mon, T.T. and Naing, Implementation of Client-Centric Consistency Control and Synchronization System for Distributed Replication (For Mobile Clients), Parallel and Soft Computing (PSC), University of Computer Studies, Yangon, Myanmar, 2010.

[3] S. Kanungo and D.M. Rustom, Analysis and comparison of concurrency control techniques. International Journal of Advanced Research in Computer and Communication Engineering, 4(3), 2015.

[4] P.A. Bernstein, and E. Newcomer, Principles of transaction processing. Morgan Kaufmann, 2009.

[5] https://www.geeksforgeeks.org/cache-coherence-protocols-in-multiprocessor-system/, August, 2022.

[6] Isolation (database systems) - Wikipedia, August 2022.

[7] M.J. Franklin, M.J. Carey, and M. Livny, “Transactional client-server cache consistency: alternatives and performance,” ACM Transactions on Database Systems, vol. 22(3), pp. 315-363.

[8] A. Adya, R. Gruber, , B. Liskov, and U. Maheshwari, “Efficient optimistic concurrency control using loosely synchronized clocks,” in Proceedings of the ACM SIGMOD Conference on Management of Data. San Jose, CA, pp. 23– 34.

[9] S. Dey, and M.S. Nair, Design and implementation of a simple cache simulator in Java to investigate MESI and MOESI coherency protocols. International Journal of Computer Applications, 87(11), 2014.

# Concurrency Control System on Transactional Replication

Saint Saint San, Kyi Kyi Win
*University of Computer Studies, Yangon*
*saintsaintsan1925@gmail.com, kyikyiwin@ucsy.edu.mm*

## Abstract

*Nowadays, there are many competitors in many other organizations and companies by using innovative technology. So, the commerce is larger and larger and it affects to open new branch offices in different locations. Therefore, it is important to handle the transaction (Write/Update) of each branch by the head office. There are two databases: original database and replica / (Key/value store) database. When the original database handles read-just responsibilities from similar applications over the information recreated from the first data set while the original database handles read/write transactional application workloads. The principal necessity is guaranteeing the utilization of the reports on the copy data set in precisely the same request they were executed in the first data set, which is called execution-defined request. The essential server executes the activities and sends duplicates of the refreshed information to the copies. This framework presents a novel concurrency control algorithm to solve the concurrency problem in the hotel reservation system. This system is implemented using C# programming language with Microsoft SQL server database engine.*

**Keywords:** replica database, primary server, novel concurrency control

## 1. Introduction

Increasingly more associations are utilizing various information base as opposed to attempting to fit one data set to all information the executive's needs. The explanation is to lessen responsibility on single expert information base. The conditional updates in the first data set due to compose exchange, are transported to the key-esteem store and applied in a similar request to ensure the right state for the copy, which is called execution-defined request. Exchanges might interleave during their execution against the first data set. The replication ought to ensure that the subsequent serialization request for exchanges in the reproduction is the very same as the serialization request in the first data set and no other serialization request is satisfactory.

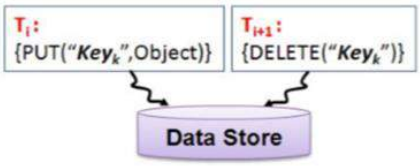



**Figure 1: Serialization Order on System State**

If $T_i$ is executed before $T_{i+1}$ then the data item (Key $_k$) will not exist in the data store. On the other hand, if $T_{i+1}$ is executed before $T_i$ the data item will exist in the data store [figure1]. Subsequently, the subsequent execution isn't right according to serialization perspective, the subsequent execution isn't satisfactory since it doesn't bring about the right state considering the predefined execution request.

## 2. Related Work

The majority of conveyed simultaneousness control calculations fall into one of three basic categories: locking computations, timestamp computations, and hopeful or affirmation computations all fall into this category. Timestamp requesting partners expects that conflicting information obtained through exchanges will be acted upon in timestamp request. Timestamp requesting partners timestamps with all recently gotten to information things. By exchanging certificate data during the commit convention, the appropriated affirmation functions. The contention exchange is restarted when appropriate authentication is used to handle messages and complete their work. This framework presents simultaneousness control for Ticket reservation framework which demonstrates that disseminated confirmation

calculation is timestamp-based, hopeful simultaneousness control calculation approach. Circulated data sets are carried out at the business community and principal office is the focal information base for all ticket deals. The Framework generally check the really focus data set contents and the deals community's information base items to have the framework information consistency. Thus, the ticket deal framework can stay away from the messy read and the lost and covered update [1].

[2] A large portion of the organizations disperse their data per branch, guaranteeing that the information is primarily refreshed locally, by the branch where they were made. Instances of such organizations could be hypermarkets and banks. Thus, it appears to be prudent to recreate all information objects, free of the organization branch where they have been made. In the event that information is reproduced, "remote" gets to can be achieved locally, further developing access times as well as accessibility, in the event of hub disappointments in the organization). Notwithstanding, each time an information object is changed, refreshes must be multicast to a given number of data set reproductions. Information replication among various destinations is seen as a method for expanding application execution and its information accessibility. This framework proposes a simultaneousness control and recuperation in a middleware engineering called COPLA (Normal Item Developer Library Access). This engineering gives industrious article state replication. This framework depends on time-stamp, it is a transformation of the Hopeful convention to this engineering. The recuperation cycle of this framework permits applications to proceed (or begin) executing exchanges at all hubs, even in the hub being recuperated.

[3] This framework is executed as dispersed framework and consequently, there might be more than one client and clashes might be happened. Clashes are constrained by Wait Die calculation in light of time-stamp. After exchanges are committed or restarted, thing status is shown once more. At the point when a client changes a read mode to the compose mode, the framework will check the ongoing mentioned thing has simultaneous access or not and has legitimate sum. Assuming the mentioned thing has simultaneous access, the framework checks the essential timestamp for the Stand by Bite the dust simultaneousness control. If the mentioned thing's fundamental timestamp is sooner than the ongoing handling client's essential timestamp, the framework awards the solicitation to stand by [Although the mentioned exchange has early timestamp yet hang tight for non-precautionary restriction]. While possibly not in this way, the mentioned exchange is kicked the bucket (for example Try not to get an opportunity to stand by due to the Stand by Kick the bucket's halt preventive plan.) After the mentioned exchange conceded to get to, the exchange will handle the thing and multicast all information update to all members.

## 3. Background Theory

The most common method of Database replication is to creating and maintain the multiple instances of a similar data set. In different areas without duplicating the entire set, it is divided information between data sets. While additional data set servers manage slave copies. One data server manages the entire copy of the data set in many data set replication operations. A single data set's at least two duplicates remain synchronized. The duplicate of the initial data set is referred to as a replica, and the first data set is referred to as an Expert data set.

**Concurrency** means that that multiple clients approach the data set at the same time. A simultaneousness control instrument's mission is to permit a number of exchanges to take place simultaneously while simultaneously ensuring the consistency of the data base.

**Consistency** means that every client has the same perspective on the data, including the obvious changes brought about by the client's own interactions and those of other clients.

### 3.1. Consistency Problems

Consistency problems causes by concurrent processing include-

- Lost or buried Updates
- Inconsistent Analysis (Non-repeatable Read)
- Uncommitted Dependency (Dirty Read)
- Phantom Reads

#### 3.1.1. Lost or buried Updates

This issue happens when at least two exchanges are perused and update on similar information thing at the offer data set. Every exchange knows nothing about different exchanges.

On the off chance that a subsequent exchange read a thing for update after the main exchange has understood it, yet before the principal exchange has committed. Whichever of the exchange commit first, that update will be lost.

#### 3.1.2. Inconsistent Analysis (Non-repeatable Read)

A transaction, on the off chance that it peruses similar information thing at least a couple of times, ought to continuously peruse a similar worth.

Non-repeatable read emerges when a subsequent exchange gets to similar information thing a few times and peruses various information each time in light of the fact that another exchange has been refreshed this thing while the subsequent exchange is perusing. Conflicting investigation includes various read (at least two) of a similar thing and each time the data is changed by another exchange; consequently, this term is non-repeatable perused.

#### 3.1.3. Uncommitted Dependency (Dirty Read)

A transaction, on the off chance that it recovers or refresh an information thing that has been update by one more exchange yet not yet dedicated by that other exchange. Filthy read resembles to conflicting examination, the thing read by the one exchange was committed by the other exchange that rolled out the improvement.

#### 3.1.4. Phantom Reads

A transaction re-executes a query, tracking down a bunch of information not equivalent to a past one-albeit the pursuit condition is unaltered. Ghost peruses may causes when inset or erase activity is performed against a column that has a place with the scope of lines being by an exchange.

## 4. The System Overview

The proposed system replicates the data change of original database in the key/value stores and prevents all read transactions from hitting the original database. All update transactions are sent to the corresponding master. The master database has a set of slaves that are its replicas, serve the read-only transactions in the system. Updates are disseminated (propagate) from master to its slave nodes by eagerly upon their arrival several updates and applying them together as shown in figure2.

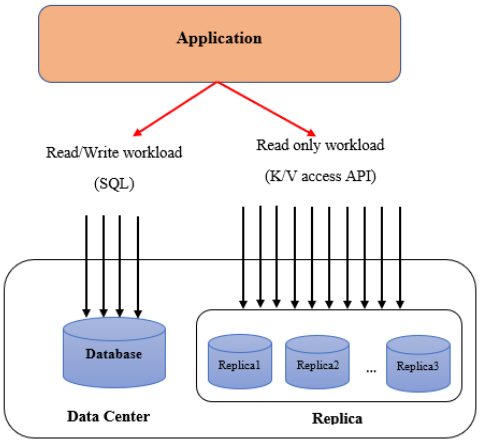



**Figure 2: System Overview**

### 4.1. The Proposed System Architecture

Relational Data Over Key-value Store: In the relational data in RDBMS into a key-value store, the data layout on these two stores are different, need to provide a mapping scheme to map the relational data layout into the key-value data layout.

| RoomID | RoomType | TotalRoom | Price |
|---|---|---|---|
| 1 | Standard Room | 20 | 40000 |
| 2 | Superior Room | 10 | 60000 |
| 3 | Single Room | 20 | 35000 |

**Figure 3: Tuples in Room Table**

Above Figure-3 is a sample structure of relational table in the proposed system. In this proposed system, the data layout of relational

data in RDBMS are different in key/value store data structure.

| Key | Value |
|---|---|
| Room_1 | {Standard Room,20,40000} |
| Room_2 | {Superior Room,10,60000} |
| Room_3 | {Single Room,20,5000} |

**Figure 4: Key/value objects for the tuples in ROOM table**

A standard relational database is the database that stores all the persistent application data and responsible for handling read/write transactional workload. The database system for certain application workloads, the database is replicated into a distributed key-value store. The replicated key-value store plays similar role to cache for the database, is used to handle the read only workload while the read/write work- load is run directly on the original database as shown in figure 4 and figure 5.

Between the relational database and the key-value store, main component of system (Replication Middleware) that are responsible for synchronizing the key-value store with the relational database.

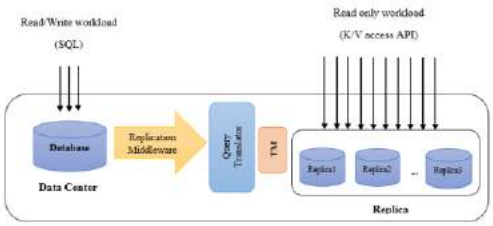



**Figure 5: Architecture of Transactional Replication Using Key/Value Store**

**Query Translator (QT):** component is responsible for translating the update only SQL statements into key/value store that can be directly executed on the key-value store. The replication workload only contains the write operations to key-value store.

**The Transaction Manager (TM):** component is used to apply the transactions to the key-value store concurrently. The TM component essentially implements the proposed concurrent replication method. When the transactions reach the TM (transaction Manager) they are in the form of key-value store as they have been translated by QT component.

**Replication Middleware:** The Replication Middleware component is responsible for shipping the transactions (Write/Update) from the relational database to the replica in the key-value store. It periodically reads the transaction log in the database; the new updated transactions are ships to the key-value store. The transactions only include write statements and there is no need to apply read statements from the relational database in the replica.

The data in the key-value store is the replication of the data in the original database. To maintain the replicated data in the key-value store synchronized with the original data in the relational database, the system uses the replication middleware. When a transaction processing is committed, the system's middleware must send committed data update to all replicas. Then, the middleware checks whether all of the replicas have committed data updated or not. If all of the replicas receive the data update, the middleware send the committed transaction to the completed transaction list. If not, the middleware will restart the replication to replicate the data update to all replicas.

## 4.2. Concurrency Control for Replication

Novel concurrency control mechanism for replicated system is different from the ordinary concurrency control systems (not replicated system) because of the execution- defined order of transactions. In an ordinary concurrency control algorithm, when a set of transactions are executed, the result of the execution is directly acceptable because of such system no need to propagate the execution result and the system can immediately accept for the new request.

To control concurrency in replicated system with ordinary control, the system must adopt the other replication mechanisms such as active or passive. However, in the propose concurrency control algorithm has to guarantee the consistency between original and key/value store. The result of committed transactions must be exactly the same result of the system replicas' (key/value store) data. So, the system halt to execute the new request until the earlier

execution result is completely updated on key/value stores.

The priority queue, which is referred to as the "CommitReqQ' in the algorithm, is responsible for keeping the order of transactions based on ascending order of their sequence numbers.

Committed transaction list: A committed transaction is the one that does not have any conflict with its predecessors. However, the updates in its buffer has not been applied to the key-value store. Such committed transactions are stored in a list which is called "Committed transaction list".

Completed transaction list: A completed transaction is a committed transaction that the updates in its buffer have been applied to the key-value store. Such complete transaction is stored in a list which is called "Completed transaction list". But the completed transaction list has a limited amount of data store. So, the completed transaction list must be periodically cleaned to store the new completed transaction for future.

### 4.3. The Novel Concurrency Control Algorithm of the Proposed System

CRQ = Commit Request Queue
CTL = Committed Transaction List (But not finish multicasting to all replicas)
CPTL = Completed Transaction List (Finished the task of multicasting to All replicas)

$T_i$ = user requested transaction variable
$T_j$ belong to the transactions in CTL
BEGIN
$T_i$ = T1 (First transaction for the CRQ);
Remove T1 from CRQ;
If (Ti is conflict with Tj) //Tj is committed but have not been multicast to al replicas
{
Ti is added to the restart list;
}
Else
{
Ti is added to the CTL;
After the transaction Ti is completed and its effect is applied to the Key-Value store. (Replicas)
}
Check the restart list;
If (Restart is not empty)
{
Restart the transaction to be committed transaction;
}
Add completed transaction to CPTL;
END

The algorithm also uses two lists:

*The committed transaction list* holds the transactions that are in COMMITTED state and have committed successfully.

*The completed transaction list* contains the transactions that are in COMPLETED state which are the committed transactions that have also been applied to the key-value store. The concurrency control algorithm starts by checking the first transaction in the CommitReqPQ. If this transaction's sequence number is not the expected sequence number the algorithm does nothing and waits until the transaction with the expected sequence number is put into the CommitReqPQ.

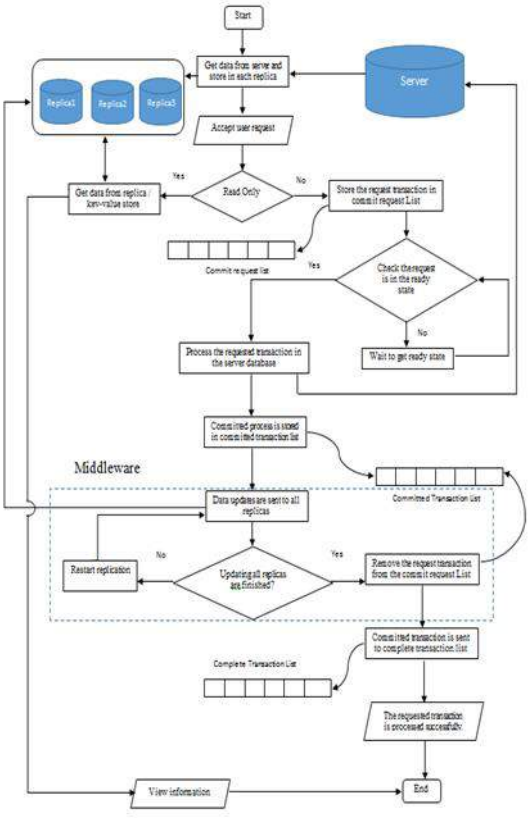



**Figure 6: System Flow Diagram**

If the transaction in the head of queue has the expected sequence number it is removed from the queue and is examined for conflict. When expected transaction is on top of the CommitReqPQ it means that all the preceding transactions have been evaluated by the algorithm and are in COMMITTED or COMPLETED state. The system flow is shown in figure 6.

In Figure 6, this is system flow diagram for proposed system and this system is used to solve the concurrency problem in the hotel reservation system. According to the above descriptions, this diagram shows how to control the concurrency to synchronized in replicas from the original database.

### 4.4. Discarding Completed Transactions

In propose concurrency control algorithm the Completed Transaction List is the last list that stores transactions. However, by processing more and more transactions this list will grow larger. Therefore, the system needs to limit the size if this list and remove the completed transactions from the list if there is no need for them.

A completed transaction Ti is stored in the Completed Transaction List. If another transaction Tj starts before the completion of Ti, and Tj has conflict with Ti, then there is a possibility that Tj did not use the updated data resulted from Ti. Thus, in order to make sure that Tj observes the results of Ti, we need to make sure that Tj starts after completion of Ti. Based on this assumption, if there is no active transaction that has started before completion of a transaction Ti and then completed Ti can safely be removed from the Completed Transaction List.

## 5. Conclusion

In the development of applications for organizations to open new branch offices in different locations is suitable to use this kind of system. This system presented an architecture based on fully replication of relational database on key-value store system where the key-value store is used for read-only transactions. In relational database to the key-value store data, this proposed architecture will propagate transaction logs and applies them in such a way that the state of key-value store is exactly the same as the relational database. To reduce the replica lag in the key- value store side, proposed a novel concurrency control algorithm that guarantees a predefined serialization order (the one same as the order in transaction log).

## References

[1] “A Lock Based Algorithm for Concurrency Control and Recovery in a Middleware Replication Software Architecture”, J.E. Armend´ariz and J.R. Gonz´alez de Mend´ıvil, Dpto. de Matem´atica e Inform´atica, Universidad P´ublica de Navarra, Campus Arrosad´ıa s/n, 31006 Pamplona, Spain, Email: {enrique.armendariz, mendivil}@unavarra.es

[2] “Concert Ticket Selling System by Passive Replication, May Phuu Ko, San San New, University of Computer Studies, Yangon.

[4] “Implementing the Update Propagation Strategies on Replicated Database”, [3] PSC 2010, December, Su Su Mon, University of Computer Studies, Yangon, implementing update propagation with Eager replication on master architecture.

[6] “Transactional Replication in Hybrid Data Store Architectures”, Hojjat Jafarpour, NEC Labs America, 2015.

# Implementation of Data Back-up and Recovery for Distributed System Using Secure Erasure Coding and Secure Data Transmission by LOM

May Htet Shan, Thandar Aung
*University of Computer Studies, Mandalay*
*mayhtetshan204812@gmail.com, thandarmyintmyathein@gmail.com*

## Abstract

*The need for distributed in the corporate world of today is growing as a result of the growth of commerce, which has an impact on the opening of new branch offices across the globe. To ensure the prevention of data loss, it is crucial to recover the database and go back to normal as soon as possible if a breakdown impacts the operation of a distributed system. The database and users are protected from unneeded troubles by the recovery process from data loss prevention. This system uses a remote data backup technique to address this issue (i.e. Secure Erasure Coding (SEC) Algorithm) in iBanking system. This system also focuses on the security concept for the back-up files stored at remote server, by using Location of Matrix (LOM) steganography technique. The aim of LOM is to hide secret data in the innocent looking carrier e.g., in normal transmissions of users.*

**Keywords**: Recovery, Secure Erasure Coding, Location of Matrix, Steganography

## 1. Introduction

Internet banking, commonly referred to as i-Banking, is a key component of payment transactions in the commercial world. The sum must have been taken out of one account and added to another during the user's payment transaction. Data consistency is crucial during this transaction processing in case there was a mistake or the transaction was terminated.

The data for both the payment and the receipt are not modified in this transaction because this system offers original data restoration depending on the checkpoint. The recovery procedure runs as the major component by recovering that transaction from a backup and then redoing work to commit in order to maintain the consistency of the data. This system's goal is to prevent data loss by using Secure Erasure Coding to restore the database to a correct state when the present state is inconsistent during the time when i-Banking transactions are being processed.

Data that is accurate and readily available when needed may thereby increase customer satisfaction.

Business enterprises now depend heavily on data. In the commercial sector, transaction processing systems are crucial for handling data transactions. The essential necessity of data exchanged or transmitted between the servers and their users is the secure data storage on dispersed settings. Steganography is one of the most effective methods for distributed systems secure communication. Steganography is the practice of composing secret messages so that only the sender and the recipient can safely decipher and transmit the information concealed in the means of communication. The safe transaction passing is maintained by this system using steganography as well.

## 2. Related Work

The banking system facilitates system simplicity and efficiently responds to user requests as computer science advances. When abort transactions are present, the system must prompt users to redo or undo them until the transaction commits. This technology can be used on a worldwide scale and will work better if customers can interact with it more. By imposing a set of rules and guidelines that outline how to prevent data loss utilizing forward and backward recovery, this system serves many jobs. Additionally, it will cut down on operation time and might meet client demands. [1].

The banking system is one of the key players in buying and selling transactions in the commercial world. This system plans to use outside transactions. A logical unit of work is called a transaction. A system that keeps track of transaction programs is known as a transaction processing system (TPS). The core of modern company operations are TPSs. The unit of recovery is indeed the transaction. In a database system, recovery refers to setting the database back to its original state when a failure has made the current state inconsistent.

Both internal and external bank transactions are possible. An internal transaction happens when a user switches an account to another client of the same bank. An external transaction happens when a user transfers funds from one customer of one bank to another. In order to deliver accurate data and prompt services, this system will apply external transactions, prevent data loss during bank transfers during transaction periods, restore the database to a consistent state when the current state is inconsistent, and prevent data loss during transaction periods. To encourage completion in the wholesale power market, Transaction Management System (TMS) granted access to their transmission system [2]. Their TMS is intended to automate, integrate, and coordinate security procedures as well as the market interface. Their system is quick and dependable enough to handle the hourly market transactions and supported the management of transmission and ancillary service. The unit of recovery is, in fact, a transaction [1]. Business organizations usually use recovery methods for data security and reliability [11].

## 3. Motivations

A crucial component of data protection in businesses, remote backup provides this purpose. Since it ensures the organization's operational functionality and continuity in the event of loss or damage to crucial data that may affect them, such as:

- Loss of the personnel database of the company
- Corruption of the database and/or application that controls automated production processes
- Loss of historical organizational data while the company is in the process of auditing or evaluating quality standards
- Information loss relating to the balance sheets and management indicators of the various organizational activities
- Computer or device theft, even by those who deal with private management information

Although data security is the primary goal of remote data backup in companies, it is also important to provide data recovery in the event of an emergency or complete information loss. Data Security is crucial in this type of communication. Consequently, it is necessary to safeguard data against hostile attacks.

## 4. Background Theory

Knowing precisely which failure categories to take into account, how frequently they will occur, how much recovery time is anticipated, etc., is necessary for the design and implementation of a recovery component. Additionally, assumptions must be made on the dependability of the supporting hardware and storage mediums as well as the relationships between various failure types. However, due to the following factors, the list of projected failures can never be exhaustive:

- At least one failure was overlooked for any group of failures that one may name.
- Some failures are quite uncommon. It might be a wise design choice to ignore these failures since the cost of redundancy required to handle them might be too costly. However, if one of them does take place, the database will be corrupted and the system won't be able to automatically recover.

The following are descriptions of the primary failure kinds [7]:

**Transaction Failure**: The previous section has made reference of the transaction of failure. The transaction program must be restarted from the beginning for a variety of reasons, either at the program's initiative or at the DBMS's behest. The ratio of self-aborting transactions inside a single application is essentially constant, depending solely on the quantity of inaccurate input data, the effectiveness of the consistency checking done by the transaction program, etc. The amount of parallelism, the level of

granularity of locking employed by the DBMS, the logical schema (there may be hot spot data, or data that are frequently referenced by many concurrent transactions), and the level of interference between concurrent activities all play a significant role in the ratio of transactions that the DBMS aborts, especially those caused by deadlocks (which is, in turn, very application dependent).

For the purposes of our classification, it is sufficient to state that transactions fail 10-100 times per minute and that recovering from these failures must happen within the time necessary for the transaction to execute normally.

**System Failure**: The system failures under consideration could be brought on by a defect in the DBMS code, an issue with the operating system, or a hardware malfunction. Each of these situations results in an abrupt termination of operations, and we suppose that the data in main memory is lost. We expect a recovery to happen in the same amount of time that would have been required for the execution of all stopped transactions because database-related secondary (nonvolatile) storage is unaffected. Recovery shouldn't take more than a few minutes if one transaction is completed in the range of 10–1 second. Based on the stability of the DBMS and its operational environment, a system failure is assumed to happen multiple times every week [5].

**Media Failure**: User must prepare for the loss of some or all of the secondary storage housing the database in addition to these more or less typical failures. Such an issue can have numerous causes, the most prevalent of which are as follows:

- bugs in the operating system routines for writing the disk,
- hardware errors in the channel or disk controller,
- head crash,
- loss of information due to magnetic decay.

Only complete redundancy, which entails a copy of the database plus an audit trail detailing recent events, may remedy such a problem. Recovery from a media failure is most likely to occur once or twice a year at the most because magnetic storage technologies are often quite dependable. Recovery of this kind will take on the order of an hour, depending on the size of the database, the media used to store the copy, and the age of the copy.

### 4.1. Remote Backup Services

The primary goal of the remote backup facility is to assist users in gathering data from any remote place, even in the absence of network connectivity or when the data is not present on the primary server. The following topics [6] should be covered by the remote backup services:

- Data Integrity
- Data Security
- Data Confidentiality
- Trustworthiness
- Cost efficiency

**Data Integrity**: Data Integrity is concerned with the server's entire structure and current state. It confirms the accuracy of the data to ensure that it is preserved during transmission and reception. It serves as a gauge for the accuracy and reliability of the data stored on the server.

**Data security**: The remote server's top objective is to provide complete protection for the client's data. And it should not be accessible to third parties or any other users or clients, whether on purpose or accidentally.

**Data Confidentiality**: It is sometimes necessary to protect the privacy of clients' data files such that, in the event that multiple users are concurrently accessing the server, any files that are private to a single client must be able to hide from other clients while being accessed.

**Trustworthiness**: The attribute of Trustworthiness must be present on the distant server. Users/clients store their confidential data on the primary server; hence it is imperative that both it and the remote backup server perform a trustworthy function.

**Cost efficiency**: The cost of the data recovery process should be reasonable so that the greatest number of businesses and clients can use the service for backup and recovery. These difficulties have been the focus of numerous techniques. This data recovery in banking system will demonstrate a back-up and recovery strategy

for distributed computing that addresses the aforementioned problems.

### 4.2. Steganography

The majority of communication in today's society takes place through technological channels. Data Security is crucial in this type of communication. Consequently, it is necessary to safeguard data against hostile attacks. Steganography, which allows for the confidentiality of communication through an unsecured channel, is the science of concealed data in another media [9]. By avoiding illegal use modification, it defends against unauthorized parties. These methods are designed to conceal sensitive information (steganograms) in otherwise innocent-looking carriers, such as user transmissions that appear to be regular. Steganography work has been done on a variety of media, including pictures, videos, text, and sounds. Perceptual transparency, resilience, and hiding capability are three crucial factors in the creation of steganographic techniques. The ideal carrier for steganograms should have two characteristics: it should be well-known, and any modifications made to the carrier in order to insert the steganogram shouldn't be "visible" to anyone who isn't familiar with the process.

## 5. The Proposed System

Only authorized users are permitted to access the online banking system. To prevent data loss during the transaction period, this system uses SEC method data recovery for data processing transactions. It serves as a protection against unplanned data loss. The system seeks to modify the balances of the sender and receiver accounts when a transfer transaction is started. Then, the sender and receiver accounts balances are created as Seed Block backup by manipulating XOR with generated random number [SEC step 1 and step 2. Detail explained in section 5.1]. Then, sender and recipient both update the original database balance in their respective accounts when a transaction commits. If the transaction is success, the original balances backup will be removed.

Unfortunately, the system displays a fail notice to the sender when a transaction is terminated before commit. To commit a transaction from a transaction backup, the system must first recover and restart. The system will use SEC's backup and adjusts the balance of the sender and receiver sites when the transaction commits. In the original balance recovering, the system will be recovered by XOR manipulating the seed value and seed block backup [Step 5 of section 5.1]. Then, restart the transaction with the roll back balances and make the transaction processing.

The system then displays a success notification to the sender. This system tests the effectiveness of the SEC algorithm in preventing data loss during online system transactions.

By using steganography, this system also regulates data security during data transmission. The cover text and secret text data are covered up using ASCII code. Steganography is also utilized in this system for secure data transmission from the user to the cloud and vice versa. This system relies on database backups for reliable storage. According to the case study's findings, LOM can be used to extract the original data with the least amount of loss.

### 5.1. Secure Erasure Coding Algorithm (SEC)

The Secure Erasure Coding method prioritizes ease of backup and recovery. It basically applies the computing world's Exclusive- OR (XOR) operation principle. For ex: - Let's say we have A and B as our two data files. A and B were XORed to make X, or X = A (XOR) B. It is quite simple to recover A data file with the aid of B and X data files if we anticipate that A data file will be destroyed and we need to obtain A data file back. A = X (XOR) B, for example. The Secure Erasure Coding Algorithm strives to offer a straightforward approach for backup and recovery. The primary server, its clients, and the Remote Server make up its architecture. Here, we start by assigning each client a distinct client id and a random integer to the server. Second, whenever a client identifier is registered on the main server, the client identifier and a random number are EXORed (XORed) to create a seed block for that specific client. Each client's created seed block is stored on a remote server.

SEC Algorithm:

**Initialization:** Main Server: $M_s$; Remote Server $R_s$;
Clients of Main Server: $C_i$; Files: $a_1$ and $a'_1$;
Seed block: $S_i$; Random Number: $r$;
Client's ID: $Client\_Id_i$

**Input:** $a_1$ created by $C_i$; r is generated at $M_s$;

**Output:** Recovered file $a_1$ after deletion at $M_s$;

**Given:** Authenticated clients could allow uploading, downloading and do modification on its own the files only.

Step 1: Generate a random number.
Int $r = rand\ ();$

Step 2: Create a Seed Block $S_i$ for each $C_i$ and Store $S_i$ at $R_s$.
$S_i = r \oplus Client\_Id_i$ (Repeat step 2 for all clients)

Step 3: If $Ci/Admin$ creates/modifies a $a_1$ and stores at $M_s$, then $a'_1$ create as $a'_1 = a_1 \oplus S_i$

Step 4: Store $a'$ at $R_s$;

Step 5: If server crashes $a_1$ deleted from $M_s$, then, we do EXOR to retrieve the original $a_1$ as; $a_1 = a'_1 \oplus S_i$

Step 6: Return $a_1$ to $C_i$.

Step 7: END.

## 5.2. LOM Method in Data Transmission

Conceals sensitive information, allowing any file type to be concealed inside ASCII Text texts. Although it can fit in 7 bits, an ASCII character in 8-bit ASCII encoding takes up 8 bits (or 1 byte). The ASCII code characters in this system proposal are changed at the LOM stage. More steps could result in a more complicated attack. However, the Stego Object's length does not alter. The embedded data is retained when saving the Stego Object in the "pdf" and "doc" file formats.

**Proposed Algorithm for Embedding:**

Output: A matrix of location (Lom).
Input: A secrete text file (Ts), cover text file (Tc)
Step1: Select the secrete text file (Ts) and the cover text file (Tc) to be uploaded.
Step2: Compute the number of characters in the secrete text file (Ts) and the cover text file (Tc).
Step3: Check if the number of characters in the cover text file (Tc) greater than the number of characters in the secrete text file (Ts), if condition is true continue to step 4, otherwise.
Step4: Conversion of the secrete text file (Ts) and the cover text file (Tc) into ASCII value and then into binary format.
Step5: For all i=1 to 7 repeat step 5 to 9
Step6: For j=1 to rows of cover text file
Step7: Matching the bits of the cover text file (Tc) with the bits of the secrete text file (Ts) is performed.
.If bits of cover text file (Tc)=0 and bit of secrete text file (Ts)=0 then save the number of zero in matrix of locations (Lom).
.If bits of cover text file (Tc)=0 and bit of secrete text file (Ts)=1 then save the number of one in matrix of locations (Lom).
.If bits of cover text file (Tc)=1 and bit of secrete text file (Ts)=0 then save the number of two in matrix of locations (Lom).
.If bits of cover text file (Tc)=1 and bit of secrete text file (Ts)=1 then save the number of three in matrix of locations (Lom) of dimensionality n rows and 7 column.
Step8: Increase the value of location and count variable by 1.
Step9: If count variable is equal to the number of the characters in secrete text file, all the data are embedded successfully.
Step10: Else data is unsuccessfully and then go to step 3.
Output: A matrix of location (Lom).

**Proposed Algorithm for Extracting:**

Output: Secret text file (Ts).
Input: Cover text file (Tc), a matrix of location (Lom)
Step1: Read the cover text file (Tc), a matrix of location (Lom).
Step2: Conversion of cover text file (Tc) into ASCII and then into binary format.
Step3: Calculate the length of a matrix of location (Lom).
Step4: For all i=1 to 7 repeat steps 5 to 6
Step5: For j=1 to length of a matrix of location (Lom).
Step6: Match the values of matrix of locations (Lom) and the matrix of cover text.
If bits of cover text file (Tc)=0 and bit of matrix of locations (Lom)=0 then

save the number of zero in extract matrix (Eom).

If bits of cover text file (Tc)=0 and bit of matrix of locations (Lom)=1 then save the number of one in extract matrix (Eom).

If bits of cover text file (Tc)=1 and bit of matrix of locations (Lom)=2 then save the number of zero in extract matrix (Eom).

If bits of cover text file (Tc)=1 and bit of matrix of locations (Lom)=3 then save the number of one in extract matrix (Eom).

Step7: Increase the value of location and count variable by 1.

Step8: Conversion of the extract matrix (Eom) from binary to ASCII format.

Step9: Conversion of ASCII format to character format.

Step10: Display the secrete text (Ts).

Output: Secrete text file (Ts).

### 5.3. Experimental Results

In order to tackle the problems that relate to time, the SEC will recover the process in the shortest amount of time possible. The storage/backup recovery component of the suggested system is what turns a failed file request transaction into a successful transaction. As a result, this system is trustworthy for recovering lost data during the processing of file requests for transactions since the original database and data backup are stored side by side.

This study also suggested a method for securing data storage that involved concealing a covert English text file that contained a secret file by creating a matrix of locations. This approach has a number of benefits. First off, the suggested strategy increases the capacity for data concealing. Second, users can conceal a greater volume of data without distorting the cover text file, making the changes that are reflected almost insignificant. Average recovery times on each type of transactions are shown in following figure 1. Each type of transaction is tested by 10 times for recovery time consuming.

On the other hand, encrypting a location matrix and adapting it to any language can increase the security of the proposed method.

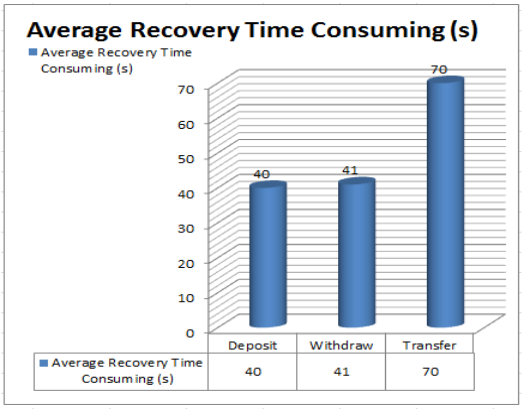



**Figure 1.**

## 6. Conclusion

Secure Erasure Coding Algorithm is a suggested remote data backup algorithm for distributed systems that enables users to gather data from any remote point in the absence of network access. When the resources allocated are comparatively adequate, it is possible to ensure that a task will always be completed by its due date, even if the workload characteristics of the activity are incorrectly predicted. In the event of deletion or if the server is destroyed for whatever cause, the Secure Erasure Coding Algorithm is strong in assisting the users in recovering the data/files. The processing of secure transactions is then supported by LOM as well.

## References

[1] Mr. G. S. Narke, “A smart data backup technique for cloud computing using seed block algorithm strategy”, Comp. Dept. BVCOERI Nasik, India, International Research Journal of Engineering and Technology (IRJET) 2015.

[2] Mahantesh N. Birje, Praveen S. Challagidad, “Remote backup and recovery review: concepts, technology, challenges and security”, International Journal of Cloud Computing, InderScience Publishers, vol. 6, issue 1, 2017.

[3] Ruchira. H. Titare, Prof. Pravin Kulurkar, “Remote Data Back-up and Privacy Preserving Data Distribution: A Review”, International Journal of Computer Science and Mobile Applications, Vol. 2, Issue. 11, November 2014.

[4] Malinowski, E. an Chakravarthy, S. (2017), Fragmentation techniques for distributing object-oriented database, in D.W. Embley & R.C Goldstein, eds, ‘Conceptual Modeling – ER ‘97’, Vol . 1331 of lecture notes in computer science, springer, PP, 347-360.

[5] Amman, P., et. al., "Recovery from Malicious Transactions", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, 2015.

[6] Ghazi Alkhatib and Ronny S. Labban, “Transaction Management in Distributed Database Systems: the Case of Oracle’s Two-Phase Commit”, Senior Lecturer of MIS, Qatar College of Technology, Doha, Qatar and Computer & Communications Engineer; Consolidated Contractors International Company Athens, Greece; Alkhatib@qu.edu.sa and r.s.labban@ieee .org.

[7] Tripathy, S. and B. Panda, “Post-Intrusion Recovery Using Data Dependency Approach”, Proceedings of the 2021 IEEE Workshop on Information Assurance and Security, pp. 156-160.

[8] NA Garg & KA Kaur. (2016). Hybrid information security model for cloud storage systems using hybrid data security scheme. International Research Journal of Engineering and Technology (IRJET). Vol, 03 Issue: 04.

[9] MA Wojciech & SZ Krzysztof (2011). Is cloud computing steganography-proof. IEEE.

[10] UD Kamred (2014). A Steganography Technique for Hiding Information in Image. International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS). ISSN (Print): 2279-0047 ISSN (Online): 2279-0055.

[11] AR Malik, GE Sikka, & HA K. Verma (2016). A high-capacity text steganography scheme based on LZW compression and color coding. Engineering Science and Technology, an International Journal.

[12] Vaishali & AN Goyal. (2014). An Implementation of 4 Bit Image Steganography for Data Security in Clouds. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 11.

# Concurrency Control in E-Tendering System using No Wait Locking with Notification and Decision Making with MAUT

Kyi Linn Lett Thar, Thandar Aung
*University of Computer Studies, Mandalay*
*Linnlett392@gmail.com, thandarmyintmyathein@gmail.com*

## Abstract

*Because of the quickening speed of technological development, governments and businesses today heavily rely on ICT for communication. The electronic publication, communication, access, receipt, and submission of all tender-related information and documents via the internet is known as e-tendering (Electronic-Tendering). producing a more efficient and effective business process for all parties involved by replacing the conventional paper-based tender processes. Concurrency will be required to manage the concurrent tender value submitting procedure due to the involvement of several users. Concurrency control is a crucial feature that every DBMS must have. Concurrency management mechanisms are used to guarantee database consistency while multiple users are concurrently accessing it. The "no wait locking with notification" approach will be employed by the proposed system to manage concurrency during the tendering process. The system will employ multi attribute utility theory (MAUT) to determine the winner when the tendering period has ended.*

**Keywords:** e-Tendering, concurrency control, no wait locking with notification, MAUT

## 1. Introduction

E-tendering is the electronic exchange of bids. E-tendering will lessen the workload for traditional tender management and increase the speed and time required to complete a purchase. A website created specifically for electronically sharing information and tender documents is known as a "e-tendering portal."

Principal and Tenderer play a crucial role in e-tendering. The individual that produces, administers, and sends electronic contract announcements is known as the principal. A tenderer is someone who submits a bid for a proposition. Electronic commerce aims to fully or partially automate each phase. Evidence suggests that automated negotiating infrastructure will be crucial to electronic commerce. Therefore, this kind of system is required to add concurrency control to the simultaneous transaction processing involving several users. Numerous building, service, and sales of commodities contracts are procured via electronic tendering methods. The fundamental structure of e-Tendering system is shown in figure 1.

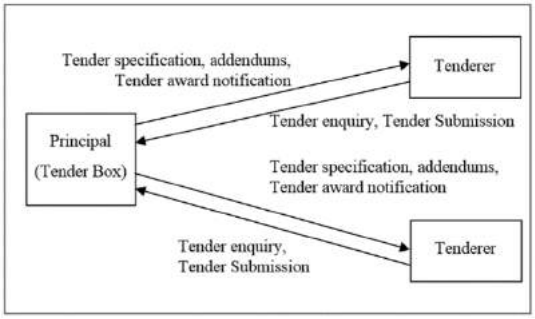



**Figure 1: Principal Based Architecture of e-Tendering**

The use of technology within the business, its capacity to transmit a huge volume of documents and information between numerous parties, convenience of use, and speed of tender filing are just a few elements that have contributed to the growing use of electronic tendering.

The selection of the best candidate must be based on a variety of elements throughout the evaluation phase of the electronic tendering. Therefore, the MAUT will use this mechanism to make decisions.

## 2. Related Work

Pessimistic are built on locking, recognize conflict as soon as it arises, and use blocking to resolve it. Locks control the concurrent accesses to shared data elements. Other transactions that need

to access a data item that has been locked by a transaction will be halted until the lock on the item has been released. A set of actions in a pessimistic system are unable to move forward because each action is waiting for a lock held by another action in the group, which might cause unbounded waiting due to blocking and result in deadlock situations. If transactions wait a long period in a blocked state, this indicates low throughput [1].

An optimistic concurrency control mechanism's fundamental tenet is that transactions should be executed in three phases: read, validation, and write. After the data object has been accessed, a conflict is found in all optimistic concurrency control (OCC) techniques. When a transaction completes its execution and requests the concurrency control manager to validate all of its accessible data objects, conflict detection and resolution are both carried out in the OCC at the certification time. It enters the commit phase, when it writes all of its alterations to the database, if it has not yet been flagged for abortion.

Concurrency control techniques based on timestamps [2] offer the advantageous characteristics of being non-blocking and deadlock-free. Because it is necessary to obtain the proper lock (by sending a lock request) before accessing an object in a pessimistic system with locking, Thus, even while reading an object, a network delay must be applied roundtrip. This pause is required to make sure that the locking requirements are upheld and that the operation reads (or modifies) the most recent version of the object. Because the cost of a message is independent of message size for objects of reasonable size, it should be noted that a network lock request is nearly as expensive as a request for a duplicate of the object. An action can read objects in an optimistic system (a timestamp-based system) without utilizing any network messages.

The number of messages that must be processed by the system can be decreased if the system is optimistic. As more messages are transmitted, both the transmission and processing times of the messages increase. In a pessimistic system, it is necessary to send one message for each lock request and additional messages at the time an action is committed. Thoughtful strategy: only send messages at commit time. Since they must contain some additional information used for validation, the optimistic commit-time messages would be larger than the pessimistic commit-time messages. The quantity of messages, as opposed to the size of each message, is more crucial. [6] Positive strategies outperform locking methods [8, 4].

The other author likewise used multi-agent technology to develop the Agent-Based Online Shopping Assistant System as a shopping assistant system. Users may find it challenging to select the ideal product within their price range when purchasing goods with several qualities and costs. The collection of pertinent things within the buyers' budget can be found through their shopping assistant system [3]. Users can save time and effort by using the online shopping method over the Internet. As a result, this thesis utilizes effective Multi Attribute Utility Theory to present the shopping assistant system for both customers and retailers (MAUT).

## 3. Background Theory

Concurrency refers to many users accessing the database simultaneously. A concurrency control mechanism's job is to maintain database consistency while allowing a number of transactions to run simultaneously. Each user should have a consistent view of the data, including any modifications that were made as a result of their own transactions as well as those of other users [10].

### 3.1. Consistency Problems

Consistency problems causes by concurrent processing include-

- Lost or buried Updates
- Inconsistent Analysis (Non repeatable Read)
- Uncommitted Dependency (Dirty Read)
- Phantom Reads

#### 3.1.1. Lost or buried Updates

When two or more transactions read and change the same data item in the same database, the issue arises. The other transactions are unknown to each one.

If an item was read for an update by a subsequent transaction after the first transaction had already done so but before the first transaction had committed. Whichever transaction commits first will result in the loss of that update.

### 3.1.2. Inconsistent Analysis (Non repeatable Read)

A transaction should always read the same value if it reads the same data item more than once.

When a second transaction accesses the same data item several times and reads different data each time because another transaction altered this item while the second transaction was reading, this situation is known as a non-repeatable read. Multiple reads (two or more) of the same item are involved in inconsistent analysis, and each time the information is altered by another transaction, hence this word refers to an unrepeatable read.

### 3.1.3. Uncommitted Dependency (Dirty Read)

A transaction, if it retrieves or updates a piece of data that has already been updated but hasn't yet been committed by another transaction. Dirty reading is similar to inconsistent analysis in that the change was performed by the other transaction after the item was read by the first transaction.

### 3.1.4. Phantom Reads

Even though the search condition is the same, a transaction re-executes a query and discovers a collection of data that is not equal to the prior one. When an insert or delete operation is made against a row that is a part of the range of rows being handled by a transaction, phantom reads may result.

## 4. Implementation of the System

The e-tendering system is what this system wants to implement. This system's multi-user capability to bid on one or more items simultaneously necessitates the need for concurrency management. The user must check the most recent updated bit amount from the auction server when they want to place a bid on an item. A bidder must choose whether to increase their offer and by how much, or to withdraw, during this phase. The auction bidder can revise their value estimations based on other bidders' offers during the middle stage.

In order to control the consistency of the clients' cached data, this system must keep the data fresh and broadcast to all other clients utilizing "No-wait locking with notification". The system needs to decide which tender will effect the outcome when tending processing reaches its end period. The selection of the system will be based on a number of factors during the determination phase. In order to choose the winner, this system will use MAUT for decision-making. Figure 2 depicts the proposed system's processing phases.

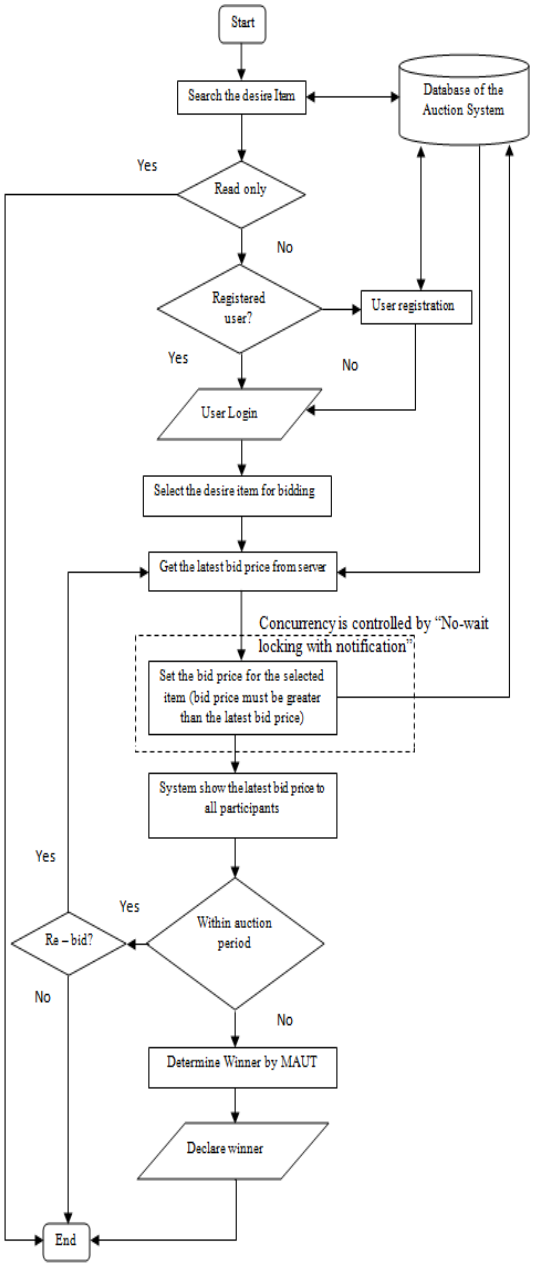



**Figure 2: The System Flow**

### 4.1. "No-Wait Locking with Notification" Algorithm

```
BEGIN
SD ← Set Number of Clients and Clients' ID;
If (type of transaction is "Read")  {
        Read Commit ();
}
Else If (type of transaction is "Write")
{
   If (No concurrent access)  {
       Write Commit();
       Check_Other_Clients (ItemID, ClientID);
// server checks number of clients and client ID
for updated item
        Send_Notification(ClientID, ItemID,
        "Notification!")
//server sends notification messages to all other
clients
    }
    Else If (Has concurrent access on the same
data item)
    {
        Check_Timestamp(ClientID=1,
        ClientID=2, ItemID, Timestamp);
        If (ClientID 1has early timestamp)
        {
            ClientID1.Write Commit ();
            Check_Other_Clients(ItemID,
            ClientID);
 // server checks number of clients and clientID
for updated item
            Send_Notification (ClientID, ItemID,
            "Notification!")
//server sends notification messages to all other
clients
        }
        Else If (ClientID 2 has early timestamp)
        {
            ClientID2.Write Commit();
            Check_Other_Clients(ItemID,
            ClientID);
// server checks number of clients and client ID
for updated item
            Send_Notification(ClientID, ItemID,
"Notification!")
//server sends notification messages to all other
clients
        }
    }
}
END
```

A transaction can be terminated in no-wait locking if it reads an incorrect object or becomes stuck in a deadlock. Notification can be incorporated into no-wait locking during deadlocks to lessen transaction aborts brought on by reading faulty data items. This is accomplished by mandating that whenever a data object is updated by a committed transaction, the server send the updated object to clients. At a later point, client transactions will read valid objects and then commit. Because a transaction can still read an invalid object before receiving updates from the server, notification cannot completely prevent reading faulty objects. Clients may potentially get simply an invalidation message from the server rather than updates.

### 4.2. Multi-Attribute Utility Theory (MAUT)

The Multi-Attribute Utility Theory (MAUT) is a normative approach for assessing objects with several conflicting qualities. One method used in Multiple Criteria Decision Making (MCDM) is MAUT.

It is a technique for making decisions when a person must weigh several different goals. Additionally [7], many systems employ it as an evaluation technique to determine the interests and preferences of consumers and assist them in configuring the desired product (s). The MAUT calculates the degree of interest (or utility) of the products in light of the user preferences by evaluating not just one user's preferences but those of multiple users. It enables the decision-makers to express their preferences for any kind of attribute and to integrate it in the target system.

In particular, MAUT is a structured technique created to handle the tradeoffs between numerous, incomparable, and conflicting aims that are represented by numerous qualities. When evaluating potential alternatives to decide which alternative(s) performs best, this strategy is advised. The purpose of MAUT is to assist decision-makers in making choices among numerous potential solutions while taking into consideration their preferences. The utility evaluation of choice alternatives is based on several attributes in instances where decision making is dependent on multiple variables. As a result, MAUT has been widely applied in the field of e-commerce for financial and economic decision-making. This system will use MAUT to

examine the winner during the decision-making process.

Multi-Attribute Utility Theory includes various utility functions (MAUT) [9]. However, the MAUT functions that are most frequently utilized are the multiplicative and additive utility functions. Utilizing the additive function is used in this system.

The overall utility of an option is determined by the weighted sums of its measures in the additive utility function of MAUT (i.e. evaluation criteria).

It is described by the following equation (1):

$$U(x_1, \dots, x_n) = \sum k_i U_i(x_i = 1) \quad \dots \text{eq(1)}$$

Where,

- $U(x_1,\dots,x_n)$ = the overall utility score of each alternative
- $U_i(x_i)$ = the utility function of the $i^{th}$ attribute
- $k_i$ = the weight of the $i^{th}$ attribute
- $0 \leq U(x_1,\dots,x_n), U_i(x_i) \leq 1$
- $k_1 + \dots + k_n = 1$

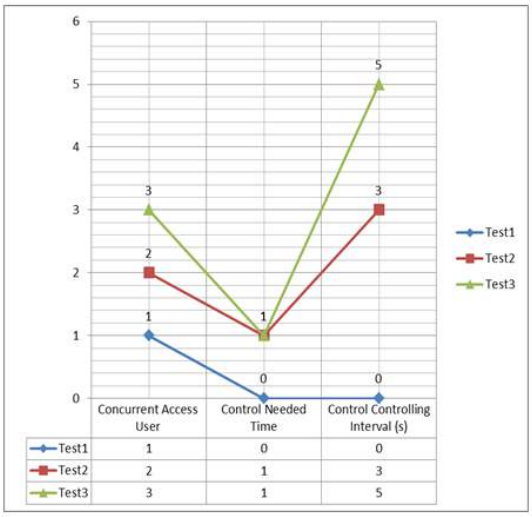
### 4.3. Processes of Multi-Attribute Utility Theory (MAUT)

There are five steps in the Multi-Attribute Utility Theory procedures. The following is a description of the processes:

1) List the characteristics that jointly describe the overall usefulness of all pertinent decision-making choices.
2) List the possibilities or alternatives to be calculated.
3) Give each attribute a weight based on how significant it is.
4) Convert the attribute scores, which are expressed in various units, into commensurate (measurably equivalent) units.
5) Describe the aggregate utility function, which combines the weights and transformed scores to determine the overall usefulness of each choice.

### 4.4. System Evaluation

In this proposed system, experiments for concurrency control are made for three times. The controlling status in each time is briefly described as shown in following figure 3.




## 5. Conclusion

In order to control the consistency of the clients' cached data, this system must maintain data freshness and multicast to all other clients utilizing "No-wait locking with notification." The system needs to decide which tender will effect the outcome when tending processing reaches its end period. For data-shipping DBMS designs, a successful concurrency management technique called "No-Wait Locking with Notification" has been introduced in order to reduce the delay for identifying data conflicts and to relieve the server bottleneck. The selection of the system will be based on a number of factors during the determination phase. In order to choose the winner, this system will use MAUT for decision-making. The best matched user preferences are assessed using the multi-attribute utility theory (MAUT). By looking for things within the human's price range and preferences, this technology can help. They can use this to save time and effort when looking up and assessing the necessary product information, and they can select the best match winners from the most suitable ones by doing so.

## References

[1] Adya, A., Gruber, R., Liskov, B. and Maheshwari, U. “Efficient optimistic concurrency control using loosely synchronized clocks,” in Proceedings of the

ACM SIGMOD Conference on Management of Data. San Jose, CA, pp. 23– 34.

[2] Carey, M.J., Franklin, M.J., Livny M. and Shekita, E. J. “Data Caching Tradeoffs in Client-Server DBMS Architectures,” in Proceedings of the ACM SIGMOD, pp. 357-366.

[3] Franklin, M.J. “Client Data Caching: A Foundation for High Performance Object Database Systems,” Kluwer Academic Publishers, Boston, MA.

[4] Franklin, M.J., Carey, M.J. and Livny, M. “Transactional client-server cache consistency: alternatives and performance,” ACM Transactions on Database Systems, vol. 22(3), pp. 315-363.

[5] Laudon, J. and Lenoski, D. “The SGI Origin: A ccNUMA highly scalable server,” in Proceedings of the 24th Annual International Symposium on Computer Architecture, vol. 25(2), pp. 241-251.

[6] M. Shanmuganathan, K.Kajendran, A.N. Sasikumar, M.Mahendran, "MultiAttribute Utility Theory – An Over View", International Journal of Scientific & Engineering Research Volume 9, Issue 3, March-2018, Faculty, Dept of C.S.E, Panimalar Engineering College, Chennai, TamilNadu, India.

[7] Maw Min and Nyein Nyein Oo, "Mobile Agent based Information Retrieval for Shopping Assistant", Proceedings of 2015 International Conference on Future Computational Technologies, (ICFCT'2015), Yangon Technological University, Yangon, 2015.

[8] Ozsu, M.T., Voruganti, K. and Unrau, R. “An asynchronous avoidance-based Cache Consistency Algorithm for Client Caching DBMSs,” in Proceedings of the Conference on Very Large Data Bases (VLDB). New York, NY, pp. 440-451.

[9] Su Myat Kyaw Lin, "Implementation of Multiple Attribute Reverse English Auciton using Multi-Attribute Utility Theory (MAUT)", University of Computer Studies, Yangon, 2014.

[10]Wang, Y. and Rowe, L.A. “Cache consistency and concurrency control in a client/server DBMS architecture,” in Proceedings of the ACM SIGMOD Conference on Management of Data. Denver, CO, pp. 367–377.

# Consistency Control in Group-Work Discussion Using Eager Invalidation

Khaing Thazin Hlaing Myint, Aye Mya Sandar
*University of Computer Studies, Mandalay*
*??@gmail.com, ??@gmail.com*

## Abstract

*People now depend on information technology for a variety of purposes (Especially, for business and other organizations). Therefore, they employ any systems connected to their work to achieve their goals. A distributed system consists of a collection of autonomous computers, connected through a network and distribution middleware, which enables computers to coordinate their activities and to share the resources of the system, so that users perceive the system as a single, integrated computing facility. This system focused on knowledge sharing in a private organization through group work discussions using distributed systems. Through the suggested system, members of proposed group-work discussion system can learn, communicate, and express their opinions from any location. The eager invalidation approach allows for parallel update transactions from several users in the group on the same document. In order to obtain reliable data, this suggested system offers consistency management utilizing eager invalidation.*

**Keywords:** Eager Invalidation, knowledge sharing, distributed system

## 1. Introduction

The distributed system is now playing a bigger role in information technology.

It consists of a group of autonomous computers that communicate with one another over a network connected by some type of hardware and share specific resources in an effort to do simultaneous calculations. The majority of real-time application fields are built using distributed systems. Data sharing capabilities are enabled by distributed systems, and this can offer the user a wealth of advantages.

Data Consistency Control plays a crucial function in Distributed Data Sharing Systems. The concurrent updated transaction is managed by the proposed system's distributed system consistency control for group project discussion utilizing the eager invalidation approach. When a data item in the group project is modified, eager invalidation may broadcast the invalidate message beforehand [5].

As a result, even when they do not request updated data checking, system users can be trusted for data consistency.

## 2. Related Work

The related works of concurrency controls are discussed in this session.

The consistency for replication in distributed systems [1] is a significant challenge. As a result, we must make sure that all copies are updated when one copy is modified in order for the duplicates to remain identical. Different consistency models for distributed shared memory systems have received a lot of attention from parallel computer designers in an effort to achieve high performance of operations on shared data. Advantages: By employing a global clock-like vector clock time synchronization mechanism, this work can manage write consistency and precisely identify which write operation was the most recent. Cons: Client-centric consistency lacks synchronous updates, which is a drawback. Increased propagation latency for data updates may result from this.

The aim of [2] is to stop concurrent users accessing a shared database from making inconsistent retrievals. By employing the train ticket sales system as a case study, this system will put home-based lazy release consistency management and vector timestamp synchronization into practice. Advantage: Those databases are subject to consistency control. This system's objective is to stop concurrent users from accessing a shared database in a way that

results in inconsistent retrievals. They can be discarded after sending copies of the data to the houses. Advantage: The home-based protocol will suffer if the home assignment to the shared pages is not properly aligned with an application's memory access pattern.

This area of research [3] focuses on creating and evaluating algorithms to address communication and data-sharing issues in highly dynamic distributed settings. Work on distributed services that offer helpful guarantees and facilitate the development of complex distributed applications. Disadvantages: Dynamic encompasses a wide variety of changes, including time fluctuations, changing client processes, changing lists of participating client processes, and changing networks. A challenging programming topic is creating distributed applications for these kinds of contexts.

## 3. Background Theory

For many people, sharing data is becoming increasingly vital, especially for companies and organizations looking to increase profits through increased productivity. As an illustration, businesses complete more work and improve the effectiveness of their peer collaboration, which is essential to achieving their corporate objectives. Students gain from working on group assignments since they are better able to collaborate with other members and complete the job more quickly [6].

### 3.1. Distributed System

A distributed system is one in which hardware or software components located at networked computers communicate and coordinate their actions only by passing messages [9]; a distributed system is a collection of autonomous computers linked by a computer network that appears to the users of the system as a single computer.

### 3.2. Distributed Database System

The Distributed Database System (DDBS) technology combines database system and computer network technologies, which on the surface appear to be diametrically opposed approaches to data processing. A distributed database is a logically connected collection of shared data that is kept on computers at various locations throughout a computer network. Users can access the database's contents from any location on the network. Users or programs must behave as though they have local access to the entire database [8] in order for a distributed database to be considered really distributed.

Each site in a distributed database system is a complete database system in and of itself, but the sites have agreed to cooperate so that a user at any site can access data anywhere in the network exactly as if the data were all stored at the user's own site. The sites are connected by some sort of communication network.

## 4. The Requirements for Controlling in Distributed System

Nowadays, Internet becomes important for people as a communication media for information repository. Therefore, giving efficient information of the region around the group is necessary. The need for consistency arises when many clients edit the same record at the same time due to needs. There are many different types of controlling strategies that can be used to regulate the distributed system's data consistency [7].

Lazy Release Consistency

- A further improvement to the release consistency is lazy release consistency.
- It is presumpted that until the acquire access is finished, the thread doing it will not need the values written by other threads.
- Lock release and datum propagate simultaneously.

### 4.1. Types of Lazy Release Consistency

There are four types of lazy release consistency [4]:

1. Lazy Invalidation
2. Lazy Update
3. Lazy Hybrid
4. Eger Invalidation
5. Eger Update

### 4.2. Lazy Invalidation

Only transmit the invalidation message in lazy invalidation when the user requests the validation check.

System: Suitable for the processing system for periodic data updates (e.g., Data backup systems)

Benefits: Processing costs are reduced because not all participants are notified right away when data is changed.

Disadvantages: Unsuitable for real-time systems that require quick updates to crucial data. (Banking system, for instance).

### 4.3. Lazy Update

Never sends an invalidation message. Lazy Update: updates the data only when the user demands it. Allow outdated data on user sites.

Benefits: Updates are available as needed by the user.

Cons: This is inappropriate for e-commerce sites and stock share trading. This is unable to support regularly updated data.

### 4.4. Lazy Hybrid

Lazy Hybrid: a fusion of lazy update and lazy invalidation. This method notifies all clients that their data has been invalidated. When a user requests data, it lazily sends the correspondingly updated data.

Advantage: Updates are delivered when the user requests them.

Send the invalidation notice to every participant as a drawback.

### 4.5. Eager Invalidation

All pages in the cache for which the eager invalidation acquiring processor gets write notices are invalidated. When updating, the participant releases the lock and notifies all other participants who own the relevant modified data item of the invalidation.

System Illustration: Appropriate for Groupware systems (example: Open Source share data editing)

Benefits: There is no need to ask for the participants to receive an invalidation message. The invalidation messages are issued after the updating participant released the lock.

Cons: When a groupware system has a large number of users, the cost of broadcasting invalidation notifications will be considerable.

### 4.6. Eager Update

In eager update, releases the lock when updating and distributes the data update to all participants who own the relevant updated data item. (No requests for data updates are required).

Examples of systems: banking systems and real-time systems.

Benefit: Compatible with banking systems. The system constantly supports timely data updates. The original data (old version) will be lost instantly because the system transmits the upgrade right away. The cost of processing updates will be significant.

## 5. The System Overview

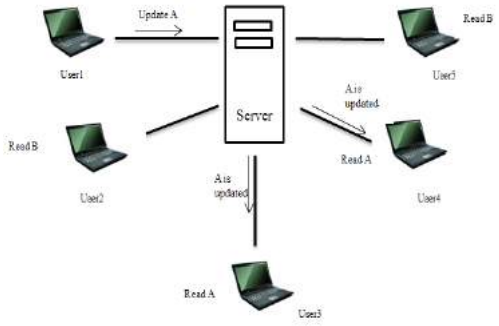



**Figure 1. System Overview**

The Eager Invalidation approach governs the distributed data consistency at each site in the system as proposed. A server organizes this system, and each location receives the server data after that. As a result, each site's client needs to be a registered user on the server. When one of the system's members updates a data item, the server can then send the appropriate data user an invalidation message. Despite the fact that read transactions at each site cannot alter data consistency, write operations can result in a condition where the data is inconsistent. This suggested system will employ eager invalidation to ensure consistency in this situation.

### 5.1. Implementation of the System

The text files will be used as the suggested system shares data. The implicit dynamism of customer behavior increases the challenges in getting precise results. To get precise results: The system first determines whether or not the local cache data and global access data on the server match. If these data contents differ, the document will be downloaded from the global server and then a copy will be kept in the local cache. If these data contents are the same, the user-selected document will be displayed.

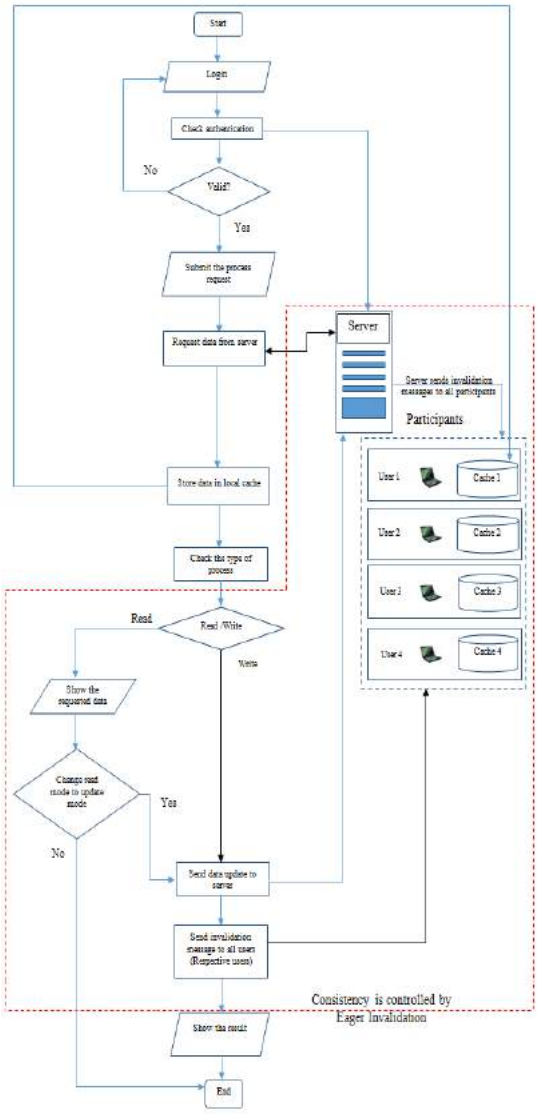



**Figure 2. The System Flow**

The system will be configured with the document's timestamp and action (read or write) at the beginning of every transaction. The system or server then confirms that the status is in a concurrent state.

When a status is concurrent, the system examines the timestamps of the concurrent transactions and instead of locking, sends messages to the conflicting sites. The conflict sites will only be accessible in read-only mode. The client's write set transaction with the earliest time stamp will be given write access. When the client's write set transaction was completed in the early time stamp, the read set transaction knew that the document was empty of data or that another user had previously opened it for writing in the later time stamp.

When updating the global and multicast data updates to other caches, the current transaction of the data contents will be committed if the status is not concurrent.

For a better understanding of the suggested system procedure, the sequence diagram that follows [Figure 3] also depicts the eager invalidation message sending between clients.

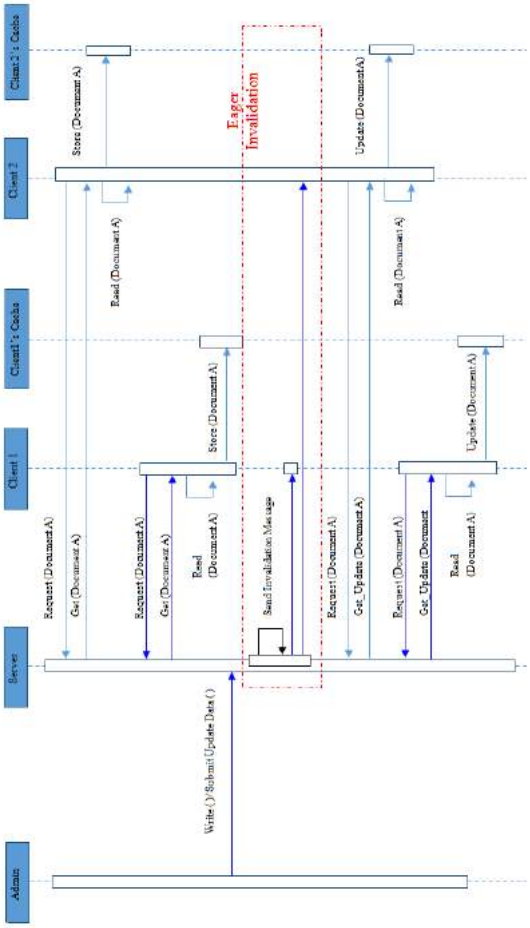



**Figure 3. Sequence Diagram**

### 5.2. Experimental Result

The experiment for consistency control will be made by three types of eager invalidation messages.

**Testing message type1:** a user processes a document (test1.txt) from its local cache by using read-only mode. Soon, another user processes the document (test1.txt) from the server by using read-only mode. Since, the system also controls the concurrent processing, the system sends the message "Another reader is online" to the late user as shown in figure 4.

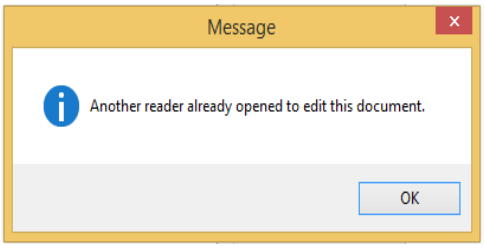



**Figure 4. Message Type 1**

**Testing message type2:** A user changes the processing edit mode to the document (test1.txt). Soon, another user also changes the edit processing mode to the document (test1.txt) from the server. Since, the system also controls the consistency, the system sends the denied message to the late user as shown in figure 5.

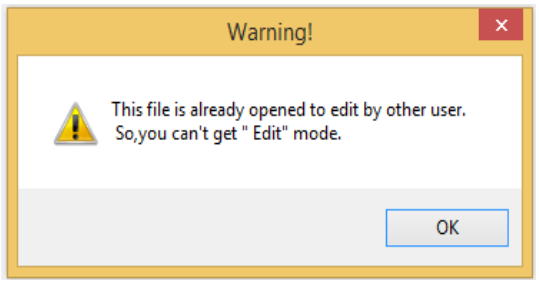



**Figure 5. Message Type 2**

**Testing message type3:** As the early edit user changes the document as a new version, the other user's data will be stale. So, the system needs to inform the other user to get the new version of the data updated by the early user to maintain the system consistency as shown in figure 6.

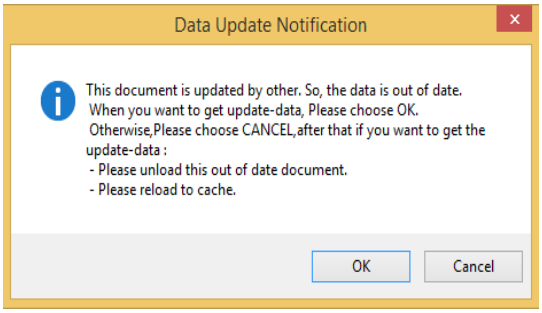



**Figure 6. Message Type 3**

## 6. Conclusion

The tendency of data accesses to reflect a locality of reference is taken advantage of by distributed data. Some data are more likely to be accessed by a specific workgroup more frequently than others. In either scenario, the concept of a database is based on two premises: Every piece of information is always accessible; no one is ever informed that a particular piece of information is unavailable; and every piece of information is under the supervision of transaction management, allowing only one person at a time to really make changes. The system makes sure that requests are queued so that they can be changed one at a time, and everyone can see the outcomes of any updates made before his or hers, if multiple persons wish to edit the same piece of shared information.

As teams or groups of students working on a project can share documents and successfully cooperate with one another, this system offers data sharing features. This enables messaging group members with updated versions of a document. Additionally, data consistency may be verified for both local and global updates via eager invalidation, preventing lost and buried updates.

The data sharing system for concurrency and consistency management is implemented by this system. Because it uses a shared database system, both the server database and the client database are necessary. Validation based on priority can be added to this system. Additionally, the central database system will be improved and its flaws will be fixed by the system that caches data on client sides and uses those cached data when the server is down.

## References

[1] “Implementation of Client-Centric Consistency Control and Synchronization System For Distributed Replication(For Mobile Clients)”, Aye Nyein Mon, Thinn Thu Naing, University of Computer Studies, Yangon, 2010.

[2] “Implementation of Home-based lazy release consistency system for a distributed application”, Zar Zar Moe, Thinn Thu Naing, 2010.

[3] “Communication and Data Sharing for Dynamic Distributed Systems” Nancy Lynch1 and Alex Shvartsman2, 2013.

[4]“A Distributed File System For Distributed Conferencing System”, Philip S. Yeager (University Of Florida), 2016.

[5] Spyros Voulgaris, Maarten Van Steen, Aline Baggio, And Gerco Ballintijn “Transparent Data Relocation In Highly Available Distributed Systems”, 2013.

[6] Philip Homburg, Maarten van Steen, Andrew S. Tanenbaum: “An Architecture for A Wide Area Distributed System”, 2006.

[7]Kjetil Norvag, Olav Sandsta, and Kjell Bratbergsengen, ―Concurrency Control in Distributed Object-Oriented Database Systems‖, Advances in Databases and Information Systems, 201 7.

[8]Maabreh K. and Hamami A., ―Increasing database concurrency control based on attribute level locking‖, on the proceedings of International Conference on Electronic Design, ICED, IEEE, pp1-4, Issue 1-3, Malaysia, Penang. Dec. 2008.

[9] Maabreh K. and Hamami A., ―Implementing New Approach for Enhancing Performance and Throughput in a Distributed Database‖, The International Arab Journal of Information Technology, Vol. 10, No. 3, May 2013.

# Image and Signal Processing

# GENDER CLASSIFICATION FROM MYANMAR (NRC) CARD WITH SUPPORT VECTOR MACHINES (SVM)

Soe Thiri Hlaing, Thin Thin Yu
*Computer University (Kalay)*
*soethirihlaing@ucskalay.edu.mm, thinthinyu.ucsm@gmail.com*

## Abstract

*Gender identification is a fundamental task of face recognition, deciding the gender according to the face image. Today, it is becoming increasingly popular for security. Gender is an important factor in social activities. In this paper, an investigation of gender classification by face from Myanmar NRC card is proposed. In this investigation, there are three parts such as face detection and extraction, facial feature extraction by PCA and classification. Initially, the facial area is detected using the Viola jones algorithm and then face region is extracted from the NRC card. In the next step, the detected face region is subjected to Principal Component Analysis (PCA) to extract facial features. For classification, these principal components are exposed to SVM classifier. This gender classification system is implemented on Matlab by using own dataset. The accuracy is obtained 95.6% at 80 number of training images. Testing error rate is about 4.4%.*

**Keywords:** Gender Identification, Principal Components Analysis (PCA), Viola jones, Support Vector Machine (SVM).

## 1. Introduction

The face is one of the most important biometric features of the human beings and normally used as identification. Each person has their own innate face and mostly a different face. Gender classification [1,2,3] using the face is very useful for human-computer interaction and control systems. In the large population, an individual's authentication process is usually time consuming. The division of the population into two parts based on gender is the possible solution of this issue. The process of gender classification according to face image is needed to use clearly distinguished features and robust classification methods.

In many years ago, the facial images are used in gender classification has become an important role. When you see a person's face, it can be easy to identify a male or a female, but it's a difficult task for a computer. The computers require some significant information to make the classification. Gender identification can be done in different appearance such as the gait, iris, hand shape and voice etc. [3]. However, the most popular techniques for gender classification were always standing on facial features. There are different characteristics between the man and woman to classify the gender.

During the process of gender identification, firstly there are face detection and then studying the facial features is included. Therefore, the machine needs to be the suitable facial features. Principal Component Analysis (PCA) method is the most frequently used for facial classification. Gender recognition can be used in numerous applications such as Identity authentication, data collection, search engine recovery and monitoring.

In this system, the selection of principal components is used as a method for extracting facial features. The structure of this paper is as follows: Section 1 introduces the analysis of facial features and gender classification. Section 2 reviews the gender classification that includes features based on geometrical and appearance features. An overview of the proposed system is provided in Section 3. And Section 4 descries the methods including face detection and extraction and the operation of facial features extraction by PCA. SVM classification techniques is discussed in Section 5. The implementation and experimental results of this system is presented in Section 6. In Section 7, the conclusion of the system is presented.

## 2. Related Works

Today's analysis of human faces is an interesting research area on gender recognition. It

is necessary to be faster and more convenient system. In this study, it is tried to determine gender from Myanmar NRC card face images. The first introduction to gender classification is described in [3]. A multi-layer neural network is used to recognize gender from facial images. They had 91.9% accurate rate on 160 images.

In [4], a hybrid approach was used. It presented a hybrid approach by combining global features with local features. The Adaboost algorithm and active appearance model AAM are used to extract global features and local features individually. They point out that it receives greater accuracy by using the hybrid method. [5] proposed "Ethnicity Identification from Face Images" Ethnicity classification was done by using LDA. Different datasets separated into Asian & Non-Asian. Ensemble LDA gives more accuracy than nearest neighbor classifier.

In [6], gender classification system was presented using linear discriminant classifier and Support Vector Machine. The experiment was done with different types of classification methods namely the cosine classifier, the linear discriminant classifier and SVM. The author of [7] used Continuous Wavelet Transforms to find features for each male and female face. The Wavelet Coefficient obtained was given to SVM for classification. The experiment was conducted in an ORL database containing 400 photos, both male and female. The kernel used for SVM is linear and the resulting rating is 98% compared to Radon Transform and Discrete Wavelet Transform.

## 3. Overview of the System

The face is one of the most important biometric aspects of humanity. Each person has their own innate facial appearance that are mostly different each other. Therefore, this system is investigated the gender classification through facial image using principal component analysis (PCA) and support vector machine (SVM).

The proposed system consists of three parts such as face detection, feature extraction from detected face by PCA and classification of male or female. The process flow diagram of this system is showed in Figure 1.

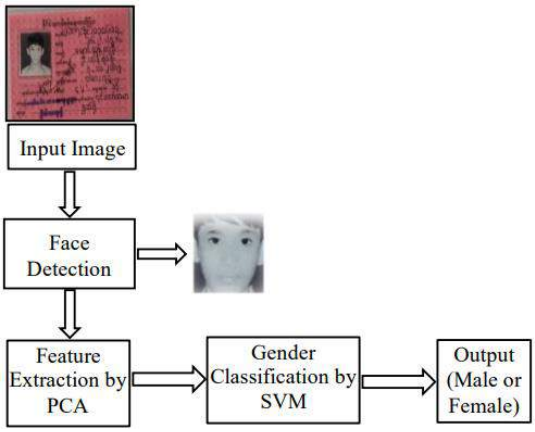



**Figure 1. Block diagram of the proposed system**

Initially, the facial area is detected using the Viola jones algorithm and then the detected face region is extracted from the NRC card. In the next step, the detected face region is subjected to Principal Component Analysis (PCA) to extract facial features. The principal components are exposed to classifier to classify the gender. Lastly, Support vector machine (SVM) classification techniques will be evaluated for gender classification. For the system to be properly examined, the system needs to be trained and tested using the face from Myanmar NRC card.

## 4. Proposed System

Face detection and gender identification play a vital role in video investigation for monitoring area and database management system. Gender classification by face from Myanmar NRC card is proposed in this system. This system is implemented by using its own Datasets. This process involves three stages. Firstly, Face detection (Viola-Jones algorithm) is used. Then PCA is applied to the face images to extract facial features. Finally, Support vector machine (SVM) classification techniques will be evaluated for gender classification. The proposed system is implemented on the MATLAB platform.

### 4.1. Face Detection

It is the process of identifying one or more human faces in images or videos. The Viola-Jones algorithm is used in this system to capture faces on NRC card images. Because of its high

detection rate, and its ability to run in real time. It is the most effective for frontal images of the face, and it can manage the rotation of 45 degrees for horizontal and vertical. This is due to its high intelligence and timely performance. The useful knowledge for facial features is the location and size of eyes and nose. And the darkness and brightness value are also important features.

#### 4.1.1 Viola-jones Object Detection Framework

The Viola-Jones algorithm was the first object detection framework proposed by Paul Viola and Michael Jones in 2001 to provide real-time competitive object search rates. The algorithm looks at several sub-regions and tries to find a face by searching for specific features in each region. Viola and Jones [4,5,6] used Haar-like features to detect faces in this algorithm.

#### 4.1.2 Haar like features

Haar like features [6] are digital images used in object recognition. It can be used to find the difference between black and light in an image. All human faces have unique properties. The eye area is deeper than its neighbors, and the nose area is brighter than the eye area. A simple way to find out which area is lighter or darker is to compile and compare the pixel values of the two areas. If one side is lighter than the other, it may be the tip of the eyebrow. Sometimes the center will be brighter than the surrounding boxes; It can be interpreted as a nose.

In Figure 2, the result of Viola Jones Face detection algorithm is shown

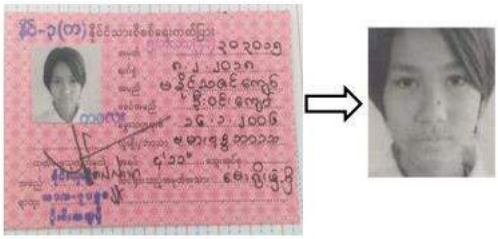


**Figure 2. Face Detection Result**

### 4.2. Features Extraction

The important thing to understand is that every face has a different pattern. The task of the face recognition system is to show that the test image belongs to a person in the database. Every image is random in nature because the lighting conditions, the orientation of eyes, facial features, hair and spectacles are different for different people. However, statistical characterization can still be carried out on this random set. They can be extracted from the original image using a mathematical tool called PCA [7,8,9].

#### 4.2.1 Principal Component Analysis

PCA is a dimensionality reduction technique, which is used to represent each image as a feature vector in a low dimensional subspace. Although the PCA is a traditional way of representing faces and it is the most widely used method in today. In this step, principal component analysis is applied to facial images to extract facial features. Flow chart of PCA Algorithm is as shown in the Figure 3.

Some features, such as set of eyes, mouth and nose, are presented with relative distances and they can help to distinguish the face. The characteristic of these features is called the principal components or the Eigen faces. Each principal component has a different robustness according to the amount of variance in its direction. One of the key features of PCA is that reconstructing any original image from the training set by integration with the Eigen faces [8,9]. Eigen faces are represented with the certain features of the original image. The Eigen faces on the Myanmar NRC card by PCA are shown in Figure 4.

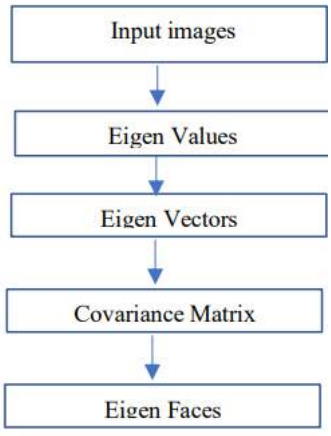



**Figure 3. Flow Chart of PCA Algorithm**

Step 1: The first step for the PCA is to get the data set.

Step 2: The next step of PCA is to calculate the mean of a given data set.

Step 3: The next step is to subtract the mean with each data of the data set, then, this produces a data set whose mean is zero.
Step 4: Then calculate the covariance of the matrix. If the data are two dimensional, then the covariance matrix will be of 2 × 2 and if the data is of N dimensional the covariance matrix will be of N × N.
Step 5: The next step is to calculate the eigenvalues and eigenvectors of the covariance matrix.
Step 6: The last step of PCA is to select the components and forming a feature vector.


**Figure 4. Eigen Faces**

## 5. Classification

Gender classification methods can enhance the performance of many other applications human recognition and video surveillance system. The classification method, Support vector machine (SVM) is used in this system. The facial features from feature extraction step are inputted to a support vector machine (SVM) classifier to classify male or female.

### 5.1. Support Vector Machine (SVM)

SVM provides a very high accuracy compared to other ratings, such as logistic regression and decision trees. It is used in many applications such as genetic classification and handwriting recognition. In this system, SVM [10,11,12] is used to identify the gender. The Eigen vector W, calculated from the feature extraction stage, is used at this classification stage. The weight values calculated from the training phase are used to train the classification algorithm and the weight values get from recognition phase are used as the input.

A Support Vector Machine (SVM) performs classification by constructing an N-dimensional hyper-plane that optimally separates the data into two categories [11]. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data-points of any class. In general, if the larger the margin, the generalization error of the classifier is lower. Without any knowledge of the mapping, the SVM can find the optimal hyper-plane by using the dot product functions in original space that are called kernels. The theory of SVM [12] is based on the idea of structural risk minimization. And special property of SVM is simultaneously minimized the empirical classification error and maximize the geometric margin.

## 6. Experimental Setup

Gender classification by face from Myanmar NRC card is proposed in this system as in figure 5. To study the system performance, the system must be trained using the face from the Myanmar NRC card. 80 images (males and females) of own dataset are used to analyze the system performance. Data collection is captured the front view of Myanmar NRC cards by using Sony camera. The images of database are under 30 years old and the images sized are 100 ×100 pixels. To evaluate the proposed system, SVM classification method is applied. The number of training images were 80 including 40 males and 40 female images. The number of different testing images were 52 including 27 male and 25 female images. Accuracy calculates the exact gender results. It measures the percentage of face images that were classified into correct gender, which accuracy was measured using formula below:

$$Accuracy\% = \frac{Number\ of\ correct\ classification}{Number\ of\ testing\ samples} \times 100$$

In the first experiment, male images, a database of 80 images of 40 people is used to evaluate the performance of proposed system. It has two images per person. In the second experiment, a database of 80 female facial images for 40 people is used to evaluate the gender classification rate of the system. It has two images for each woman. In each case, the number of female faces classified as male and the number of males classified as female are noted. The percentage accuracies for male and female face classification are shown in Table 4.1. These results again indicate that there is considerably more error in classifying female faces. In the case of male, around 4.4 % of the males are wrongly classified as females. In the case of female images, a larger error observed and around

15.00% of the female faces are classified as males.

**Table 4.1 shows the results of classification by SVM**

| Dataset | Accuracy | Error Rate (%) |
|---|---|---|
| Male (80 images for 40 people) | 95.6% | 4.4 % |
| Female (80 images for 40 people) | 85.0% | 15.0% |
| Female (100 images for 60 people) | 93.6% | 6.4% |
| Combined (male and female,180 images for 100 people) | 92.4% | 7.6% |

The experiment was also repeated on female images dataset by increasing the number of objects. Therefore, the number of objects with short hair faces is increased that contained 100 female images for 60 people in this experiment. In this case, the percentage accuracy for female was around 94%. In the next experiment, female and male (mix) dataset is used to study the performance of the proposed system. A database of 180 images for 100 people (40 males and 60 females) is used in this experiment. The percentage accuracies rate for male and female classification of the system is 92%.


**Figure 5. Gender Classification System**

In this experiment, the classification results of 'Male' type are also lower than the 'Female' type of gender.

**Advantages**: There is no need for direct contact with any like other biometric system such as finger print, voice and signature etc.
**Disadvantages**: Face recognitions are not able to perform well in the variation of illumination. Face recognition systems are not always accurate.
**System limitation**: In poor lighting, low resolution images does not work well.

## 7. Conclusion

This system is proposed the human gender classification by face images from Myanmar NRC card. The experimental evaluation of proposed system confirms the good performance. According to the experimental results, SVM classifier achieves as high as 89.06% gender classification accuracy for 80 subjects. For further experiments, the proposed system will be tested on the larger value of dataset and will be implemented the other machine learning algorithm such as KNN and Neural Network to compare with the proposed method.

## References

[1] Amit Jain, Jeffrey Huang, "Integrating Independent Components and Support Vector Machines for Gender Classification", ICPR, pp. 558-561, 2004.
[2] Baback, Mand Ming.H.Y Yang 'Gender Classification with Support Vector Machines' 'Proceedings of the 4th IEEE International Conference on Face and Gesture Recognition, March, 2000.
[3] Chai, T. Y., Rizon, "Facial Features for Template Matching Based Face Recognition". American Journal of Applied Sciences, vol. 6, no.11, pp. 1897-1901, 2009
[4] David Shaw,'Support Vector Machines for Classifying Face Data', Due:12/14/2009.
[5] D. Mohammad, A. Alqudah and O. Debeir, "Face Detection using Viola and Jones Method and Neural Networks" IEEE International Conference on Information and Communication Technology Research, pp 40-43,2015.
[6] G.Kavitha, I.Laurence Aroquiaraj "Face Detection and Gender Classification using Facial Features" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, November 2019.
[7] H. B. Kekre udeep, D. Thepade, C.Tejas , "Face and Gender Recognition Using Principal Component Analysis" International Journal on Computer Science and Engineering (IJCSE) Vol. 02, No. 04, 2010, 959- 964

[8] K.J. Diamantaras and S.Y. Kung, “Principal Component Neural Networks: Theory and applications”, John Wiley and Sons, Inc., 1996

[9] R. Sharma and M. Patterh (2015), “Age invariant face recognition using k-pca and k-nn on indian face age database,” International Journal of Computer applications, vol. 126, no. 5.

[10] S. Avinash, C. Mridul, K.S Deepak, C. Yogesh, “Gender Classification using Facial Embeddings: A Novel Approach” International Conference on computational Intelligence and Data Science (ICCIDS 2019).

[11] Smriti Tikoo, Nitin Malik “Detection of Face using Viola Jones and Recognition Using Back Propagation Neural Network”, Research gate, on 31 January 2017.

[12] Steven M. Holand, "Principal Component Analysis (PCA)", in Department of Geology, University of Georgia, Athens, 2008, GA 30602-2501.

# Face Recognition System using Principal Components Analysis and Back Propagation Neural Network

Su Sandar Win
*University of Computer Studies, Yangon*
*susandarwin750@gmail.com*

## Abstract

*Face Recognition System is a computer application which is applied to identify or verify a face image with the digital image type, based on digital image processing that is a popular area of research. The Face Recognition System uses to be efficient in criminal verification, data privacy, home video surveillance systems etc. Various innovative face recognition systems have been developed so far using a wide range of algorithms. The good method for face recognition using Principal Component Analysis and Back Propagation Neural Network is presented in this system. In this system feature extraction is used Principal Component Analysis (PCA) and then a Back Propagation Neural Network (BPNN) is trained to work as a classifier to gain the recognized facial image.*

**Keywords:** Face Recognition, Principal Component Analysis (PCA), Back Propagation Neural Network, Eigenvalue, Eigenfaces.

## 1. Introduction

The key research topic of this era in the area of computer science and biometric field is face recognition and this become getting more and more popular subject. It can take the content of many rules, and can be used widely. Specially it has wide and effective objective for the field of communication and social networking.

The fundamental rule for the face recognition system which is used to compare the face image of a person that is to be identified with the image saved in the database and it can produce the nearly similar the image of face as output. It can be comprehended the method of BPNN and PCA for useful and reliable recognization.

There are three main stages in this system such as Pre-Processing, Principal Component Analysis and Back Propagation Algorithm are presented.
The pre-processing stage can be applied for the two reasons as follow:
1) for interfering system, the face image is used for reduce noise and possible coil view.
2) for prove the obvious facts of facial features, it has to transform the image into a different space.
The reasons why using the Principal Component Analysis is as follows:
1. Reducing the data dimension to get more tractable limits.
2. Capturing the face image's specific features.
3. Eliminating the features redundant.
The outputs produced from Principal Component Analysis are feed to Back Propagation Neural Network which work as a recognizer to gain the verify the image of selected person.

The Back Propagation Neural Network is developed by using these four steps. For the 1st stage the weights of network must be randomly generated. Besides, an input signal release with each input unit (X) in feed forward operation and each input unit (X) is changed as each of the hidden units such as $Z_1$, $Z_2$ ...., $Z_n$. The output layer calculates with the activation function and the target output compare to the output. The mean square error is calculated and updated the bias and weights. For solving non-linear problems, it has to use the number of activation functions.

## 2. Related work

Face recognition system is has been developed for these two basic ways. The one way is used on extracting the vectors of facial features from the fundamental piece of the face such as eyes, nose, mouth and chin with the help of deformable templates and wide mathematics. The system creates the basic parts of the face using the gathered information as a feature vector. The

system shows the whole face image's information. Eigen face, which in terms of a best coordinate system represent as the individual face image. These are the average covariance of a whole of face's the eigen functions.

In [1] proposed, this system PCA and BPNN have worked well in the constrained and learned environment. And also it has found a good result for small orientation and low variance. PCA and BPNN is good and efficient for face recognition system. Now this will be developed for more complex and dynamic environment where noise, illumination, lighting etc. creates difficulties.

In [2] proposed, the result is compared with several techniques and proposed technique gives a better recognition rate then the other techniques. The Eigenfaces method is very sensitive to head orientations, and most of the distinct occur for the images with large head orientations.

In [3] proposed, this system unifies the capability of fuzzy set theory to obtain the degree of belonging of different pixels of a face image to different classes. Common vector method is obtained to reduce the number of samples used in training then traditional PCA has been used for recognition task.

## 3. Proposed Algorithm

For Face Recognition, the PCA and BPNN algorithms are presented in this system. The one popular method is the Principal Component Analysis (PCA) for the face recognition system. This method has been applied in the field of biometric identification and recognition. The obvious objective of PCA is to resize the large dimension of the image features to the small dimension of space of the data that are usually described the necessary information.

The working principal of the human brain is the basic idea of back propagation neural network (BPNN). and that algorithm carry out the suitable interconnection weights for a path of interconnected units. There exits three layers. The input layer gives the signal to the hidden layer nodes. The feature detector is worked by the hidden layer node with the function of activation and the final output is given by the output layer to the outside. The output unit and hidden units also have biases which is proved. The interconnection weights are connected for the input, hidden and output layers. For the hidden layer and output layer, which provide with the bias, for calculate the net input.

### 3.1. PCA Algorithm

1. Firstly, the training set take a set of M images and it must be supposed which every signal face owns the N*N dimensions. This transforms each face image to dimension $1*N_2$ image vector.

2. It has to subtracted the image vectors to the average face.

(1)

3. The following is the expression calculated of the average face from the above step,
(1) Where $G_1$, $G_2$, ……$G_M$ is a set of training images.

4. The average vector differs to each image.

(2)

5. After being subtracted the mean value, the system apply the eigen faces and M' eigenfaces are thought with the biggest of the eigen values. In cause of calculate the co-variance matrix, it produces the eigen faces with the following equation:

$$C = \frac{1}{M} \sum_{n=1}^{M} \phi_n \phi_n^T = A \cdot A^T \quad (3)$$

where the matrix $A = [\Phi_1, \Phi_2, \ldots\ldots\ldots, \Phi_M]$
where the matrix $A^T = [\Phi_1^T, \Phi_2^T, \ldots\ldots, \Phi_M^T]$

6. To gain a weight vector W of contributions of each and every eigen-faces to a image of the face, the image of the face is converted into its eigen-face facts projected onto the space of the face by a operation.

$$W_k = u^T_k \quad (4)$$

7. For the recognition, the number of eigen-faces of (k=1,.., *M'*, where *M'* £ *M)* is applied. A basis set for face images treate the eigen-faces, it is because of the weights vector form such as {W = $[W_1, W_2, \ldots, Wm]$} that shows the contribution of single eigenface in representing the face image. The best algorithm for deciding that the face supports the good description of an unknown face image input, is to calculate the face of image "*k"* that reduces the value of Euclidean Distance.

### 3.2. Euclidean distance

The calculating method for Euclidean distance between two values in two-dimensional space is

the measurement of a coordinate points joining the two values.
The calculating of two distance value are as follows in figure. When the value of points ($x_1$, $y_1$) and ($x_2$, $y_2$) be the two-dimensional space on the Euclidean Distance.

The square root of $(x_1 - x_2)^2 + (y_1 - y_2)^2$ (5)

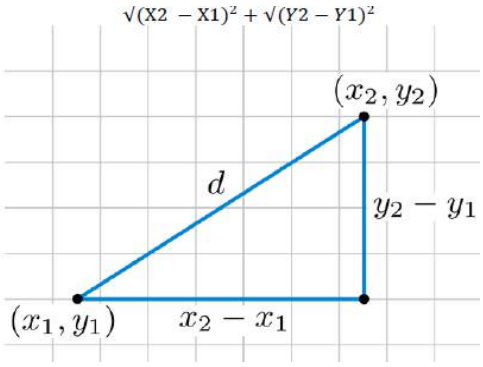



**Figure 1. Euclidean Distance Formula**

### 3.3. Artificial Neural Network

A artificial neural network is a set of connection with the input neurons and output neurons which every single relation has a related weight value with its formula. It helps to build predictive model. The method develops inspire with the human brain system. It supports to comprehend the operation image identification using neural network. This system purposes the back-propagation neural network (BPNN) in image recognition.

### 3.4. Back Propagation Neural Network

The back-propagation neural network (BPNN) is a widely used formula in the stage of training and testing including the feed forward formula of neural networks. In this system, Back Propagation Neural Network is created for the one-input layer, the two-hidden layer and the one-output layer. This system of input layer composes ten-neurons which are derived from the feature extraction method and the input layer is added bias neuron. The hidden layer consists of six neurons plus bias neuron because the network has only one-output neuron, so the number of neurons in the hidden-layer is two-thirds of the number of neurons in the input-layer. The output-layer consists of one neuron because the network is attempting to work as a classifier.

The back-propagation neural network-(BPNN) of the training stage, testing stage carry out the three stages of follows:

1. calculating the feed forward formula for the both of training stage, testing stage
2. calculating the back-propagation formula for the related error values
3. calculating the adjustment of weight value

Calculating the formula of the Feed-Forward, every input neuron gets an input value and delivers it to every hidden neuron, which in turn computes with the function and passes it on to each output neuron, which calculate again the function to gain the produced output.

Calculating training stage, testing stage, an appropriate error of the net output is calculated. The factor of error is gained which is applied to delivers the value of error back to the hidden-layer. The updated weights are calculated. Similarly, the factor of error is formulated for neurons. And then, after obtaining the error factors, the system must immediately calculate the weights. In this system, output-layer has one-neuron. For this system, there exists an eleven-input neurons, one-output neuron and seven - hidden neurons.
The Back-propagation Neural Network algorithm are as follows.

$$O_n = \frac{1}{1+ \exp [- (\sum_{xiwi} - t)} \quad (6)$$

where t represents bias neuron which value gives o or 1 in turn.

Determining Errors: $E = O_{actual} - O$ (7)

where $O_{actual}$ represents real output which value is 1 and O represents target output at the output layer.

$$E_n = O_n (1-O_n) \sum_j w_{nj} E_j \quad (8)$$

Finding new weights:

$$w^i_{jk} (new) = w^i_{jk} (old) + \alpha E_k^{i+1} x_{jk} \quad (9)$$

where α is the rate of coefficient for training which is limited to the value (from 0.01 to 1.0).

This calculated results through the above formula are produce as output which is recognize or not the face image person in the database.

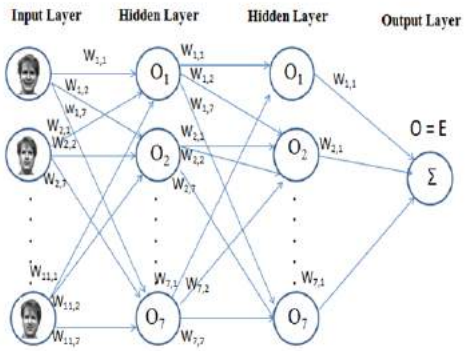



**Figure 2. Illustration of Back Propagation Neural Network**

## 4. System Design and Experimental Results

In this system includes two parts. The 1st stage is the taking the features of face image and the 2nd stage is the identification in order to compare with the feature values of stored in the database.

### 4.1. Proposed System design

This system shows the implementation design for the face identification system for android mobile phone platform. The system method combined two key stages which are stage of training and stage of testing.

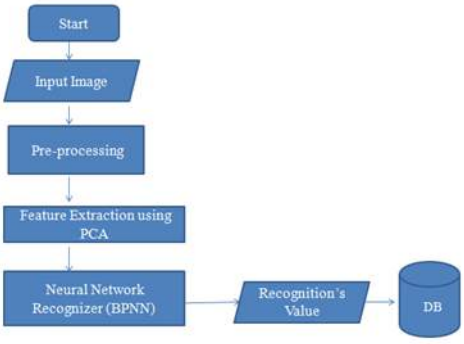



**Figure 3. The system flow diagram of the training stage**

In training stage, the face image is acquired from the capture or datasets in the "Input image" step. The face images are preprocessed and the features of this image is extracted using PCA algorithm and the output of "Feature Extraction using PCA" step is have to feed to the "Neural Network Recognizer (BPNN)" and then the value produce from the previous step which will be saved to the Database.

The training stage of purposed system is shown in Figure 3.

In Stage of testing, the face image is acquired from the capture or datasets in the "Input image" step. The face images is preprocessed and the features of this image is extracted using PCA algorithm and the output of "Feature Extraction using PCA" step is have to feed to the "Neural Network Recognizer (BPNN)" and then the value produce from the previous step will be compared with the value which is the output value saved in the database from the above training stage. If the compared two values are almost similar, the system show the message "Know user" and if not so, the system show the message "Unknown user".

The testing stage of purposed system is shown in Figure 4.

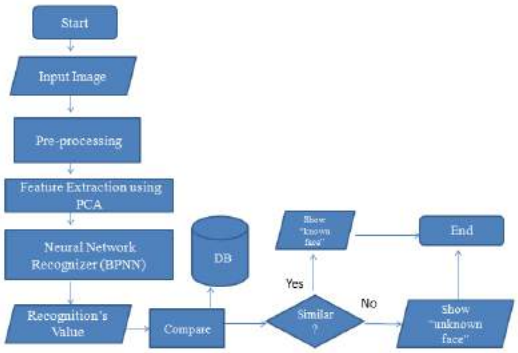



**Figure 4. The system flow diagram of the testing stage**

### 4.2. Dataset

The Dataset of Yale Faces (6.4MB) consists of 500-grayscale faces in the format of gif with 50 individual persons. This dataset includes 10 face for a person, for a one person with various expression of face and conditions: center side, front side, light view, side with glass, happy face view, left side, side with no glasses, normal side, right side, etc.

The example image of dataset is shown in Figure 5.

**Figure 5. The example of face images from Yale face dataset**

### 4.3. Experimental Results

In this system, the system was used 300 face images of thirty persons as training dataset and 200 face images of twenty persons as testing dataset which was used the yale face dataset.
The following are some examples of trained the face image for the system.

**Table 1. Face image in training stage**

| No | Face Image | PCA | BPNN |
|---|---|---|---|
| 1 |  | 0.62762 970101 | 0.89501 002615 |
| 2 |  | 0.09313 504702 | 0.35899 425358 |
| 3 |  | 0.71591 791742 | 0.24772 901110 |
| 4 |  | 0.87048 05470 | 0.70818 558489 |
| 5 |  | 0.90881 937941 | 0.56664 055793 |

According to the implementation of the system, the following are example of the recognition values produced from testing stage by using some face images which are trained in the above training stage.

**Table 2. Face image in testing stage**

| No | Face Image | PCA | BPNN |
|---|---|---|---|
| 1 |  | 0.90881 937941 | 0.56664 055793 |
| 2 |  | 0.42729 803702 | 0.41949 306559 |
| 3 |  | 0.0058 19048 | 0.8660 02507 |
| 4 |  | 0.4764 69389 | 0.1953 26625 |
| 5 |  | 0.24780 21051 | 0.29173 40777 |

## 5. Conclusion

This system develops for human face verification using the robust feature extraction method called PCA and efficient recognition method called BPNN.

The BPNN network algorithm is applied for verifying the side of front view or nearly front view and tabulate the necessary output. It will be able to expend with video image and the various kind of orientation.

## References

[1] A. K. Jain, B. Klare and P. Unsang, “Face recognition: Some challenges in forensics,” in Proc. IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, March 21-25, 2011, pp. 726-733.

[2] Dibberi, 4 Jan 2005, “Back propagation” ,https://en.Wikipedia.org/wiki/Back propagate ion, 20 September 2015.

[3] David Monzo, Alberto Albiol, Antonio Albiol, Jose M.Mossi, “A Comparative Study of facial landmark localization methods for Face Recognition using HOG descriptors”, Proceedings of IEEE 2010.

[4] Daijin Kim Sang-Ho Cho, Bong-Jin Jun2005. Face recognition on a mobile device. In Proceedings of International workshop of Intelligent Information Processing.

[5] Eugene Weinstein, Purdy Ho, Bernd Heisele,Tomaso Poggio, Ken Steele,and Anant Agarwal 2002, Handheld Face

Identification Technology in a Pervasive Computing Environment.

[6] Etisalat Coll. of Eng., Sharjah,United Arab Emirates 2005, A GPRS-based remote human face identification system for handheld devices.

[7] Joseph Lewis, University of Maryland, Bowie State University,January 2002,Biometrics for secure Identity Verification: Trends and Developments.

[8] Ms. Varsha Gupta and Mr. Dipesh Sharma, "A Study of Various Face Detection Methods", International Journal of Advanced Research in Computer and Communication Engineering), vol.3, no.5, May 2014.

[9] N. Revathy, T. Guhan, "Face recognition system using backpropagation artificial neural networks", International Journal of Advanced Engineering Technology, vol.3, no. 1, 2012.

[10] Prachi Agarwal, Naveen Prakash, "An Efficient Back PropagationNeural Network Based Face Recognition System Using Haar Wavelet Transform and PCA" International Journal of Computer Science and Mobile Computing, vol.2, no.5,pg.386 – 395, May 2013.

[11] Sheifali Gupta , O.P.Sahoo , Ajay Goel, Rupesh Gupta ,A2010 New Optimized Approach to Face Recognition Using EigenFaces.

[12].http://web.mit.edu/emeyers/www/face_ databases. html#yale

# Audio Steganalysis System based on Mutual Information Approach

Su Su Hlaing, Yawai Tint
*University of Computer Studies, Magway*
*susuhlaingucsy@gmail.com, yawai.ywt@gmail.com*

## Abstract

*Steganography is a technique whereby we put the existence of a message to question by simply covering it up within another file image or video. This system used the input audio signals as hidden message with audio signal which is embedded based on the stenographic tool and pure audio signal. The proposed system tested the steganalysis technique on audio signals embedded with "InvisibleSecrets" audio stenographic tools which are available from the Internet. The aim of the system is to develop the steganography detection system using the concept of data analysis approach. To evaluate the performance of proposed system, three types of audio signal are analyzed with difference number of bit stream. Overall detection accuracy for various types of music is around 94% and experimental results revealed that have the highest detection rate in audio steganalysis system.*

## 1. Introduction

Steganography is the method of embedding secret message in a conventional, non-secret, file or message in order to avoid detection; the secret data is then taken out at its target. The aim of steganography is to embed the hidden message within cover media such as audio, image and video file. Numerous steganography approaches and software have been broadly applied. Correspondingly, steganalysis techniques are technologically advanced to detect the existence of hidden information. Detection of steganography is the scientific technology to analyze if a medium carrier some hidden messages or not and if possible, to evaluate what data are embedded in the carrier files.

There are two approaches for analysis of steganography, which are technique specific steganalysis and universal steganalysis. The first one is to against the targeted steganographic technique which have good accuracy and second one is used for wide variety of technique. However, ever since universal steganalysis is well suitable to the practical setting, it attracted more awareness and many effective steganalyzers are proposed.

## 2. Proposed System for Steganography Detection

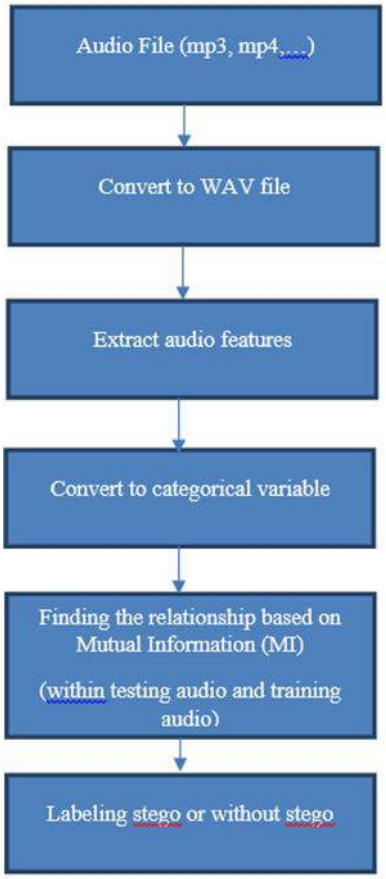



**Figure 1. Process Diagram of Proposed System**

This study extract the difference type of audio features from difference audio file, which are Mel-Frequency Ceptral Coefficient [1], Zero Crossing Rate, and Short Time Energy features. The extracted features are used to convert the dummy categorical variables for the next stage of

audio steganography detection. In the process of steganography detection, Mutual Information (MI) is utilized based on the extracted categorical variables of audio features. The proposed system tested the steganalysis technique on audio signals embedded with one of the audios steganographic tools which are available from the Internet. Steganographic tools will hide information in audio files. The data is compressed for encrypting the message within the audio bit stream.

There are many steganographic and steganalysis algorithm which are developed for growing attention on multimedia security [2]. The aim of the study is to apply one of the data analysis approaches in audio steganalysis. Bivariate mutual information is equal to zero if the two features are independent and there is no relationship each other.

## 3. Related Methodology

### 3.1. Feature Extraction

In the process of audio analysis, the selection of audio features and properties are emerged in important role. For audio detection and classification systems mostly used two processing stages: feature extraction followed by detection and classification. In this study, three types of features *"Mel-Frequency Cepstral Coefficient, Zero Crossing Rate, and Short Time Energy"* which are computed from MP3 signal.

#### 3.1.1. Mel-Frequency Cepstral Coefficient

Mel-frequency cepstral coefficients [1] are non-parametric representations of audio signal, which is the human auditory perception system. The term "mel" defines the measurement of the perceived frequency. The mapping between the frequency scale (Hz) and become aware of frequency scale (mels) is approximately linear below 1 kHz and logarithmic at higher frequencies. The following formula can evaluate the relationship.

$$\mathrm{F_{mel}} = 2595.\log_{10}\left(1+\frac{\mathrm{F_{Hz}}}{700}\right) \qquad (1)$$

where, $F_{mel}$ is the collected frequency in mels and $F_{Hz}$ is the frequency in Hz. The bandwidth and the spacing of these critical-band filters are invariable values, 300 mels and 150 mels within the mel-frequency domain. Y(n) indicate the power spectrum of an audio stream, F[k] is the power in k-$^{th}$ critical band and N represent the number of the critical bands in mel scale, ranging usually from 20 to 24. Then,

$$F[k] = \sum_{j=0}^{f/2-1} W_k(j).Y(j),\ k = 1, \ldots\ldots, N \qquad (2)$$

where, $W_k$ is the critical-band filter. The aspiration of MFCC is indicated by L and evaluate the MFCCs from logarithm and cosine transforms as follows

$$C[n] = \sum_{k=1}^{M} \log(S[k]) \cos\left[(k-0.5)\frac{n\pi}{M}\right],\ n = 1, \ldots L \qquad (3)$$

#### 3.1.2. Short-term energy:

In this steganography detection system, *short-term energy and zero-crossing rate (ZCR)* are used for the feature extraction step:

$$E = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n) \qquad (4)$$

#### 3.1.3. Short-term zero crossing rate (ZCR):

ZCR is based on the evaluation of time domain zero crossings [4]. If {x(0), x(1),……, x(N-1)} is the short term frame, then two features are given by

$$ZCR = \frac{1}{N}\sum_{n-1}^{N} \frac{|sgn\{x(n)\} - sgn\{x(n-1)\}|}{2} \qquad (5)$$

### 3.2. Dummy Coding

This coding is developed by Cohen and Cohen in 1983, which is the simplest coding structure that supposed to examine group mean differences. A dummy variable is a numerical variable for representing group behavior. For a categorical variable with multiple levels (n), n-1 numbers of dummy variables are required to represent it. Steganography detection system assumes that the independent features are numerical data. These are converted to the categorical variable for reducing the effect of redundancy within the extracted features of audio signals.

Dummy coding has the valuable advantages for representing categorical variable. This is suitable for nominal data and used to get dichotomous data. The most efficient one is "to

avoid a biased assessment of the impact of an independent variable, as a consequence of omitting another independent variable that is related to it". A second benefit to such a coding structure is its ease of interpretation.

The extracted features of MFCC, STE and ZCR are used to convert the categorical variable by using dummy coding for the analysis of mutual information.

### 3.3. Mutual Information Approach

Evaluating the "amount of information", i.e. the "loss of uncertainty" that give the comprehension of the other, and vice versa during the process of mutual information (MI) within two random features. Entropy is used to measure the concept of uncertainty between a random features. Not only the discrete variables but also continuous variables can be evaluated with entropy approach [3]. The entropy $H(Y)$ of a random feature Y with probability density function (pdf) $p_Y$ is defined by

$$H(Y) = -\int p_Y(y) log p_Y(y) dy \quad (6)$$

The entropy of a random feature $Y$ when the value of some other random feature X is known is the conditional entropy:

$$H(Y|X) = -\int p_X(x) \int p_Y(y|X = x) log p_Y$$

$$(y|X = x) dy dx \quad (7)$$

The mutual information is the difference between the entropy of a features and the conditional entropy.

$$I(X,Y) = H(Y) - H(Y|X). \quad (8)$$

Given two random featurs x and y, their mutual information is defined in terms of their probabilistic density functions $p(x), p(y) and p(x,y)$:

$$I(x;y) = \iint p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy \quad (9)$$

The extracted features $x_i$ are used, individually, to have the highest mutual information $I(x_i; c)$ with the target class $c$, reflecting the largest dependency on the target class during the process of max relevance approach.

## 4. Experimental Result

For the tested steganographic tool, two datasets are assembled: training and testing (Tr and Ts) which are downloaded from YouTube. Each data set contains 180 WAV audio signals of 10s length. All signals are sampled at 44.1 kHz. Each dataset contains 90 positive (stego) and 90 negative (cover) audio sample. All stego-audio signals are generated by hiding data from different text files.

The stego signals produced by invisible secret tool. This system are implemented with Matlab programming. The following figure show the user interface of steganalysis system.

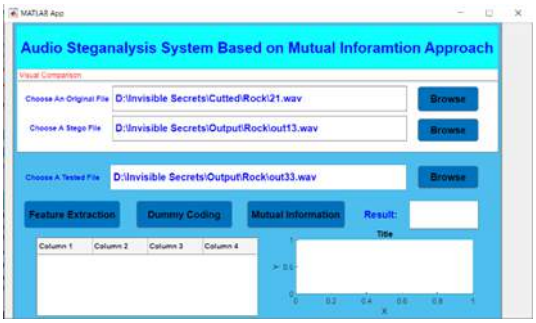



**Figure 2. User Interface for steganography detection system by using MATLAB**

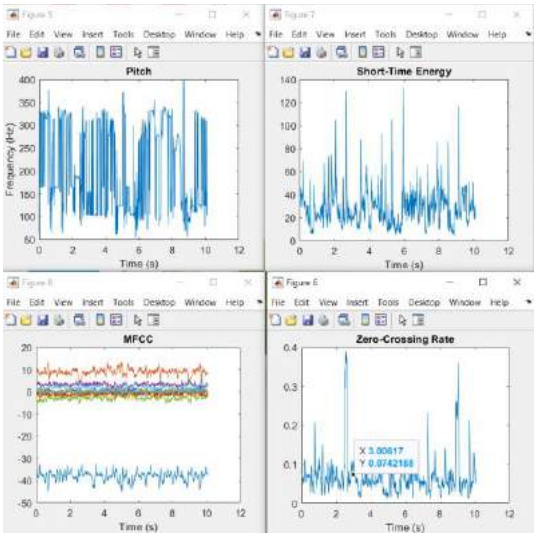



**Figure 3. Feature extraction for tested audio file**

Figure 3 describes the feature extraction process of the steganography detection system.

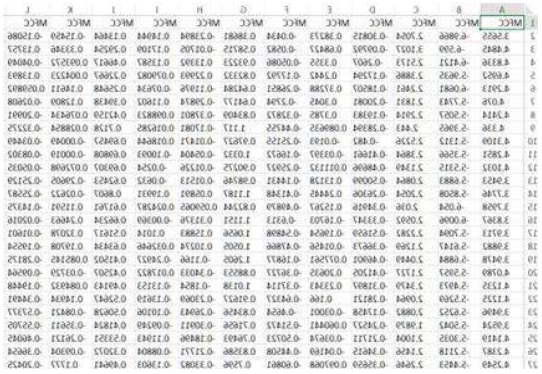


**Figure 4: Sample Extracted Audio Features**

The three type of features are extracted from the tested audio file which can be converted to dummy categorical variables by using dummy coding. This categorical variables are used for evaluating the mutual relationship. Table 1 describes the dummy coding results for converting extracted audio features to categorical variables.

**Table 1. Dummy Coding of Audio Features**

| MFCC | MFCC (Coding) | ZCR | ZCR (Coding) | STE | STE (Coding) |
|---|---|---|---|---|---|
| -1.33204 | 1 | 0.053951 | 3 | 0.000671 | 4 |
| -1.49417 | 1 | 0.05396 | 3 | 0.000389 | 2 |
| -1.64355 | 1 | 0.035514 | 2 | 0.000664 | 4 |
| -1.81382 | 1 | 0.036092 | 2 | 0.000247 | 2 |
| 1.573448 | 2 | 0.06811 | 4 | 0.000234 | 2 |
| 2.042204 | 3 | 0.066647 | 4 | 0.000178 | 1 |
| 2.279364 | 3 | 0.028541 | 2 | 0.000092 | 1 |
| 1.972043 | 2 | 0.026904 | 2 | 0.000102 | 1 |
| 1.668896 | 2 | 0.06135 | 4 | 0.000474 | 3 |
| 1.425379 | 2 | 0.0603 | 4 | 0.000139 | 1 |

### 4.1. Performance Analysis

Figure describes the detection accuracy under difference number of bits. In the stenographic tools, hidden message are embedded by difference encoding scheme so this cause to vary detection accuracy over difference bit numbers. Maximum number of bit are embedded in audio signal that is more accurate than signal with minimum number of bit. Invisible secret embeds only one bit of secret message in each frame in a consecutive manner and starting with the first frame of audio signal. Therefore, for a short message is used for data hiding scheme, the algorithm just use the few number of frames are affected by a variation in the quantization step.

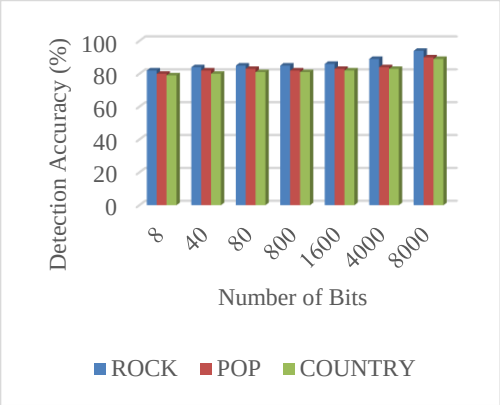



**Figure 5. Detection Accuracy with Different Number of bits**

In the analysis of performance, to evidence the effectiveness of this steganography detection scheme by using receiver operating characteristic (ROC) curve. Figure 4 describes the ROC curves as the detection threshold $^{Th}$ is varied.

The result revealed that large number of bit are hidden in cover audio file, true positive rate is nearly one and false positive rate is decreased.

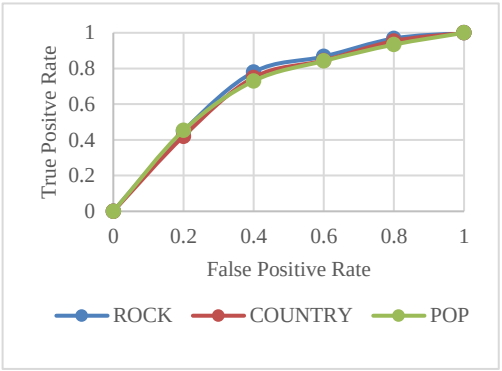



**Figure 6. ROC Analysis for different audio files**

## 5. Conclusion

This Audio steganalysis system based on Mutual Information approach has been applied in the system. This system showed the efficacy and efficiency in applying audio feature extraction based MI approach for analyzing audio

steganography detection. The steganography is promptly detected by MI approach and the performance of this system can be evaluated in different bit numbers with different audio files. The experimental result revealed that the proposed system is efficient for steganalysis of Invisible Secret tool and to improve the accuracy of the detectors. For analyzing the categorical data, Mutual Information can be employed very effectively and systematically. Dummy variables are used to make the resulting application easier to implement and distinguish in steganography detection system. For future work, the place of hidden message in audio signal by using the combination of different types of audio features with machine learning algorithm.

## References

[1] C. Kraetzer and J. Dittmann, "Mel- Cepstrum based steganalysis for voip stegography," Proc SPIE, vol 6505, p. 650505, 2017.

[2] N. F. Johnson, S. Jajodia. Exploring steganography: Seeing the unseen. IEEE Computer 31 (2), 2018, 26 – 34.

[3] Peng, H., Long, F., & Ding, C (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE
Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238.

[4] Q.Ding X. Ping "Steganalysis of Analysis-by-synthesis Compressed Speech" 978-0-7695-4258-4/10 $26.00 © 2010 IEEE DOI 10.1109/MINES.2010.148

[5] R. Chandramouli, M. Kharrazi, N. Memon. Image steganography and Steganalysis: Concepts and practices. International workshop on Digital Watermarking, 2019, 204 – 211.

[6] R, Krenn, "Steganography and Steganalysis", An article, January 2004.

[7] Rumsey D (in press), Statistic Essential for Dummies, Wiley publishing
Inc, 2010.

# Analysis of Edge Detectors for the VIN Number Area Segmentation

Khine Htoo, Amy Tun
*University of Computer Studies, Yangon, Myanmar*
*kohtookhine@gmail.com, amytun@ucsy.edu.mm*

## Abstract

*The quality of the edge detection work can make or break the sequencing process, good or bad. This works just emphasis on the VIN image. This step is the most important process to find the text area of an image which contains the VIN number. In this paper, experimental study is carried out to make analysis the performance of 3 famous edge detectors: canny, prewitt, and sobel operator. Various VIN images are tested with these operators and required decision can be made for choosing appropriate edge detector in the process of VIN number recognition process. Experimental result shows that Canny edge detection algorithm is better than the other two detectors in almost all the VIN images.*

.

**Keywords:** Edge detection, Prewitt method, Sobel method, Canny method, VIN image.

## 1. Introduction

Finding the best performance of edge detection algorithm to perform on various VIN image types is important task. The better quality of edge detection process, there will be the more robust text area detection process in the VIN image. The type of the vehicle is classified from the image of vehicle identification number (VIN). The vehicle identification number (VIN) is a unique code, including a serial number, used by the automotive industry to identify individual motor vehicles and VIN is sometime called car's chassis number.

After some preprocessing processes of a VIN image, edge detection is carried out to locate the possible text area. Mamta Joshi [2] made a comparison task among the Canny, Prewitt, and Sobel Edge detectors using various image types and concluded that Canny algorithm can find more reliable edges than the others. Here in this work, same detectors are analyzed to choose the edge detector in the VIN number identification system. Tamilselvi Nagasankar and B. Ankaryarkanni [3] did the performance analysis on the five commonly used edge detection algorithms: Prewitt, Sobel, Robert, Log and Canny using MATLAB tool with the various image types. They said that Canny edge detector has the better performance than the others. The main objective of this paper is to compare and analyze the performance among the various edge detection techniques. Analysis only on the one detector is also carried out using various parameters.

Both gradient based edge detector (Prewitt and Sobel) and Laplacian based edge detector (Canny) are analysis in this work to use in VIN number segmentation.

## 2. Edge Detection

An edge of the object in an image is the discontinuity of the brightness or intensity.[4] Edge detection is the technique used to identify the regions in the image where the brightness of the image changes sharply. This sharp change in the intensity value is generally observed or estimated using the first-order derivative finding maximum and minimum values [1].

So many comparisons have been done between commonly used edge detectors like Sobel, Canny, Prewitt, Roberts, Laplacian and Zero Crossing. In this work, comparison between 3 edge detectors is emphasized.

### 2.1. Prewitt Edge Detector

The Prewitt operator was developed by Judith M. S. Prewitt. It detects both horizontal (along the x axis) and vertical (along the y axis) edges. Steps of Prewitt operator are as follow:

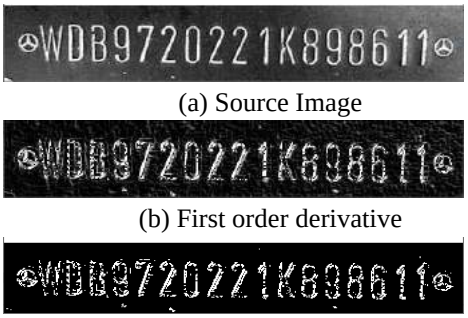
- Read the image.
- Change to gray image.

- Create 2 dimensions byte and double array.
- Define the mask in desired size.
- Extract the edges along X-axis.
- Extract the edges along Y-axis.
- Sum these two detected edges.
- Display the image.

Prewitt operator uses two masks for horizontal and vertical direction.

$$[\text{X-axis}] = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}, [\text{Y-axis}] = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Example detected image is shown in the following Figure 1.



(a) Source Image

(b) First order derivative

(c) Binary Image

**Figure 1. Result of Prewitt Operator**

## 2.2. Sobel Edge Detector

Similar to the Prewitt operator, Sobel operator uses two kind masks to detect an edge in an image. It also detects both horizontal (along the x axis) and vertical (along the y axis) edges. To calculate the first order derivative, following derivative masks are used for horizontal and vertical direction:

$$[\text{X-axis}] = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, [\text{Y-axis}] = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}.$$

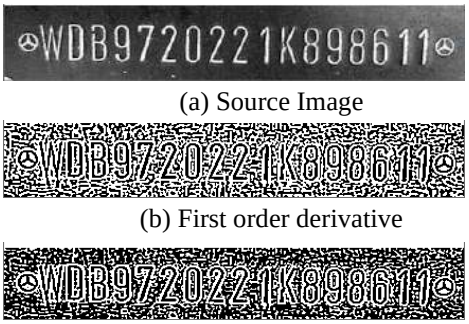
If we define $\boldsymbol{A}$ as the source image, and $\boldsymbol{G}_x$ and $\boldsymbol{G}_y$ are two images which at each point contain the horizontal and vertical derivative approximations respectively, the computations are as follows equation:

$$\boldsymbol{G}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A, \boldsymbol{G}_y \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \quad (1)$$

At each point in the image, the resulting gradient approximations can be combined to give the gradient magnitude, using:

$$G = \sqrt{\boldsymbol{G}_x^2 + \boldsymbol{G}_y^2} \quad (2)$$

Example result of Sobel is shown in the following Figure 2.



(a) Source Image

(b) First order derivative

(c) Binary Image

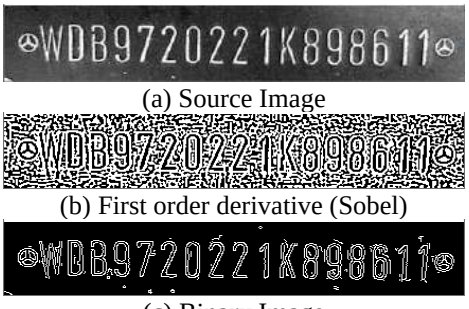
**Figure 2. Result of Sobel Operator**

## 2.3. Canny Edge Detector

The canny edge detection first removes noise from image by smoothening. It then finds the image gradient to highlight regions with high spatial derivatives. The algorithm then tracks along these regions and suppresses any pixel that is not at the maximum (non-maximum suppression). Algorithm of canny detector has 4 major steps as follows:

- Reduce Noise using Gaussian Smoothing.
- Compute image gradient using Sobel filter.
- Apply Non-Max Suppression or local maxima
- Finally, apply Hysteresis thresholding which that 2 threshold values (upper and lower).

Example result of Sobel is shown in the following Figure 3.



(a) Source Image

(b) First order derivative (Sobel)

(c) Binary Image

**Figure 3. Result of Canny Operator**

## 3. Analysis of Each Detector

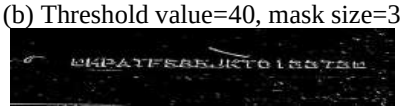
In this section, each detector is analyzed using various parameters: such as matrix size, threshold value, etc.

### 3.1. Prewitt Parameters

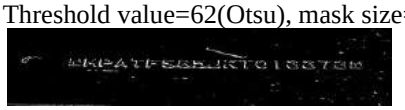
Likewise processing steps of Sobel detector, first order derivatives $G_x$ and $G_y$ are calculated using Prewitt masks mentioned in the section 2.1. Analysis on various threshold values is also carried out for the Prewitt edge detectors. Example result is shown in the following Figure 4.

(a) Source Image

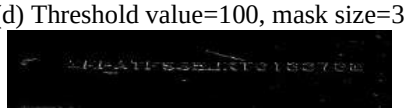
(b) Threshold value=40, mask size=3



(c) Threshold value=62(Otsu), mask size=3



(d) Threshold value=100, mask size=3



(e) Threshold value=150, mask size=3

**Figure 4. Result of Prewitt Operator**

Result shows that Otsu threshold value is good enough to use for VIN images.

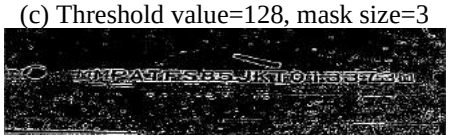
### 3.2. Sobel Parameters

Parameters of mask size 3,5,7,9, etc. can be used in Sobel edge detection process. On the first order derivatives output image is calculated from $G_x$ and $G_y$ using Eq. 1. And various threshold values can be applied to get binary output image.

(a) Source Image

(b) Threshold value=64, mask size=3

(c) Threshold value=128, mask size=3



(d) Threshold value=128, mask size=5

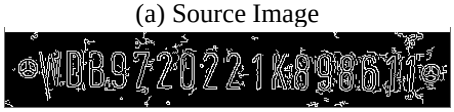
**Figure 5. Result of Sobel Operator**

The greater mask size, the better the output image quality. Otsu global threshold value can also be used to get the binary image.

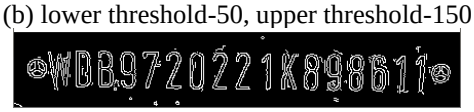
### 3.3. Canny Parameters

Performance of Canny edge detector on various parameters can be clearly seen in the following figure:

(a) Source Image



(b) lower threshold-50, upper threshold-150



(c) lower threshold-100, upper threshold-200

(d) lower threshold-200, upper threshold-250

**Figure 6. Result of Canny Operator**

Results shown in Figure 6, it says that lower threshold value should be between 150 and 200.

## 4. Analysis on Three Each Detectors

For the best result of character segmentation, text area extraction is also important process to get the real VIN number area. Finally, choosing a reliable edge detection method merely leads to the good VIN text area extraction. In this work, performance analysis is examined to choose the best edge detector for text area extraction on both good and bad quality VIN images using Mean Squared Error (MSE).

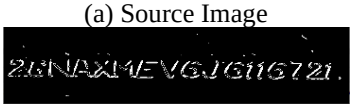
The MSE represents the average of the squares of the "errors" between resulted edge image and ground truth image. This is the amount differences between two images. The smaller

difference or error between two edges of image, it is the better accuracy of edge detector. The following analysises show the results of MSE values.

### 4.1. Analysis on Normal Quality Images

MSE results in edge detection for normal quality images are shown in Figure 7.

(a) Source Image



(b) Prewitt (MSE - 19.126)

(c) Sobel (MSE - 19.168)

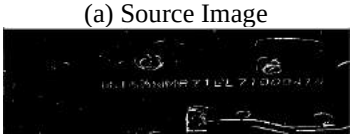
(d) Canny (MSE - 16.234)

**Figure 7. Result of edge detectors**

### 4.2. Analysis on Bad Quality Images

Edge detector analysis for the bad quality VIN image is also carried out to choose the best edge detector. Figure 8 shows the experimental results of various edge detection methods on the bad quality VIN image. As the result, Prewitt and Sobel edge detectors give the rather unpleasant output than Canny Edge Detector for the bad quality and non-uniform brightness image. The quality analysis of edge detectors shows that the best choice is “Canny Edge Detector”. MSE values for 3 detectors are shown in the following test works.

(a) Source Image



(b) Prewitt (MSE - 13.985)

(c) Sobel (MSE - 13.752)

(d) Canny (MSE - 4.801)

**Figure 8. Result of edge detectors**

## 5. Conclusion

According to the results of section 3 and section 4, it can be obviously seen that canny edge detection algorithm which has minimum MSE value can find more accurate text boundary in an image. Prewitt and Sobel edge detectors can also find edge of all the objects in the image. However, to extract VIN text area, dilation process is carried out after edge detection step and, in this case, result of Prewitt and Sobel detectors cannot be used to differentiate between text area and other object area. Although canny is better for choosing, its complicity and understandable stage of algorithm should be considered.

## References

[1] Chinu and Chhabra, “Overview and Comparative Analysis of Edge Detection Techniqurs in Digital Image Processing”, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 10 (2014), pp. 973-980.

[2] Mamta Joshi, “Comparison of Canny edge detector with Sobel and Prewitt edge detector using different image formats”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, ETRASCT' 14 Conference Proceedings.

[3]Tamilselvi Nagasankar, B. Ankaryarkanni, Performance Analysis of Edge Detection Algorithms on Various Image Types, Indian Journal of Science and Technology, Vol 9(21), June 2016.

[4] Tamar Peli, David Malah, “A study of edge detection algorithms”, Computer Graphics and Image Processing”, Volume 20, Issue 1, September 1982, Pages 1-21.

# Soil Classification for Agriculture Crops using K-Nearest Neighbors (KNN)

Shwe Yee Win, Thin Lai Lai Thein
*University of Computer Studies, Yangon*
*suuzaanna@gmail.com, tllthein @ucsy.edu.mm*

## Abstract

*Soil is the basis of our earth's agroecosystems which provide us with fiber, food, fuel etc. Soil classification helps predict soil type and performance for growing agricultural crops that provide us with food. Soil classification is essential for a farmer can know soil type, and plants the suitable crops depending soil type. The aim of this research is therefore to develop a method that automates soil classification by applying image processing techniques. In the proposed soil classification method, soil classification is performed by using color and texture of a soil image as features and by using the K-Nearest Neighbors (KNN) as a classifier. After classification, the system provides the list of crops and vegetation which can easily be grown in the predicted soil type. Soil RGB images dataset applied to our soil classification system contains "clay", "clay loam" and "sandy loam" (Red Earths and Yellow Earths) soil images taken in plantations and farms in Lashio township in Myanmar, and collected from Internet. The overall accuracy of the proposed method is about 88% for all 3 soil types: clay, clay loam and sandy loam.*

**Keywords**: soil classification, color and texture features, K-Nearest Neighbors (KNN), machine learning (ML), artificial intelligence (AI)

## 1. Introduction

Soil is an important ingredient of agriculture, and many kinds of soil exist. Each soil type can have various kinds of features and crops grown on different soil types. The features and characteristics of different types of soils are needed to find out so as to know which crops can grow healthier in the specific soil types [1].

Soil classification comprises steps like image acquisition, feature extraction, image preprocessing, and classification. Statistical features such as HSV histogram, standard deviation and, mean, and low-pass filter, Gabor filter, and color quantization [2], pH, Zinc and Potassium, chemical features [1], and color features, texture features, drainage class features, and terrain features [3], can be applied as features to soil classification.

Machine learning (ML) algorithms can be useful to soil classification because it is progressed significantly in recent years. ML is still a challenging and emerging research field in agricultural data analysis. Several ML methods such as Support Vector Machines (SVM), Bagged Trees, weighted K-Nearest Neighbor (k-NN), Gaussian Kernel-based [1], K-means clustering and Self-organizing Maps (SOM) [3] are applied to soil classification.

The organization of paper is as follows. Section 1 introduces the importance of soil classification, steps and machine learning methods needed to implement it. The related works are discussed in section 2. Section 3 describes the methods for the proposed soil classification system. Section 4 presents the design of the proposed system. The experiments and results of the proposed soil classification system is discussed. Conclusion is described in the last section, section 6.

## 2. Related Works

With the advent of artificial intelligence (AI) and ML, various methods have already been proposed to facilitate soil classification task in literature. The previous research works related to this research in literature are discussed and reviewed in this section.

In [1], S. A. Z. Rahman and K. Chandra Mitra have proposed a model or a method that can predict soil series together with land types, and according to prediction results, their model can suggest suitable crops for a certain soil type.

Gaussian kernel-based Support Vector Machines (SVM) and the weighted KNN, and Bagged Trees are applied to soil classification in their research. The proposed SVM based method achieves better results than many existing classification methods according to the experimental results. The remark for their research is that prior survey on soil series of a country, chemical and geographical features are must be performed.

In [3], Sofianita Mutalib and S. Abdul-Rahman have used k-means and self-organizing map (SOM) in the classification model. The inputs to their model are texture, color, terrain and drainage class. The classification rate for the SOM is 91.8%, and k–means, 79.8%, respectively. The remark for their research is that prior survey on soil texture, color, terrain and drainage class features are must be first measured, and the classification rate of the model based on k–means is about 79.8 %.

In [2], Srunitha. K. and S. Padmavathi, have explained soil types classification using support vector machine. The texture features from soil images are taken using color quantization technique, Gabor filter and the low pass filter. The statistical parameters: Standard deviation, mean amplitude and HSV histogram, are also extracted. The remark for their research is that, although the proposed method works effectively with sand and clay soil types, it provides poor results for the peat soil with 58.7% accuracy.

## 3. Methods applied to the Proposed Soil Classification System

Methods used in the proposed soil classification system are presented in this section. In features extraction method, color features: mean and standard deviation, and texture features: energy and contrast are extracted from soil RGB images. In soil classification method, features similarity between the features dataset and the tested features dataset is calculated using KNN.

### 3.1. Features Extraction

Features extraction can be used in many different domains such as diagnosis, identification, clustering, classification, detection and recognition. Image features extraction is used to get much information as feasible from the image. There exist many feature extraction methods, which may depend on color features, geometric features, statistical features, and texture features [4]. Feature is crucial in image processing. The different features of an image are domain specific features, or shape, texture and color [5]. Features extraction is to simplify resource numbers needed to represent a large dataset accurately [6].

#### 3.1.1. Texture Features

Texture is assumed as one of the most important features extracted from any image. The second order statistical features from any image are achieved by using a gray level co-occurrence matrix (GLCM). Some of the Haralick texture features are contrast, energy, homogeneity, entropy and correlation [5], and they are summarized and calculated as follows: [7]
Contrast is the measure of the local variations in the gray-level co-occurrence matrix.

$$Contrast = \sum P_{i,j}(i-j)^2$$

Entropy is the quantity of energy.

$$Entropy = \sum P_{i,j}\ x\ logP_{i,j}$$

Homogeneity computes the not-zero in the GLCM, and it is the inverse of the weight of the contrast. The value of the homogeneity is ranging from 0 to 1.

$$Homogeneity = \sum \frac{P_{i,j}}{1+(i-j)^2}$$

Energy computes the local homogeneity, and the value of the energy is ranging from 0 to 1.

$$Energy = \sum (P_{i,j})^2$$

Where: P = Normalized GLCM, i = row and j = column.

In this research, energy and contrast of a normalized GLCM transformed from a soil image are used as texture features in features extraction of the system.

#### 3.1.2. Color Features

Greyscale texture features are popular in image processing domain, and provide enough information to solve many different tasks. Many

researchers, however, have started to take color information and features into consideration for the human eye perceives any image as a combination of shape, color and texture [8]. Therefore, the combination of texture and color features is applied to our proposed soil classification method.

Color moment measures the color distribution in any image. Two types of color moments or features: mean and standard deviation, are used as color features in our proposed method.

The first color feature or moment, mean, can be represented as the average or mean color in the image, and it can be calculated as follows:

$$\mu = \sqrt{\frac{\sum_{i=1}^{n} x_i}{N}}$$

Where: μ =mean, $N$ = the number of pixels in the image, and $x_i$ = value of the $i^{th}$ pixel in the image.

The second color feature or moment is the standard deviation, and it can be calculated as follows:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Where: $\sigma$ = standard deviation, $N$ = the number of pixels in the image, μ =mean, and $x_i$= value of the $i^{th}$ pixel in the image.

### 3.1.3. Color and Texture Features Extraction in the Proposed System

The color and texture features of soil RGB images are extracted in the system by using the above color and texture features methods, and these features are as shown in Figure 1.

| Image Name | Image | Features | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RM | RS | GM | GS | BM | BS | RE | RC | GE | GC | BE | BC |
| Clay |  | 174.67 | 53.44 | 85.28 | 35.32 | 62.12 | 28.89 | 0.10 | 0.52 | 0.14 | 0.43 | 0.17 | 0.40 |
| Clay Loam |  | 89.06 | 51.04 | 83.04 | 44.35 | 79.71 | 42.49 | 0.05 | 1.13 | 0.06 | 1.11 | 0.06 | 1.10 |
| Sandy Loam |  | 167.25 | 47.01 | 122.33 | 42.83 | 87.98 | 38.59 | 0.06 | 1.07 | 0.06 | 1.06 | 0.07 | 0.99 |
| Test |  | 180.30 | 42.55 | 93.30 | 30.07 | 66.38 | 25.62 | 0.09 | 0.68 | 0.12 | 0.60 | 0.15 | 0.56 |

R=Red, G=Green, B=Blue, M=Mean, S=Standard Deviation, E=Energy, C=Contrast

**Figure 1. Color and Texture Features from Soil RGB Images**

As depicted in Figure 1, there are 12 features for a soil RGB image and the range of values of 12 features is different from each other. The different in range of features can lead to a challenging classification task for KNN. The solution to this challenging task is to normalize features values by using the normalization technique used in the following equation:

$$V_n = \frac{V_c - V_{min}}{V_{max}}$$

Where: $V_n$ = normalized value, $V_c$ = current value, $V_{min}$ = minimum value, and $V_{max}$ = maximum value. Normalization makes the values of all the features ranging from 0 to 1. The normalized features are as shown in Figure 2.

| Image Name | Image | Normalized Features | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RM | RS | GM | GS | BM | BS | RE | RC | GE | GC | BE | BC |
| Clay |  | 0.60 | 0.45 | 0.31 | 0.26 | 0.29 | 0.20 | 0.10 | 0.17 | 0.20 | 0.13 | 0.20 | 0.12 |
| Clay Loam |  | 0.23 | 0.43 | 0.30 | 0.35 | 0.42 | 0.34 | 0.02 | 0.43 | 0.04 | 0.43 | 0.03 | 0.47 |
| Sandy Loam |  | 0.57 | 0.39 | 0.53 | 0.33 | 0.48 | 0.30 | 0.03 | 0.40 | 0.05 | 0.41 | 0.04 | 0.42 |
| Test |  | 0.62 | 0.34 | 0.36 | 0.20 | 0.32 | 0.16 | 0.09 | 0.24 | 0.16 | 0.21 | 0.17 | 0.20 |

R=Red, G=Green, B=Blue, M=Mean, S=Standard Deviation, E=Energy, C=Contrast

**Figure 2. Normalized Color and Texture Features from Soil RGB Images**

## 3.2. Soil Classification

The above normalized features are built as a features dataset, and they are used in comparing with the normalized features of a tested soil image by using KNN for soil classification. The KNN algorithm is one of the most widely used machine learning algorithms for classification due to its simplicity and easy implementation. In many domain problems, it is also applied as the baseline classifier.

The distance between two points is calculated by Euclidean distance, a similarity measure, in KNN. The distance between two points: p and q with n elements is measured as in the following equation:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

After measuring the distance between the testing data and the training data using the above equation, the k number of nearest neighbors to the testing data is then selected, and the majority class or label of the selected neighbors will become the predicted class or label for unknown testing data.

As an example of our proposed soil classification, normalized features shown in Figure 2 are used to classify a tested input soil image using the following steps:

Step 1: Find the distance (d1) between clay image and test image.

$d_1 = \sqrt{(0.60 - 0.62)^2 + \cdots + (0.12 - 0.20)^2} = 0.2035$

Step 2: Find the distance (d2) between clay loam image and test image.

$d_2 = \sqrt{(0.23 - 0.62)^2 + \cdots + (0.47 - 0.20)^2} = 0.6519$

Step 3: Find the distance (d3) between sandy loam image and test image.

$d_3 = \sqrt{(0.57 - 0.62)^2 + \cdots + (0.42 - 0.20)^2} = 0.4925$

Step 4: Sort the distances in ascending order.

$$distances = [0.2035, 0.4925, 0.6519]$$

Step 5: Find the nearest distance(s) based on the number of nearest neighbor value, k. If k = 1, the nearest neighbor or distance is,

$$nearest\ distance = 0.2035.$$

Step 6: The nearest distance, 0.2035, is the distance between test image and clay image. According to the nearest distance, the test image in Figure 2 is classified as a clay soil image by the proposed system.

## 4. Proposed System

The overview design diagram of the system is shown in Figure 3. The proposed system is implemented as the soil classification system by using KNN. The main point of the proposed system is color and texture features extraction and soil classification based on these extracted features using KNN as a classifier. In this system, there are three main steps.

In the first step, pre-preprocessing, which consists of acquiring soil images and image resizing, is performed. The soil images are resized as size of 250x250x3. This system accepts various image types such as Portable Network Graphics, PNG (.png), Joint Photographic Experts Groups, JPEG (.jpg, .jpeg), Tagged Image File Format, TIFF (.tif, .tiff) and JPEG File Interchange Format, JFIF (.jfif).

In the second step, extraction of color and texture features from all soil images is done, and the features dataset is built using the extracted features vectors as described in Figure 1 and 2.

In the final step, a tested soil image is inserted into the system and the features from this image are also extracted. The tested soil image is classified based on distances between the features vector of the tested soil image and the features vectors in the features dataset by using KNN as a classifier.

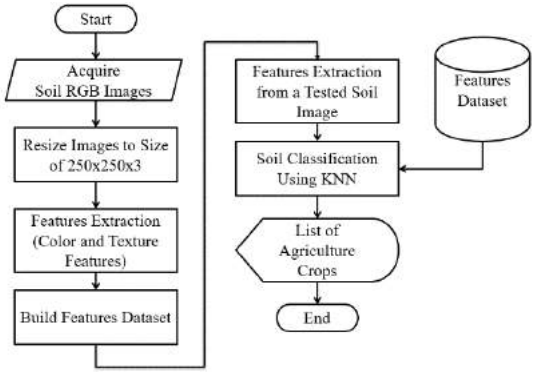



**Figure 3. Overview Design of the Proposed Soil Classification System**

After soil classification, the system also provides a list of agriculture crops that can grow easily on the soil in the tested image.

## 5. Experiments and Results

In Figure 4, soil classification on a tested image is performed.


**Figure 4. Soil Classification on a Tested Soil Image**

The features from the tested image are also needed to extract and they are normalized, and soil classification is performed based on normalized features vector of the tested image and features vectors in the normalized features dataset by using KNN. Moreover, a list of agriculture crops that is suitable for the output soil type is also shown in Figure 4.

The average processing time for soil classification on 75 soil images is about 0.32 millisecond.

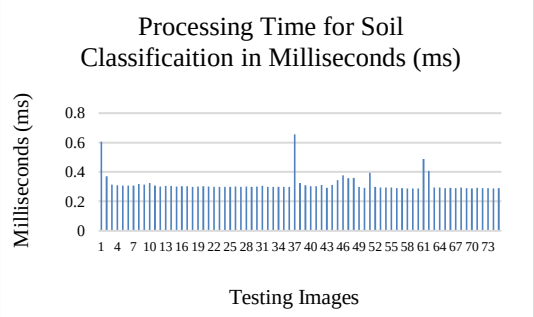



**Figure 5. Processing Time for Soil Classification**

The processing time for soil classification on each soil image is as illustrated in Figure 5. However, the processing time for soil classification can also change depending on the size of the features dataset because KNN, a lazy learner, is applied to the system as a classifier. The time complexity of the proposed soil classification can be denoted as $O(nd+k)$ because of KNN, in which n is feature vector numbers in the features dataset, d is the dimension of features in a features vector, and k is the number of nearest neighbors. The processing time can also vary depending on the software and hardware specifications of a computer.

To evaluate the accuracy of the system, 75 soil images including 30 clay soil images, 22 clay loam soil images and 23 sandy loam soil images are tested in the system. As a result, number of correct tests is 66, and number of misclassified tests is 9 and number of all tests is 75. Therefore, soil classification accuracy is 88 % and the percent error is 12 %. The soil classification accuracy of the system is illustrated using confusion matrix as shown in Figure 6.

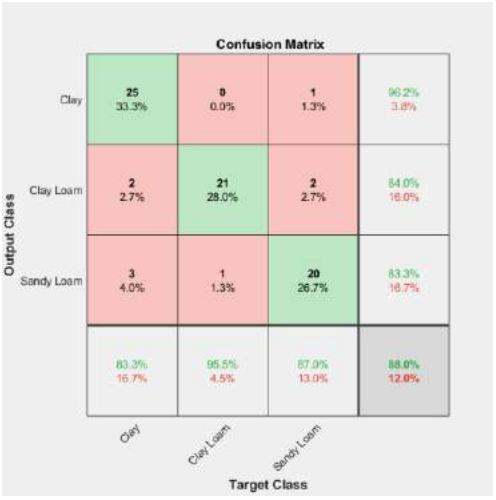



**Figure 6. Confusion Matrix depicting Soil Classification Accuracy**

The soil classification accuracy of the proposed system based on KNN is compared with other two classifiers: Decision (Fine) Trees and Kernel Naïve Bayes. The normalized features dataset is first trained using the other two classifiers, and the normalized tested features dataset is then used to evaluate their accuracies. These processes are conducted by using the Classification Learner App of the MATLAB. The accuracy of the KNN classifier is 86.7 % in the Classification Learner App. The accuracy of the Decision (Fine) Trees classifier is 78.7 % in the Classification Learner App. The accuracy of the Gaussian Naïve Bayes classifier is 76 % in the Classification Learner App. The accuracy comparison among the proposed system and the other two classifiers is depicted in Figure 7.

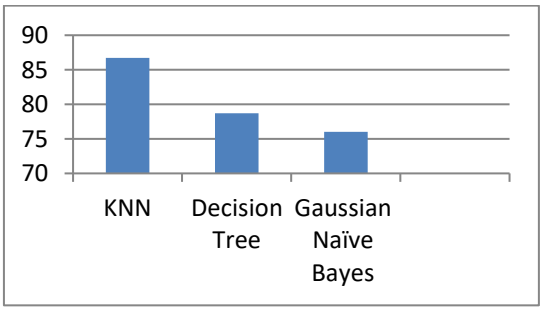



**Figure 7. Accuracy Comparison with Other Classifiers**

## 6. Conclusion

Traditional soil classification methods in laboratory or in-situ are expensive, experts- and labor-intensive, and time-consuming. With the rapid advancement of computer technology, many researchers in the field of AI have been

trying to automate soil classification in order to reduce human efforts as much as possible. This thesis has also proposed a method and implemented a system that can efficiently and automatically classify soil types using trending computer technologies such as machine learning and image processing. In addition, the system can provide farmers a list of agriculture crops that can easily grow in their farms and plantations.

## References

[1] S. A. Z. Rahman, K. Chandra Mitra and S. M. Mohidul Islam, "Soil Classification Using Machine Learning Methods and Crop Suggestion Based on Soil Series," 2018 21st International Conference of Computer and Information Technology (ICCIT), IEEE, 2018, pp. 1-4, DOI: 10.1109/ ICCITECHN. 2018.8631943.

[2] Srunitha. K, S. Padmavathi, "Performance of SVM Classifier for Image based Soil Classification", International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016, IEEE, ISBN: 978-1-5090-4620-1.

[3] Sofianita Mutalib, S. Abdul-Rahman, "Soil Classification: An Application of Self Organizing Map and k-means", 2010 10th International Conference on Intelligent Systems Design and Applications, Malaysia, 2010.

[4] Wamidh K. Mutlag, Shaker K. Ali, Zahoor M. Aydam, Bahaa H.Taher, "Feature Extraction Methods: A Review", Journal of Physics: Conference Series, 2020, DOI:10.1088/1742-6596/1591/1/012028.

[5] J. C. Kavitha and A. Suruliandi, "Texture and color feature extraction for classification of melanoma using SVM", 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), IEEE, 2016, pp. 1-6, DOI: 10.1109/ ICCTIDE. 2016. 7725347.

[6] P. Mohanaiah, P. Sathyanarayana, L. Guru Kumar, "Image Texture Feature Extraction using GLCM Approach", International Journal of Scientific and Research Publications, Volume 3, Issue 5, ISSN 2250-3153, May 2013.

[7] Sundos Abdulameer Alazawi, Narjis Mezaal Shati, Amel H. Abbas, "Texture features extraction based on GLCM for face retrieval system", Periodicals of Engineering and Natural Sciences, Volume 7, No. 3, pp. 1459-1467, October 2019, ISSN 2303-4521.

[8] Miroslav Benco, Robert Hudec, Patrik Kamen cay, Martina Zachariasova, Slavomir Matsuka, "An Advanced Approach to Extraction of Color Texture Features based on GLCM", 2014 International Journal of Advanced Robotic Systems, 2014, DOI: 10.5772/58692.

# Myanmar Road Sign Recognition System using Convolutional Neural Network and Support Vector Machine

Tin Zar Htun, Aung Nway Oo
*University of Information Technology, Yangon, Myanmar*
*tinzarhtun@uit.edu.mm, aungnwayoo@uit.edu.mm*

## Abstract

*Modern advanced automated driving systems' most vital role is to identify road signs since it improves drivers' safety and comfort. But, with the diversity of road traffic patterns, it is still a challenging task. A subclass of deep learning models termed deep convolutional neural networks (DCNN) is widely used for vision-related tasks like detecting road signs. For the recognition of road signs in actual traffic situations, a unique two separate methods were proposed. In this proposed approach, road sign features are extracted using a pre-trained CNN model (AlexNet), which is then recognized using an SVM algorithm. The Myanmar Road Sign Dataset is created to train and evaluate the proposed method in the experiments. According to the experiment results, two separate methods with the pretrained CNN model (AlexNet) and SVM achieve high level accuracy with a 96.64% rate. But, the pretrained CNN model (AlexNet) gives 94.03% accuracy. These experimental results demonstrate that the proposed two separate methods performed better rather than the individual one.*

**Keywords**- Deep Convolutional Neural Network (DCNN), Deep Learning, Convolutional Neural Network (CNN), Road Sign recognition, AlexNet, SVM

## 1. Introduction

Automated road signs detection and recognition allows for little human involvement while cars move while adhering to modern traffic laws and regulations. In general, the main reason caused the traffic accidents is a failure to recognize road signs, a lack of knowledge of the regulations and driver distractions. Most of these incidents might be prevented by following the instructions on the road scene [1]. Since it is crucial for both the development of autonomous driving systems and for enabling drivers to understand and follow road signals, the road signs recognition has grown in prominence in the vision - based industry. Road signs are immediately recognizable by the human eye since they come in a variety of distinctive shapes and a variety of colors. Computers can identify and categorize road signs using the same features.

Each road sign differs from the others in terms of color and design [2]. Myanmar road signs have the following hues: red, blue, green, and yellow. They also typically have circular, rectangular, diamond or triangular designs. The signs selected for this study include backgrounds that are red, yellow, and blue, and their shapes are circular, rectangular, and diamond-shaped. The red, yellow, and blue items in the image have been located using the RGB color space. To expand the quantity of datasets, augmentation is utilized.

The following are the contributions that this paper made:

- A Myanmar Road Sign Dataset consisting of 25 classes which are significantly distinct and large number of road signs categories than the previous research is created.
- Evaluation of classification accuracy on proposed model on our created dataset.

In this paper, the combination of the pretrained CNN model (AlexNet) and SVM were proposed. This paper's remaining sections are labeled as follows. Section 2 presents the most recent studies on methodologies for identifying and detecting road signs. Section 3 provides information about the system overview, which also includes feature extraction with AlexNet, a pretrained CNN model for the feature extraction phase and SVM classifier for the recognition phase. The experiment's findings are presented in Section 4. The findings and suggestions for additional research are summarized in Section 5.

## 2. Related Work

Road signs recognition is a crucial part of the autonomous vehicle system. Road sign recognition has been the subject of extensive study. But there is still no perfect solution in this field due to the inconsistency of the types and patterns of road signs in various countries. The authors created a new dataset with images that are located in Beijing in various weather conditions and reported that the recognition accuracy is 80% based on template matching method [3]. Due to the development of technology, deep learning techniques can perform better than feature-based computer vision techniques [4]. So, the authors demonstrated recognition accuracy on Malaysia traffic sign dataset. Experimental results shows that most pre-trained models have an accuracy of 90%, however the DenseNet169 model has an accuracy of 98.33%. However, CNN has a lot of architectures. To get better accuracy, the researchers have been tested using VGG16, VGG19, AlexNet and ResNet50 on German Traffic Sign Recognition Benchmark dataset (GTSRB) [5]. Among these pretrained models, AlexNet model has a better accuracy than all the other implemented models. Image processing has made significant developments in the detection and recognition of traffic signs over the past ten years. However, there are still challenging in recognition since many traffic signs include variations, weather conditions, illumination, occlusion, deformation and color fading of signs and so on. The authors suggested a two-stage strategy for traffic sign detection and recognition to get around these problems [1]. The first stage is feature extraction which is performed in the HSV color space using histograms of oriented gradients (HOG) and SVM. After that, CNN is applied for traffic sign recognition. According to the experimental findings, SVM and CNN work together to create high level accuracy rather than working separately.

So, we proposed two separate methods for road sign detection and recognition system. In this system, we use combination methods with AlexNet that is the best classifier according to previous research and SVM.

## 3. System Overview

Preprocessing, feature extraction, and classification are typically used in the recognition of road signs. This section describes our proposed system architecture for recognizing road signs using two separate methods. First, we present to extract the road signs features using AlexNet, then we present the recognition stage of extracted road signs using SVM on Myanmar Road Signs Dataset.

### 3.1. Image Preprocessing

Preprocessing is a crucial stage in the recognition of road signs. The purpose of this initial step is to prepare the image ready for further processing. The input image is an RGB color image. The image has been downsized to 227*227 pixels. After then, the training data set will be expanded using the augmentation method. The number of training images increases as a result.

### 3.2. AlexNet Architecture

In 2012, this network structure was presented by Alex Krizhevsky [6]. It is a more complex and comprehensive CNN model than LeNet and won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), the most challenging ImageNet challenge for visual object recognition. To categorize the 1.2 million high-resolution photos in the ImageNet into the 100 different classes, a massive, deep convolutional neural network was trained. With top-1 and top-5 error rates of 37.5% and 17.0%, respectively, it has significantly outperformed the prior state-of-the-art. It achieved cutting-edge recognition accuracy when compared to every other computer vision and machine learning approach currently in use.

The architecture of AlexNet used in this system is shown in Table 1. Originally, it contains 8 layers (Convolutional layers make up the first five layers, and fully connected layers make up the remaining layers). The convolution and max pooling are carried out by the first convolution layer where 96 different kernels are used that are 11*11 in size. When performing the max pooling operations, a 3*3 kernel with a stride size of 2 is used. With 5*5 kernels, the identical procedures are carried out in the second

layers. The third, fourth, and fifth convolutional layers, with respective kernel counts of 384, 484, and 296 use 3*3 kernels. A Softmax layer is utilized at the end after two fully connected (FC) layers with dropout are used. There are 650,000 neurons and 60 million parameters in the neural network.

In this proposed system, AlexNet uses for feature extraction process. The preprocessed images were fed as input for AlexNet. The top six layers in AlexNet are used for feature extraction process. The 4096 feature vectors produced by AlexNet were provided as input to the SVM classifier.

**Table 1. An AlexNet architecture with 6 layers deep**

| Layer | Input | Parameters | Output |
|---|---|---|---|
| Conv1 | 3*227*227 | Kernels 11*11; Stride 4; Padding 0 | 96*55*55 |
| Pool1 | 96*55*55 | Kernels 3*3; Stride 2; Padding 0 | 96*27*27 |
| Conv2 | 96*27*27 | Kernels 5*5; Stride 1; Padding 0 | 256*27*27 |
| Pool2 | 256*27*27 | Kernels 3*3; Stride 2; Padding 0 | 256*13*13 |
| Conv3 | 256*13*13 | Kernels 3*3; Stride 1; Padding 1 | 384*13*13 |
| Conv4 | 384*13*13 | Kernels 3*3; Stride 1; Padding 1 | 384*13*13 |
| Conv5 | 384*13*13 | Kernels 3*3; Stride 1; Padding 1 | 256*13*13 |
| Pool5 | 256*13*13 | Kernels 3*3; Stride 2; Padding 0 | 256*6*6 |
| Fc6 | 256*6*6 | Kernels 6*6; Stride 1; Padding 0 | 4096*1 |

### 3.3. Support Vector Machine

In this proposed system, SVM is applied in the recognition step on the extracted features of AlexNet. Support Vector Machine is a class of supervised machine learning algorithm that may be used to both classification and regression applications [7]. The objective of the support vector machine is to find the best hyperplane in an N-dimensional space that distinctly classifies the data points. Figure 1 shows some patterns that are members of two classes: Class A and Class B.

There are a variety of different hyperplanes that might be used to divide these two groups of data points. By evaluating the largest margin, or the maximum distance between data points of both classes, the optimum hyperplane can be identified. The margin is the separation between the support vectors and the hyperplane. In figure 1, the black solid line is the best hyperplane.

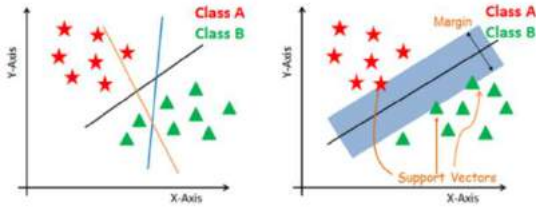



**Figure 1. The best hyperplane in SVM**

Figure 2 illustrates that some issues cannot be resolved with a linear hyperplane (left-hand side). As seen on the right, SVM in this situation employs a kernel approach to convert the input space into a higher dimensional space. Equation 1 is used to plot the data points on the x- and z-axes.

$$z = x^2 + y^2 \quad (1)$$

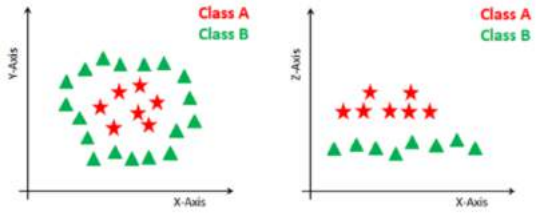



**Figure 2. Linearly non-separable dataset**

Using a kernel, an input data space is transformed into the desired form. An input space with low dimensions is transformed into a space with higher dimensions using the kernel.

Linear kernel:

$$K(x, xi) = sum(x*xi) \quad (2)$$

Polynomial kernel:

$$K(x, xi) = 1 + sum(x*xi)\ d \quad (3)$$

Radial Basis kernel:

$$K(x, xi) = \exp(-\gamma * sum(x-xi2)) \quad (4)$$

Where d is the degree of the polynomial and γ is a kernel parameter.

### 3.4. Road Sign Recognition System

For this proposed system, two separate methods have been proposed. Deep neural network will be used to detect the shape and color of road signs. Another method will be used to classify what the road sign is. Dataset is split into 20% for testing and 80% for training. For data augmentation, random rotations, translations, shearing, and zooming were applied to training images. Figure 3 shows the architecture of the proposed system.

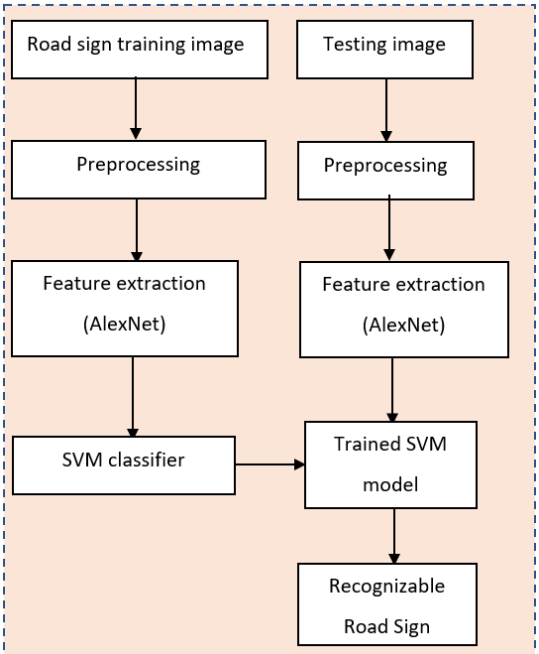



**Figure 3. Architecture of the proposed System**

## 4. Experimental Results

The performance of the proposed methods was evaluated in this section using Myanmar Road Sign Dataset. The MATLAB programming language R2022b version is used to implement the proposed methods. The system is implemented by an Intel ® i5-8250U running at 1.6GHz, 8GB of RAM, a 256GB SSD, and a NVIDIA GeForce MX130 GPU. The timeit function in MATLAB is used to measure how long the proposed system takes to run.

### 4.1. Datasets

There are 25 classes in the Myanmar Road Sign Dataset, and each class has between 11 and 160 images. Before augmentation, there are total of 2142 images for training and 536 images for testing. After augmentation, the training data would become total of 8568 images. Images were captured under a range of lighting settings and viewpoints. The dataset contains images that range in size from 32 by 32 to 225 by 225 pixels. Some of the images are taken from social media in other nations. Figure 4 illustrates the sample images of Myanmar Road Sign Dataset.

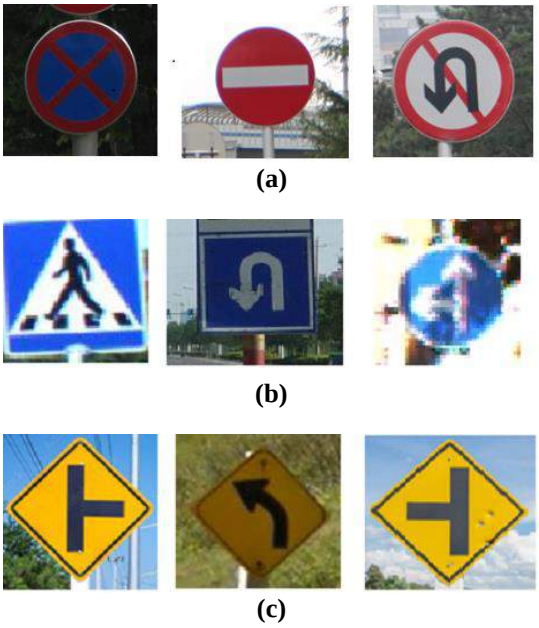



**Figure 4. Example of Myanmar Road Sign Dataset**

### 4.2. Evaluation

A confusion matrix is used to evaluate the proposed methods. The proposed methods' performance, recall, accuracy, and performance were calculated. The outcomes of the experiment are calculated using the following formulae. The evaluation results are displayed in Tables 2 and 3.

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (7)$$

where,

FP = False positive, FN = False negative, TP = True positive, TN = True negative

### 4.3. Results

Table 2 and 3 list the result of the proposed method. According to experimental results, AlexNet and SVM, the two separate algorithms that have been proposed, demonstrate that they achieve highest accuracy (See Table 2). It can also process data more quickly than individual ones (AlexNet) (See Table 3).


**Figure 5. Results of the proposed System**

**Table 2. Comparison of Recall and Accuracy**

| Model | Recall | Accuracy |
|---|---|---|
| **AlexNet, SVM** | 92.45 | 96.64 |
| **AlexNet** | 85.25 | 94.03 |

**Table 3. Comparison of Performance and Precision**

| Model | Performance | Precision |
|---|---|---|
| **AlexNet, SVM** | 80s | 89.17 |
| **AlexNet** | 25m 25s | 85.49 |

The proposed method has a better performance than the individual AlexNet. The main reason is that AlexNet is used only for feature extraction and SVM is used for training and classification in this proposed system. That is why, the input data for SVM is the extracted features of AlexNet. So, even when it has to be trained again, the input will be extracted features. On the other hand, no matter how many times it has to be trained, the input data will be the image. For the above reasons, the proposed method has a better performance.

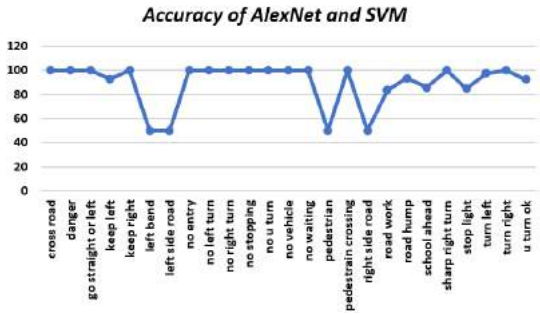



**Figure 6. Classification Accuracy of the proposed System**

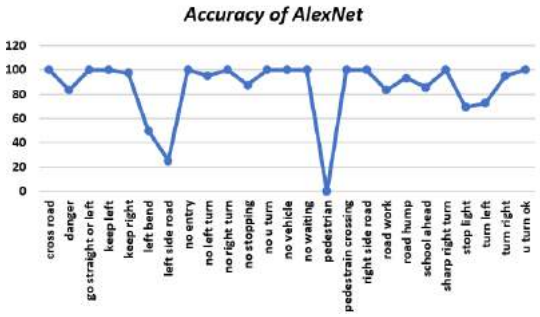



**Figure 7. Classification Accuracy of AlexNet**

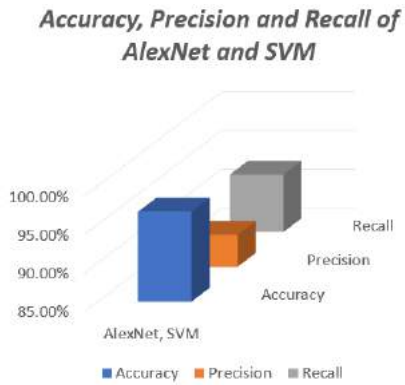



**Figure 8. Confusion matrix of AlexNet and SVM classification results**

Figure 6 shows the accuracy result of proposed methods in each class. The result of individual AlexNet for each class is shown in Figure 7. According to figure 6 and 7, the accuracy of left bend, left side road, pedestrain and right-side road have a lower accuracy than the other classes as there are few images in training dataset and some of the images have low resolution. Figure 8 shows the accuracy, precision and recall of the proposed methods. Accuracy, precision and recall of AlexNet are shown in figure 9. Figure 10 shows the comparsion results of proposed methods and individual AlexNet.

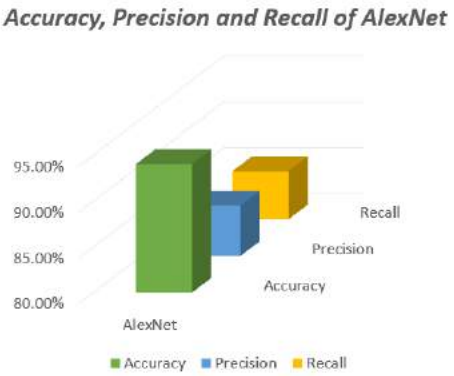



**Figure 9. Confusion matrix of AlexNet classification results**

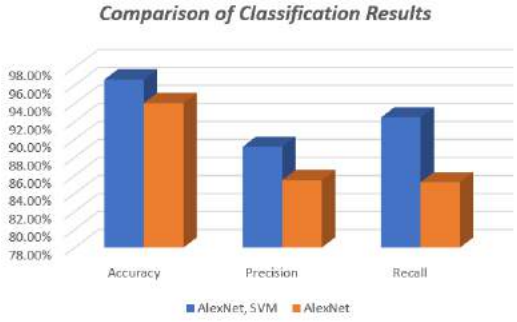



**Figure 10. Comparison of Classification Results**

## 5. Conclusion

Systems for detecting and recognizing road signs are essential for Myanmar. Road sign violations by drivers have a substantial impact on the persons who suffer from accidents on the roads each year. So, this research developed a system that can identify and categorize a collection of 25 road signs in various environments. This proposed system consists of two-stage: feature extraction and recognition. In the first stage, pretrained CNN model (AlexNet) is used to extract the road sign features. In the second stage, the extracted road signs are identified using SVM classifier. According to the experimental results, the proposed system performed well for the detection and recognition of road signs, obtaining a better recognition rate when compared to the outcomes of individual AlexNet model on the Myanmar Road Sign Dataset. In a future work, the created dataset will improve to include more significant road signs and the application will be implemented using other pretrained CNN models.

## References

[1] Ahmed Hechri and Abdellatif Mtibaa, “Two-stage traffic sign detection and recognition based on SVM and convolutional neural networks”, IET Image Process, 2020, pp. 1-8.

[2] Md Tarequl Islam, “Traffic sign detection and recognition base on convolutional neural networks” IEEE, 2020, pp. 2-3.

[3] Jiayuan Yu, Huiling Liu and Huayan Zhang, “Research on Detection and Recognition Algorithm of Road Traffic Signs”, IEEE, 2019, pp. 2-6.

[4] Dickson Neoh Tze How, Khairul Salleh Mohamed Sahari, Yew Cheong Hou and Omar Gumaan Saleh Basubeit, “Recognizing Malaysia Traffic Signs with Pre-Trained Deep Convolutional Neural Networks”, 4th International Conference on CRC, 2019, pp.1-5.

[5] Soulef Bouaafia, Seifeddine Messaound, Amna Maraoui, Ahmed Chiheb Ammari, Lazhar Khriji and Mohsen Machhout, “Deep Pre-trained Models for Computer Vision Applications: Traffic sign recognition”, 18th International Multi-Conference on SSD, 2021, pp. 1-6.

[6] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks “, 2012.

[7] Corinna Cortes and Vladimir Vapnik, “Support Vector Networks”, 1995.

[8] Domen Tabernik and Danijel Skocaj, “Deep Learning for Large-Scale Traffic-Sign Detection and Recognition”, IEEE Transactions on Intelligent Transportation Systems, 2019, pp. 1-3.

[9] David Cotovanu, Cristian Zet, Cristian Fosalau and Marcin Skoczylas, “Detection of traffic signs based on Support Vector Machine classification using HOG features”, EPE, 2018, pp. 1-2.

[10] Zain Nadeem, Abdul Samad, Zulkafil Abbas and Janzaib Massod, “A Transfer Learning based approach for Pakistani Traffic-sign Recognition; using ConvNets”, IEEE, 2018, pp. 1-2.

[11] Marco Magdy William, Pavly Salah Zaki, Bolis Karam Soliman, Kerolos Gama Alexsan, Maher Mansour, Magdy EI-Moursy and Kerolos Khalil, “Traffic Signs Detection and Recognition System using Deep Learning”, ICICIS, 2019.

[12] Omar BELGHAOUTI, Wahida HANDOUZI and Mohamed TABAA, “Improved Traffic Sign Recognition Using Deep ConvNet Architecture”, 4th International Conference on CIM, ScienceDirect, 2020, pp. 3-4.

[13] Banhi Sanyal, Ramesh Kumar Mohapatra and Ratnakar Dash, “Traffic Sign Recognition: A Survey”, AISP, 2020, pp.3-4.

[14] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Senior Member, IEEE, Hui Xiong, Fellow and Qing He, “A

Comprehensive Survey on Transfer Learning", arXiv, 2020, pp. 4-25.

[15] Wang Canyong, "Research and Application of Traffic Sign Detection and Recognition Based on Deep Learning", International Conference on Robots & Intelligent System, IEEE 2018, pp. 1-2.

# Distracted Driver Detection Based on Convolutional Neural Network

Thandar Oo, Amy Tun
*University of Computer Studies, Yangon*
*thandarict17@gmail.com, amytun@ucsy.edu.mm*

## Abstract

*In the recent years, the major cause of accidents is due to driver's distracted actions. Most of car accidents involve driver distraction under different forms such as talking on the phone, texting, operating the radio, drinking and talking with passenger and so on. In most cases of distractions, a driver just keeps only one hand or even no hand on the steering wheel. Therefore, detecting driver's hands on the wheel aims at several goal, i.e. the levels of pay attention of a driver to the road and using two hands or not while driving. In this system, deep learning method is used to detect and classify these driver's action This system is realized by using Fine-tuning-AlexNet Convolutional Neural Network to train and classify the driver's distracted behaviors. The system is implemented by Python programming language.*

## 1. Introduction

A road traffic accident claims the lives of almost 1.3 million people each year, according to research by the WHO (World Health Organization). Driving while distracted might raise the likelihood of a car accident. Distracted driving includes a variety of activities, including as texting, chatting on the phone, using a navigation system, and eating while operating a vehicle. Deep learning techniques, which identify and categorize a driver's behaviors whether they are distracted or not, play a significant role in helping to eliminate such distracted activities.

In order to classify between different driving behaviors, the proposed system uses Fine-tuning-AlexNet CNN architecture. This system receives a picture of a driver from dataset as input, and then outputs a particular kind of driver action. To build such a system, a dataset, well-trained transfer learning model and evaluation method for accuracy have to be applied. By controlling the driver concerned with his/her distracted action, the road accidents may be reduced.

## 2. Related Works

In 2022, M. Aljasim and R. Kashef [1] presented ensembled-based driver distraction detection with recommendation (E2DR) system to improve the detection accuracy and provide recommendations. In the E2DR model, two deep learning models are aggregated in a stacking ensemble. Recommendation layer is provided for distracted behaviors recommendations. The E2DR model enhances the generalization of the detection process and reduces overfitting.

In 2022, M. U. Hossain [2] used CNN based method to detect distracted driver and identify the cause of distractions like talking by means of face and hand localization. CNN, VGG-16, ResNet50 and MobileNetV2 have been adopted for transfer learning. This model is trained with ten driver conditions and analyzed the results.

## 3. Image Preprocessing

Image preprocessing includes the following:

- Rescale image: Image is scaled down by factor 255 before feeding to the model. 0 is black and 255 is white.
- Resize image: Resize process consists Stretch to, Fill (with center crop) in, Fit within, Fit (reflect edges) in, Fit (black edges) in, Fit (white edges) in.
- Normalization: When normalizing data, each pixel is first subtracted from its mean before the result is divided by standard deviation [3].

## 4. Image Classification

Image classification is the technique of identifying and labeling groups of pixels or vectors within an image according to predetermined rules. Supervised classification is the process of

identification of classes within a remote sensing data with inputs from and as directed by the user in the form of training data. Unsupervised classification is the practice of automatically identifying natural groups or structures within a set of remote sensing data [7].

## 5. Convolutional Neural Network (CNN)

CNNs are composed of multiple layers of artificial neurons. These neurons are mathematical functions. When taking an input image in ConvNet, each layer generates several activation functions that are passed on to the next layer [4]. CNN includes the following layers:

- Input layer: In the event where colored images of 640*480 pixel resolution are scaled down to 224*224 pixels to shorten training time.
- Convolutional layer: It needs input data, a filter, and a feature map, among other things. The input will be a color image with three dimensions—height, width, and depth. Moving over the image's receptive fields, a feature detector checks to see if the feature is there.
- Pooling layer: In this layer, dimensionality reduction takes place. These layers decreased the input volume's 2D dimensions to minimize over fitting or inefficient computing.
- Fully connected layer: The data that was extracted by earlier levels is combined in this layer to create the finished product. This layer makes a classification determination using the output of the convolution/pooling process. [5].

### 5.1. Activation Function

By generating a weighted total and then including bias with it, the activation function determines whether or not a neuron should be turned on. These functions are as follows [6]:

- Softmax function: It is used in the output layer of the classifier to attain the probabilities and to define the class of each input.

$$S(y_i) = \frac{exp^{(y_i)}}{\sum_{j=1}^{n} exp^{(y_i)}} \quad (1)$$

where, y is an input vector. The $y_i$ is the $i^{th}$ element of the input vector. The exp $y_i$ is the standard exponential function applied on $y_i$. The $\sum_{j=1}^{n} exp^{y_i}$ is the normalization term and the n is the numbers of classes.

- ReLU function: Rectified linear unit activation function (ReLU) will output the input directly if it is positive, otherwise, it will output zero.

$$f(x) = \max(0, x) \quad (2)$$

By resolving the vanishing gradient issue, this function enables models to learn more quickly and perform better.

### 5.2. Fine-Tuning -AlexNet Model

In this system, the Alexnet Model pre-trained on the ImageNet dataset, the winner of the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition, is fine-tuned by replacing the last fully connected layer and retraining the output parameters. FT-AlexNet Model is composed of five convolutional layers with a combination of max pooling followed by three fully connected layers and uses ReLU Activation in each of these layers except the output layer similar to the Alexnet Model. To prevent overfitting, it is also used the dropout layers. The input image size is 227x227x3.

In the first convolution layer, 96 filters of size 11x11 with stride 4 are used to extract features. The output feature map is 55x55x96. The first Max pooling layer of size 3x3 and stride 2 is used. Then the resulting feature map is the size of 27x27x96. In the second convolution, the filter size is reduced to 5x5 and has 256 such filters. The stride is 1 and padding 2. The output size is 27x27x256. Then, a max-pooling layer of size 3x3 with stride 2 is used. The resulting feature map is 13x13x256. FT-AlexNet model is shown in Figure 1.

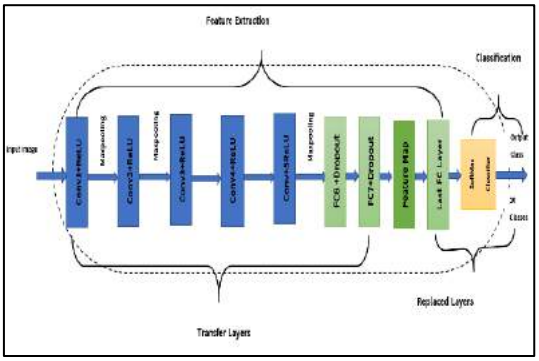



**Figure 1. FT-AlexNet Model**

The third convolution uses 384 filters that are 3x3, stride 1, and padding 1, respectively. The final feature map has the dimensions 13x13x384.

The fourth convolution utilizes 384 3x3 filters. The padding and stride together are 1. The output size, 13x13x384, stays the same.

The final convolution layer has 256 of these filters and is 3x3 in size. The final feature map has the dimensions 13x13x256. It uses the third max-pooling layer, which is a 3x3 stride layer. The feature map is 6x6x256 as a result. The first dropout layer is used which the drop-out rate is set to be 0.5. The first fully connected layer is with a ReLU activation function. The size of the output is 4096. Next layer comes another dropout layer with the dropout rate fixed at 0.5. Then, second fully connected layer is with 4096 neurons and ReLU activation. The last fully connected layer or output layer with 10 neurons is used in the FT-AlexNet Model and the activation function is Softmax [7]. The network has a total of 58322314 (about 58 million) trainable parameters.

## 6. Proposed System Design

This system is proposed as the districted driver detection system using FT-AlexNet CNN as shown in Figure 2. In Fine-tuning AlexNet architecture, the last fully connected layer with 10 neurons is used to classify the 10 driving classes. Firstly, the user must load the districted driver image. Then, the user can choose the training or testing phase. To classify the districted driver class, this system performs the training process.

Before training process, pre-processing steps including "rescale", "resizing" and "normalization" processes are required. Then, this system trains the fine-tuning AlexNet CNN model. In this network, five convolution layers and three fully connected layer with dropout and SoftMax activation function are used. The network is trained using the training data. After finishing training process, this system saves the prediction model.

In the testing process, the user must input the desired image. Then, this system performs pre-processing step. Then, this system classifies the tested image by using the prediction model that is obtained from the CNN. Finally, this system displays the relevant districted driver class.

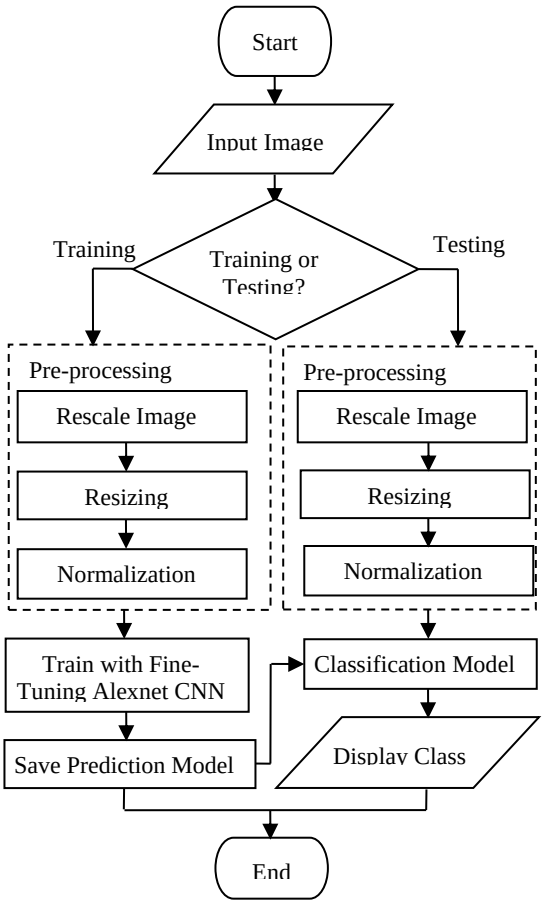



**Figure 2. Proposed System Design**

### 6.1. Distracted Driver Dataset

This system uses the "StateFarm" dataset as districted driver dataset [3].

**Table 1. Districted Driver's Classes**

| Class ID | Class Name |
|---|---|
| Class 1 | Driving safely |
| Class 2 | Texting with the right hand |
| Class 3 | Talking on the phone with right hand |
| Class 4 | Texting with the left hand |
| Class 5 | Talking on the phone with left hand |
| Class 6 | Operating the radio |
| Class 7 | Drinking |
| Class 8 | Reaching behind |
| Class 9 | Dressing hair and makeup activities |
| Class 10 | Talking to passengers |

"StateFarm" dataset includes 10 classes, and each image is classified among distracted driver classes. The dataset includes 102150 images of drivers. These districted driver's classes are shown in Table 1. Figure 3 shows the classes of driver postures.

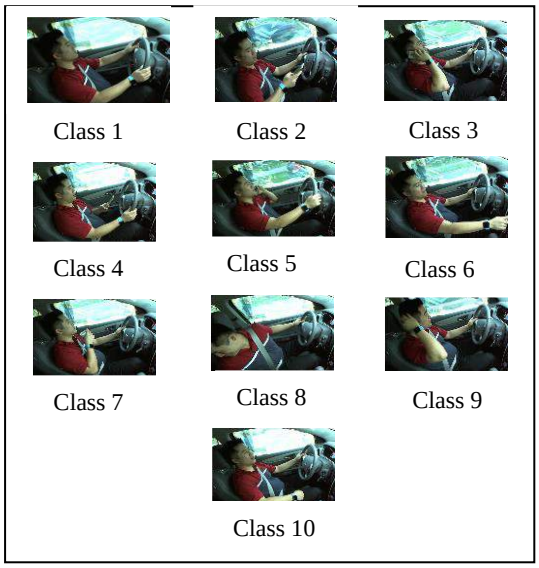



**Figure 3. 10 Classes of Driver Postures**

The dataset is split into train set that contains 79,726 images and test set that contains 22,424 images. In the dataset, all images are 640x480 pixels images for 26 different drivers. The ten classes are used for classifying driver behavior.

## 6.2. Distracted Driver Classification

The proposed districted driver detection and classification system is implemented by using python programming language.

For classification, the testing image must be chosen from anywhere. This image is shown in Figure 4. In this sample, the batch size for input image before pre-processing step is "32, 227, 227, 3". But, after finishing pre-processing step, this system produces the normalized image that has the "32, 10" batch size. Preprocessing step is shown in Figure 5.


**Figure 4. Testing Image**

```
Found 22424 images belonging to 10 classes.
Batch Size for Input Image :  (32, 227, 227, 3)
Batch Size for Output Image :  (32, 10)
Image Size of first image :  (227, 227, 3)
Output of first image :  (10,)
```

**Figure 5. Preprocessing Step**

This system trains the "79726" distracted driver images by using fine-tuning AlexNet CNN model. Table 2 shows this model architecture.

By using the AlexNet CNN model, this system detects driver and classifies this driver into the distracted driver class. Figure 7 shows the classification for distracted driver.


**Figure 7. Classification for Distracted Driver**

## 6.3. Experimental Results of the System

To measure the performance of distracted driver detection and classification process, this system is tested by using 102150 images that contains the 79726 training images and 22424 testing images. The performance evaluation methods are as follows:

**Table 2. Fine-Tuning AlexNet CNN Model**

| Layer | Filter Size | Strike | Size of feature Map | Activation Function |
|---|---|---|---|---|
| Input | - | - | 227x227x3 | - |
| Conv1 | 11x11x96 | 4 | 55x55x96 | ReLU |
| Max Pool1 | 3x3 | 2 | 27x27x96 | - |
| Conv2 | 5x5x256 | 1 | 27x27x256 | ReLU |
| Max Pool2 | 3x3 | 2 | 13x13x256 | - |
| Conv3 | 3x3x384 | 1 | 13x13x384 | ReLU |
| Conv4 | 3x3x384 | 1 | 13x13x384 | ReLU |
| Conv5 | 3x3x256 | 1 | 13x13x256 | ReLU |
| Max Pool3 | 3x3 | 2 | 6x6x256 | - |
| Droup out1 | rate=0.5 | - | - | 6x6x256 |

| | | | | |
|---|---|---|---|---|
| Fully Connected1 | - | - | 4096 | ReLU |
| Droupout1 | rate=0.5 | - | 4096 | - |
| Fully Connected2 | - | - | 4096 | ReLU |
| Last Fully Connected | - | - | 10 | Softmax |

- Accuracy: It is the percentage of correct classification of test dataset.

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fn+Fp} \quad (3)$$

- Precision: It is the positive predictive value.

$$Precision = \frac{Tp}{Tp+Fp} \quad (4)$$

- Recall: It is the ratio of correctly recognized images to the number of relevant images in dataset.

$$Recall = \frac{Tp}{Tp+Fn} \quad (5)$$

- F-score: It is basically harmonic mean of recall and precision.

$$F - Score = \frac{2(Precision*Recall)}{Precision+Recall} \quad (6)$$

- Error rate: It is the total numbers all incorrect predicted images to the total number of images in dataset.

$$Error\ rate = \frac{Fp+Fn}{Tp+Tn+Fn+Fp} \quad (7)$$

Overall accuracy results are shown in Table 3. Experimental results of the system are shown in Figure 8. Then, the results about each districted driver's class are shown in Figure 9.

**Table 3. Overall Accuracy Result**

| Name | Results of the System |
|---|---|
| Precision | 0.8851 |
| Recall | 0.9617 |
| F-score | 0.9216 |
| Accuracy | 0.9984 |
| Error Rate | 0.0016 |

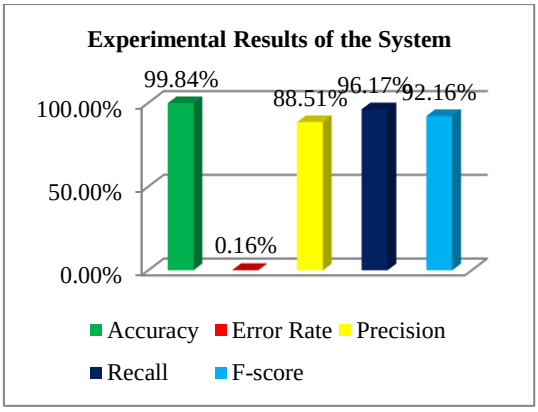



**Figure 8. Experimental Results of the System**

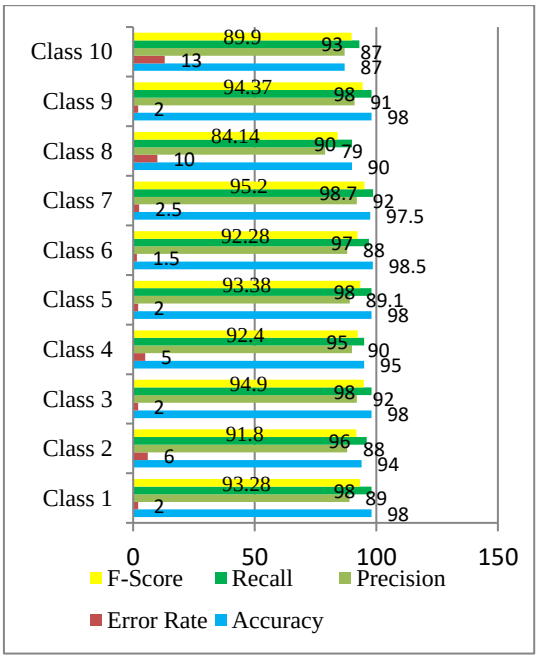



**Figure 9. Experimental Results for Each Class**

## 7. Conclusion

For distracted driver detection, this system used the fine-tuning Alexnet CNN architecture that solves the issue of road accidents because of driver distractions. This system used the images of StateFarm dataset from the Kaggle. This system is trained with FT-AlexNet CNN model by using the training dataset and learns the image features and saves the prediction model. In testing, this system classified the testing image of driver's behaviors with higher performance. The accuracy of proposed system is 99% and loss is 0.16% for most images of dataset. However, the accuracy for the images of two driving classes (talking to passenger class and reaching behind class) is less than other classes. For future work, FT-AlexNet architecture can be developed and used to detect the real-time images or raw images.

## References

[1] B. Nikhil, "Image Data Pre-processing for Neural Networks", *Chatbot Conference*, 2017.

[2] J. R. Rekkala, *Mobile Usage Detection of Driver using CNN (Convolutional Neural Network)*, California State University, Northridge, 2021.

[3] Kaggle (2016). State farm distracted driver detection. URL:Https://www.kaggle.com/c/state-farm-distracted-driver-detection Accessed: 2021-03-20.

[4] M. Aljasim and R. Kashef, "E2DR: A Deep Learning Ensemble-Based Driver Distraction Detection with Recommendations Model", *MDPI Journal*, vol. 22, no 5, 2022.

[5] M. H. Alkinani and W. Z. Khan, "Detecting Human Driver Inattentive and Aggressive Driving Behavior Using Deep Learning: Recent Advances, Requirements and Open Challenges", *IEEE*, 2020.

[6] M. U. Hossain, "Automatic driver distraction detection using deep convolutional neural networks", *Elsevier*, vol. 14, 2022.

[7] S. Masood, A. Rai and A. Aggarwal, "Detecting Distraction of Drivers using Convolutional Neural Network", *Elsevier*, 2017.

[8] W. Hao and W. Yizhou, "The Role of Activation Function in CNN", IEEE, 2020.

# Solid Trash Segregation System Using Convolutional Neural Network

Moh Moh Thet Aung, Amy Tun
*University of Computer Studies, Yangon*
*mohmohthetaung@ucsy.edu.mm, amytun@ucsy.edu.mm*

## Abstract

*All over the nations, trash management is a critical and ongoing problem. Many nations have distinctive waste management laws. The proper management of the trash is not currently using. The government may make substantial money from properly recycling garbage. Therefore, this proposed system is an automation method to accurately classify trash. Among the duties necessary for reuse, garbage sorting is a crucial step in enabling economical recycling. In this study, this system attempts to identify individual waste items in images and categorize them into appropriate categories. The Kaggle Garbage dataset is applied and divided into train, valid and test set. Training data is augmented by using augmentation methods and the testing and validation sets are resized according the input size of appropriate models. The simple convolutional neural network (CNN) and ResNet50 are used to classify the waste types. The system is implemented on the Google Colab and the highest accuracy on testing data is achieving 96%.*

**Keywords**: Solid Trash Segregation, CNN, ResNet50

## 1. Introduction

Many trash products are produced because these are used in daily. Some waste products have hazardous effects. By 2025, this amount was projected to increase to almost 21,012 tones/day with 0.85 kg per person (World Bank, 2015). But according to recent estimates, wastes are produced almost 20,000 tons of trash in every day as describing in 2017 [8]. Additionally, various solid wastes can really be reused again as recycling in different places. Most of the recycling trash are cardboard, paper, plastic, metal and glass. In this study, these are separated by using deep learning techniques such as simple CNN and ResNet50. The dataset is augmented by using augmentation methods.

After augmentation, the data is trained by classification model and the prediction model is saved. This prediction model is used to test the testing data. These advances are presently being used in research by various studies applying them to categorize. In Section 2, the related works are described. Methodologies of system are explained in Section 3. And then, step by step implementation is included in Section 4. The evaluation results are analyzed in Section 5. Finally, the system is concluded in Section 6.

## 2. Related Work

According to the paper [7], SVM was used to classify waste images. This paper proved that SVM is better performance than CNN. The dataset was categorized into only training 70% and testing 30%. Testing accuracy is achieving 63% by using SVM with scale invariant feature transform. This method was achieving better accuracy than CNN. In this experiment, CNN (AlexNet) was achieving accuracy 22% when testing data. According to this result, CNN was less in accuracy than SVM. In this experiment, the data was trained 60 epochs and the learning rates are changing in every 5 epochs.

In another related paper [7], namely waste classification using CNN algorithm which was one of the research projects of numerous classifications. In this study, it could be only used for classifying garbage based on their materials, organic and recyclable. The dataset comprised two components: a training set and a testing set. In this experiment, ReLU, Sigmoid, and Tanh activation functions were used and compared to quantify if a classification was successful or failed. The data was trained by using 50 epochs and 0.001 learning rate. According to their confusion matrix, the accuracy was achieving about 85.5%. Accuracy is not enough for classification because 800 training

and 200 testing images. Moreover, it can only classify two classes.

## 3. Methodologies

In this experiment, two types of models are used for classification. They are simple CNN and Resnet50.

**Simple CNN model:** In order to examine the performance variance between two models and evaluate the effectiveness of a fundamental CNN. A simple CNN architecture is firstly developed as seen in Figure (1). The input size is 300x300. For extracting of the images, convolutional layers are used. The filter size is (3x3) and rectified linear unit is used for non-linearization. After that, maximum pooling is used for reducing the training parameters and dimension of the input feature map. By doing so, it might be possible to prevent overfitting and maintain important traits after converting layers.

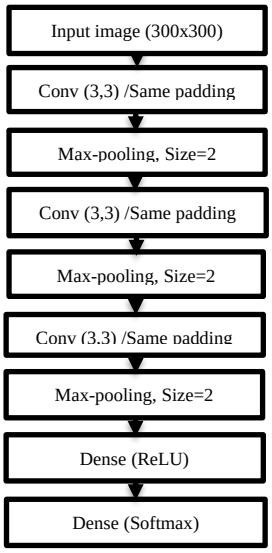



**Figure 1. Simple CNN architecture**

**Resnet50 Model:** This is also known as residual neural networks, are intended to fix the efficiency deterioration issue of convolutional neural network. It's a modern powerful version of CNN models. To handle the issue of performance degradation,it offers shortcuts connection between layers. This shortcut connections are also called “skip connection” [1][3].

There are two major categories of shortcut are identity block and convolutional block. Identity Block is defined as the situation where the same dimension applies to both the input and output activations. When the input and output are different, this is known as a convolutional block. In final layer, the SoftMax activation function is used. Figure (2) demonstrate the architecture of ResNet50 model. There are 50 convolutional deep layers and the input size is 256 x 256. The final layer, also called SoftMax layer classify the six types of trash.

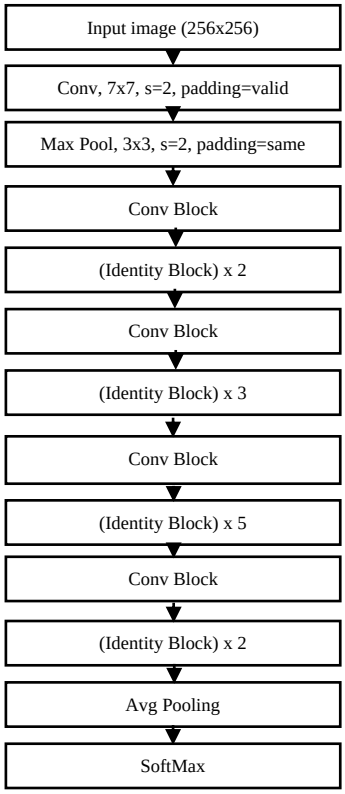



**Figure 2. ResNet50 architecture**

## 4. Proposed System Implementation

To implement this experiment, the following subtasks are organized.

**Collecting Data:** An appropriate dataset from Kaggle which is chosen for the objective of garbage categorization. An appropriate dataset is obtained from Kaggle that is appropriate for the purpose of garbage categorization.

**Data Preprocessing:** Before training the models, the training data set is needed to be augmented by using augmentation methods. The validation set is used to validate while training process and these images are needed to resize according to the acceptable inputs of models. By using test set, the accuracy performances are calculated.

**Model Training:** To improve results and model accuracy, the different alternative models

are applied and fine-tune the hyperparameters. These models, CNN and ResNet50 models are trained after hyperparameter tuning.

**Performance Evaluation:** Performance of model during its training is visualized. Test data is used to evaluate the effectiveness of models following the training processes and compare outcomes of the performances.

### 4.1. Dataset

The Kaggle website provided the data that was used in the study, which users can download for free. The dataset, which is openly available and goes by the term "Kaggle Garbage Classification," consists of 6 different classes [2]. These categories include trash, glass, paper, metal, cardboard, metal, and plastic. This dataset contains 2527 different photos, 403 of which belong to the cardboard category, 501 glass, 410 metal, 594 paper, 482 plastics, and 137 waste.

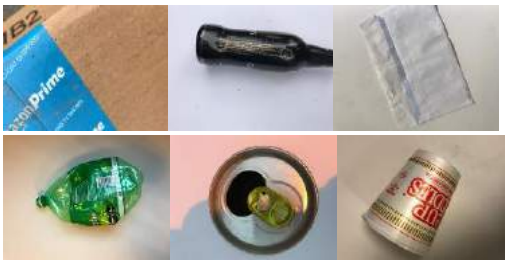


**Figure 3. Sample dataset of each category**

This dataset was applied to display the features of each category and was chosen suitable for our analysis. It was essential to use a clear dataset. The dataset is shown in the following. These are 512x384 pixels. This dataset is split into training80%, validation10% and testing10%. Each class of sample dataset is presented in Figure (3).

### 4.2. Data Preprocessing

**Resizing image**: The input images are resized before training. The size of the image should be suitable for requiring inputs of the different models. Therefore, the input images are needed to resize the dimensions that are concerned with appropriate models.

**Zooming image**: The zoom enhancement either erratically zooms in or out on the image. The lower and upper limits could be specified in a list with two values. If a float value is provided instead, zooming will take place between (1-zoom range,1+zoom range).

**Flipping image**: Vertical flip refers to flipping the rows of pixels in the input image, whereas horizontal flip refers to flipping the columns of pixels in the input image.

**Shifting image**: To randomly convert images vertically or horizontally, there are two ranges called width shift and height shift (as a percentage of total width or height).

**Rescaling image**: Before performing any extra processing, the input will be multiplied by a value called "rescale." The RGB parameters in image range from 0 to 255, but given a normal learning rate, such values would be too high for using models to handle. As a result, we scale original images by a factor of 1/255 to get values between 0 and 1.

**Image Shearing**: Shear describes an axis-based visual distortion that is typically used to create or correct perception angles. Usually, it's used to improve images so that computers can examine things from different angles much like people do [7].

### 4.3. Experimental Output

In the implementation steps, the dataset is categorized into three portions such as train, valid and test set. The trainable parameter of simple CNN is over 2millions and the trainable parameters of ResNet50 is over 23millions. The training process is done in 50epochs and trained on Google Colab's GPU. This is a web-based integrated development environment for Python. The training time is about an hour.

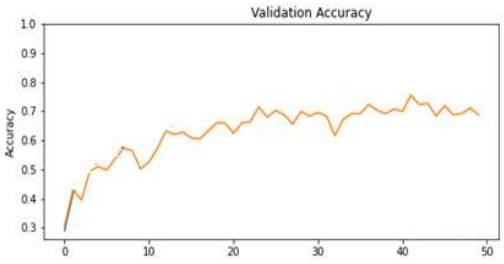



**Figure 4. Accuracy of simple CNN during training process**

The accuracy and loss of simple CNN during training process are shown in the Figure (4) and (5). In consequently, the accuracy and loss of

ResNet50 during training process are shown in the Figure (6) and (7).


**Figure 5. Loss of simple CNN during Training process**

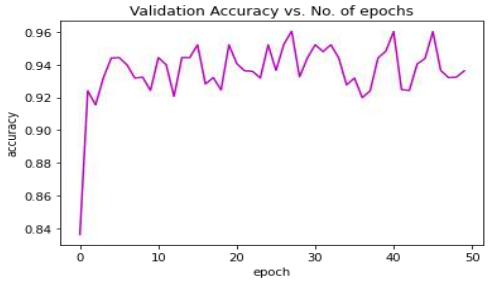



**Figure 6. Accuracy of ResNet50 during training process**

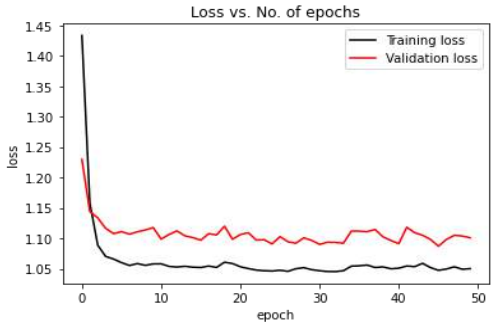



**Figure 7. Loss of ResNet50 during training process**

The Figure (8) shows the result of classification label on testing data.


**Figure 8. Classification Result**

## 5. Performance Evaluation

To assess the effectiveness of models, the test data is used for calculation. The performance calculations are accuracy, recall, precision and f1 score. The following equations are applied to calculate these. These results of each class have been shown in the following figures [4].

**Accuracy:** The percentage of accurate predictions for the testing dataset is meant by accuracy.

$$\textbf{Accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}} \quad (1)$$

**Precision**: This is measured by using percentage of instances that are truly relevant (also known as true positives) out of all the examples that were predicted to fall into a particular class.

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}} \quad (2)$$

**Recall**: Recall is the proportion of examples that were accurately forecasted to be members of a class relative to all of the actual members of the class.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \quad (3)$$

**F1 Score**: the precision and recall high chances. Therefore, this score takes into account both false positives and false negatives.

$$\textbf{F1 score} = 2 * \frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \quad (4)$$

**Macro Average**: is computed using the arithmetic mean of all the per-class F1 scores.

**Weighted-average**: F1 score is calculated by taking the mean of all per-class F1 scores while considering each class's support.

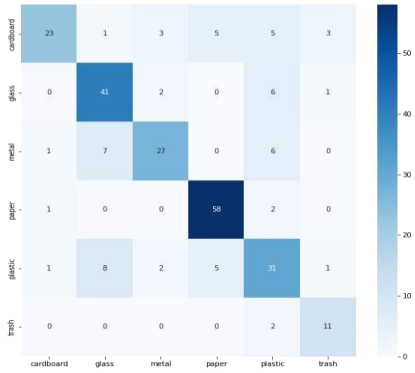



**Figure 9. Confusion matrix of simple CNN**

Figure (9) and (10) show the confusion matrix of testing data to analyze the accuracy of each category of two models. According to these results, the error rating of simple CNN is higher than ResNet50.

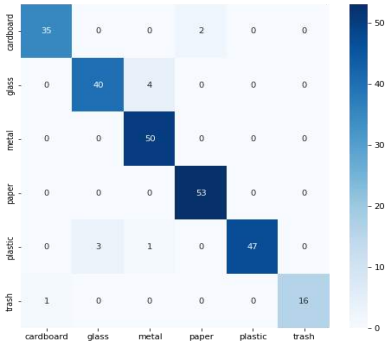



**Figure 10. Confusion matrix of ResNet50**

By comparing the performance of two models in Figure (11) and (12), the evaluation results of each category using simple CNN is lower than using ResNet50. ResNet50 is getting the highest accuracy of 96% in classification.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.57 | 0.70 | 40 |
| 1 | 0.72 | 0.82 | 0.77 | 50 |
| 2 | 0.79 | 0.66 | 0.72 | 41 |
| 3 | 0.85 | 0.95 | 0.90 | 61 |
| 4 | 0.60 | 0.65 | 0.62 | 48 |
| 5 | 0.69 | 0.85 | 0.76 | 13 |
| accuracy | | | 0.75 | 253 |
| macro avg | 0.76 | 0.75 | 0.74 | 253 |
| weighted avg | 0.76 | 0.75 | 0.75 | 253 |

**Figure 11. Performance evaluation of simple CNN**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| cardboard | 0.97 | 0.95 | 0.96 | 37 |
| glass | 0.93 | 0.91 | 0.92 | 44 |
| metal | 0.91 | 1.00 | 0.95 | 50 |
| paper | 0.96 | 1.00 | 0.98 | 53 |
| plastic | 1.00 | 0.92 | 0.96 | 51 |
| trash | 1.00 | 0.94 | 0.97 | 17 |
| accuracy | | | 0.96 | 252 |
| macro avg | 0.96 | 0.95 | 0.96 | 252 |
| weighted avg | 0.96 | 0.96 | 0.96 | 252 |

**Figure 12. Performance evaluation of ResNet50**

## 6. Conclusion

A crucial component in the waste management process is trash sorting. This study demonstrates that many deep learning techniques could be used to classify garbage with higher accuracy. A dataset of rubbish images on Kaggle, multiple CNN-based models were trained for this challenge. When dataset is divided into training 80%, validation10% and testing 10%, the accuracy of simple CNN is achieving 75%. And the highest accuracy on testing data is 96% by using ResNet50 in training epoch50.

## References

[1] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola, “Dive into Deep Learning”, Release 1.0.0-alpha0, Jul 30, 2022.

[2] Dataset_https://www.kaggle.com/datasets/asdasdasdas/garbage-classification.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep Residual Learning for Image Recognition; arXiv:1512.03385, 2015.

[4] Kartik Nighania, “Various ways to evaluate a machine learning model’s performance”, Towards Data Science, Dec 30, 2018.

[5] M. Yang and G. Thung, "Classification of trash for recyclability status," CS229 Project Report, vol. 2016.

[6] Mohammad Diqi, “Waste Classification Using CNN Algorithms” 1st International Conference on Science and Technology Innovation (ICoSTEC), ISBN: 978-623-331-338-4, February, 26 2022.

[7] Niteesha Balla, “Five Data Augmentation Methods in Deep Learnings”, Data Aspirant, August 31, 2020.

[8] S. Kaza, L. Yao, P. Bhada-Tata, and F. Van Woerden, What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050. The World Bank, 2018.

# Myanmar Sign Language Recognition System using Support Vector Machine (SVM) and Kernel Principal Component Analysis (KPCA)

[1]Eaint Thu Thu Khaing, [2]Yi Mon Shwe Sin
[1]*University of Computer Studies, Taunggyi,* [2]*University of Computer Studies, Yangon*
*eaintthuthukhaing@ucsy.edu.mm, yimonshwesin@ucsy.edu.mm*

## Abstract

*Sign Language is the essential language for deaf and dumb people. Sign language became a boon for the physically challenged people to express their thoughts and emotions. A system that can translate is needed when a normal person wants to talk with a dumb or deaf person. Our proposed system was built to classify 11 static number signs and 30 static and opened finger spelling consonant signs. In our proposed system, there are three main processes, namely, preprocessing, features extraction and classification for extracted features. One of the feature extraction methods in image processing, namely, Kernel Principal Component Analysis (KPCA) is combined with Supportive Vector Machine (SVM) to create a Myanmar sign language recognition system. For classification of the extracted features, Supportive Vector Machine (SVM) is used. Among many kernels, Gaussian Radial Basis Function (RBF) is used together with SVM to classify non-linear data in higher dimension. As a result, Kernel Principal Component Analysis with Supportive Vector Machine have the highest accuracy compared with Principal Component Analysis.*

**Keywords**: Sign Language Recognition, Feature Extraction, Kernel Principal Component Analysis, Supportive Vector Machine

## 1. Introduction

The history of Sign Language appeared from Western societies since 17th century. Sign Language is a visual method to communicate among deaf people. Unlike oral languages, Signs are expressed by using different hand shapes to represent words.

In Myanmar, 673,126 of population are disable in hearing according to 2014 Myanmar national census. 4.6% of population are disable, 1.3% of the population are deaf and hearing impairment and only 0.0006% of the deaf have a university level. There are four Deaf schools in Myanmar, namely, Mary Chapman School for the Deaf, Yangon (est.1904), School for the Deaf Children, Tamwe, Yangon (est.2014), School for Deaf Children, Mandalay (est.1964) and Immanuel School for the Deaf, Kalay (est.2005).

Most of the deaf people cannot read and write well. Only a few deaf people can involve in social inclusion due to the communication difficulties. To overcome these barriers, researchers make much research in computer vision, image processing, natural language processing etc. deal with Sign Language translation and recognition. But Sign Language Recognition remains a challenging task in research areas. The proposed system aims to recognize 30 static consonant signs (from ကြ (ka) to အ (a) except two dynamic signs ဇ (ja) and ဈ (jha) and one closed finger spelling consonant sign ဏ (nna)) and 11 static number signs (from 0 to 10)) by using Kernel Principal Component Analysis and Supportive Vector Machine (SVM). Among many feature extraction methods, non-linear principal component analysis, namely Kernel Principal Component Analysis (KPCA) is used to extract maximum variance non-linear components from data. In the classification stage, Supportive Vector Machine is used to make best hyperplane that can segregate different classes. This study employs 6150 hand gesture photos for 30 static alphabets and 11 static numbers from 30 different people.

This paper is organized into five sections. Section 2 will discuss about the related works of this paper. Background theory and preparation of dataset are described in section 3. Section 4 describes the design and implementation of the system. Finally, this paper is concluded in section 5.

## 2. Related Work

Since the last two decades, researchers had been paying attention in the area of sign language recogniton. Unfortunately, every research had limitions and required many expensive devices such as sensors or 3D cameras, gloves.

[1] Ni Htwe Aung, Su Su Maung, Ye Kyaw Thu developed a sign language recognition system for Myanmar numbers from zero to ten.They also investigated the performance of three different Supportive Vector Machine(SVM) classifiers: SVC with polynomial kernel, SVC with linear kernel and LinearSVC. In the preprocessing stage, recorded videos are transformed into multiple frames. That converted frames are cropped for only regions and resized into 128*128 resolution. And then those images are changed into grayscale format. In the feature extraction stage, distinctive invariant features are extracted from converted grayscale image data by using Scale Invariant Feature Transform (SIFT). In the last stage, classification stage, extracted features are fed into three different classifiers for recognition of hand signs. SVM with polynomial kernel achieved the accuracy score of 88%.

[2] Face Recognition between Two Person using Kernel Principal Component Analysis and Support Vector Machines developed by Ivanna K. Timotius, Iwan Setyawan, and Andreas A. Febrianto compared the performance by the using Kernel Principal Component Analysis and Support Vector Machine with a pair of Kernel Principal Component Analysis and Nearest Neighbor classifier. Image data are first transformed into feature space by using kernel function to reduce dimension. Features obtained from feature extraction stage are fed into SVM to classify. Finally, KPCA with SVM had 99.05% of accuracy and KPCA with NN had 97.14% in accuracy.

[3] Myint Tun, Thida Lwin developed Real-time Myanmar Sign Language Recognition System using PCA and SVM. In that system, input video stream is detected by Viola-Jones Object Detection Framework and converted into YCbCr color space. After threshold segmentation, they made removing background from the received binary image and marking hand region for feature extraction. For feature extraction, Principal Component Analysis (PCA) is used to extract maximum variance feature by reducing dimension. Extracted features are fed into Support Vector Machine to recognize hand signs. Experimental results showed that the system gave the successful recognition accuracy of static sign gestures of MSL alphabets with 89%.

## 3. Methodology and Dataset

### 3.1. Feature Extraction

The process of feature extraction expresses the appropriate shape of information included in a pattern to make easy the processing of classification by a formal procedure. In the field of image processing, feature extraction is used for dimensionality reduction. The objectives of feature extraction are to get the most appropriate information from original dataset and to represent that information into low dimensional feature space as feature vectors. Not much information is useful for process of classification. In large and multi-dimensional dataset, such as image datasets, signal processing, etc. only just a few data can be transformed into feature vectors for classification by using feature extraction methods. This reduction leads to build model with less computational cost and can also speed up the learning and generalization step in machine learning processes.

### 3.2. Kernel Methods

In many big data applications, like signal processing, geostatistics, krigin, inverse distance weighting, 3D reconstruction, bioinformatics, chemoinformatics, information extraction and handwriting recognition, kernel methods are mostly used to achieve better performance. Data can be divided into two forms: linear and non-linear. For non-linear data, kernel methods can be used to transform data into high dimensional feature spcae in order to achieve better pattern classification. The backbone of every kernel method is kernel function that performs dot product in transformed space.
Mathematical definition:

$$k(x,y)=(f(x),f(y))$$

where, K is the kernel function, x and y are inputs in n dimensional. f is a mapping function from n-dimensional feature space from m-dimensional feature space. <x,y> can be denoted as the dot

product. m is usually much larger than n.Some commonly used kernel functions are:

1. Linear: $K(x,y) = x^T y$
2. Polynomial: $K(x,y) = (1 + \frac{x^T y}{\sigma^2})^d$
3. Radial Basis Function:

$$K(x,y) = exp\ (-\frac{||x-y||^2}{\sigma^2})$$

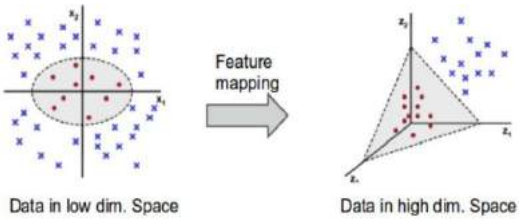



**Figure 1. Feature mapping from two to three dimension**

### 3.3. Kernel Principal Component Analysis

Feature extraction methods can be used to reduce the redundant data by capturing the maximum variance data. Among many types of feature extraction methods, Kernel Principal Component Analysis is the best suite for multi-dimensional dataset. Kernel Principal Component Analysis (KPCA) is an extension of Principal Component Analysis (PCA) for non-linear feature extraction. In Principal Component Analysis (PCA), a subspace is created using orthogonal basis vectors related to the maximum variance directions. These directions can be calculated from the eigenvectors of the covariance matrix. PCA cannot work for features extraction of non-linear data efficiently. An extension method of PCA, Kernel Principal Component Analysis (KPCA) can solve that problem by using kernel functions. Based on the action of that function, KPCA projected the input space to the high-dimensional feature space through non-linear mapping. After that, PCA can be performed on the data that exists in high-dimensional feature space in order to produce non-linear principal components. The process of mapping is shown in figure 2.

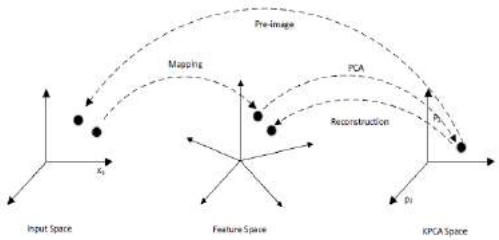



**Figure 2. Mapping Process of KPCA**

Let's assume Φ(x) is the mapping function to transform data into a high-dimensional space, N is the number of samples and d is the dimension . $w_i$ (i=1,…,d) is eigenvector in transformed feature space, where D(D>>d) is dimension vector and $\lambda_i$(i=1,…,d) is eigenvalues related to the above eigenvector. PCA in that space can be defined as:

$$\boldsymbol{\phi(X)\phi(X)^T = \lambda_i w_i} \quad (1)$$

$w_i$(i=1,…,d), the eigenvector, can be described linearly by using the sample set $\phi(X)$ as:

$$\boldsymbol{w_i = \sum_{k=1}^{N} \alpha_i\ \phi(X_i) = \phi(X)\ \alpha} \quad (2)$$

After substituting, it can be defined as:

$$\boldsymbol{(X)^T \phi(X)\phi(X)^T\ \phi(X)\alpha = \lambda_i \phi(X)^T\ \phi(X)} \quad (3)$$

We can replace $\phi(X)^T\ \phi(X)$ with kernel matrix K which is a symmetric matrix:

$$\boldsymbol{k\alpha = \lambda_i \alpha} \quad (4)$$

Where, $\alpha = (\alpha_i, \dots, \alpha_n)^T$ is the eigenvector of kernel matrix K.

For a new sample for testing, $x_{new,}$ reduction of dimension can be defined as:

$$\boldsymbol{\bar{x}_{new} = [\alpha_i, \dots, \alpha_N][k(x_1, x_{new}), \dots, k(x_N, x_{new})]^T} \quad (5)$$

### 3.4. Support Vector Machine (SVM)

In every recognition system, the process of classification is the main component used to predict the actual class from a given set of data. It can be performed on both structured or unstructured data. There are many machine learning methods for classification, namely, k-nearest Neighbor (KNN), Random Forest, Supportive Vector Machine and so on. Among them, Support Vector Machine (SVM) is one of the most popular Supervised Learning algorithms. It can be used for both Classification and Regression problems. The goal of SVM is to create the best line or decision boundary that can separate n-dimensional space into classes. In the applications of face detection, image classification and text categorization, SVM is mostly used for classification.

There are two types of Support Vector Machines:

Linear SVM is mainly used for linearly separable data which means that the data that can be separated by using a single straight line, mostly called Linear Support Vector Classifier.

Non-Linear SVM can be applied for unstructured data. If a dataset cannot be classified by using a straight line, non-linear SVM classifier must be used to create the best hyperplane by using appropriate kernel functions like polynomial kernel of degree h, Gaussian radial basis kernel function and sigmoid kernel.

The formula of decision boundary for Support Vector Machine can be expressed as follow:

$$g(x) = w^T x + b = 0$$

where, w is vector normal to hyperplane and b is the offset.

### 3.5. Dataset

Data are collected according to the format of Myanmar consonants (from ကခ(ka) to အ(ah)) and Myanmar numbers (from 0 to 10) from dictionary book published by Department of Social Welfare, Ministy of Social Welfare, Relif and Resettlement. Sign images are taken from 30 different people with different backgrounds by using Canon PowerShot SX620 HS with the resolution of 5184x3888, saved them as “jpg” format. Images that include only hand region are cropped manually and put into the proprocessing step for featrue extraction. In the dataset, there are 150 one-handed and two-handed images for each of 11 different number images and 30 different consonant images. Samples of signs pictures that represent Myanmar numbers and Myanmar consonants are displayed in figure 3 and 4 respectively.

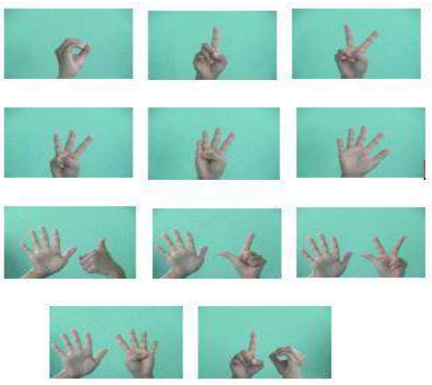


**Figure 3. Images for Myanmar Number**

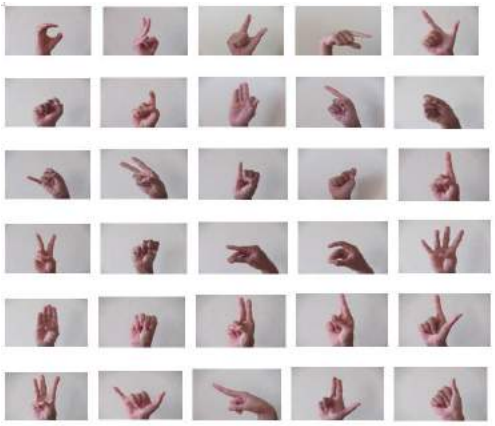


**Figure 4. Images for Myanmar Consonant**

## 4. System Design and Implementation

The flow chart for the proposed system is shown in figure 5. In the proposed system, there are two sections, namely, training and testing. In the training process, images of static sign gesture are preporcessed with the aim to provide the quality of the images higher. Processing step is the main part in every image classification task to avoid undesired distortions on image features.

In this system, sign images are converted into grayscale format and resized into 128x128 images. And these images are combined to extract array file and stored as .npy file. After the preprocessing step, feature extraction process is started by using Kernel Principal Component Analysis (KPCA) with radial basis function. The radial basis function with appropriate values of gamma and C is used to constuct kernel matrix. From that kernel matrix, non-linear principal component that retains the most imporant information are extracted. The explained_variance function is used to find the less principal components in order to speed up the training process. For classification stage, the non-linear features received from feature extraction step are trained by using Support Vector Machine (SVM). Before the model creation, gridsearchcv is used to find the best decision boundary that can classify different classes correctly.

The major aim of the gridserahcv is to determine the ideal parameter values in a grid from a set of parameters. The SVM model is created with 80% of training and 20% of test. Then the created model is stored as joblib file for later classification. In the testing process, input image is cropped manually, converted into grayscale format and resized into 128x128 images like as

images in training. Reduced features of a test image can be classified by using created SVM model and the result is shown as text. The interface of the system is shown in figure 6 and 7. User can start the recognition process by choosing the desired image with browse button. The classification result is shown with label. Image entered is preprocessed first by making hand region cropping, resizing and grayscale conversion. After that step, features are extracted for classification and text result is shown with a label.

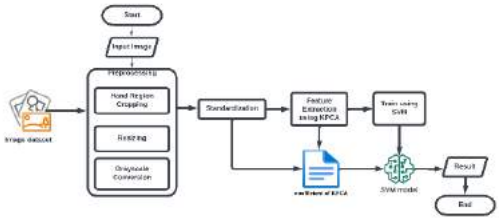


**Figure 5. Process flow of proposed system**


**Figure 6. Recognition of number 10**


**Figure 7. Recognition of consonant ဖ(pha)**

## 5. Conclusion

The proposed system can be used for recognition of Myanmar consonants and Myanmar numbers including 30 static consonant signs and 11 static number signs. This system can be used to recognize sign gestures without the need of expensive resources. This system can also be able to speed up the performance of machine learning process by keeping maximum variance features. There are some limitations with dataset. But the accuracy is 82% after testing the performance of model by splitting the dataset into 80:20.

## References

[1] Ni Htwe Aung, Su Su Maung, Ye Kyaw Thu, "Sign Language Recognition for Myanmar Number Using Three Different SVM Classifiers", NCSE 2019, 27th - 28th June 2019, Yangon

[2] Myint Tun, Thida Lwin, "Real-time Myanmar Sign Language Recognition System using PCA and SVM", International Journal of Trend in Scientific Research and Development (IJTSRD), Volume 3 Issue 5, August 2019.

[3] Ni Htwe Aung, , Ye Kyaw Thu, , Su Su Maung, , Swe Zin Moe, Hlaing Myat New, "Transfer Learning Based Myanmar Sign Language Recognition for Myanmar Consonants", Journal of Intelligent Informatics and Smart Technology, VOL. 4, APRIL 2020

[4] Adithya V, Rajesh R., "A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition", Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communications (CoCoNet'19), 10.1016/j.procs.2020.04.255

[5] Face Recognition between Two Person using Kernel Principal Component Analysis and Support Vector Machines developed by Ivanna K. Timotius, Iwan Setyawan, and Andreas A.

[6] A. Kulkarni, P. Halgekar, G. R. Deshpande, et al., "Dynamic sign language translating system using deep learning and natural language processing", Turkish Journal of Computer and Mathematics Education, Vol.12 No.10(2021), 129-137

[7] P. S. Santhalingam, P. Pathak, J. Košecké, H. Rangwala, et al. "Body Pose and Deep Hand-shape Feature Based American Sign Language Recognition". In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE. 2020, pp. 207–215.

[8] M. M. Rahman, M. S. Islam, "A new benchmark on american sign language recognition using convolutional neural network". In: 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI). IEEE. 2019, pp. 1–6.

# Low-Light Image Enhancement with ResNet Architecture and Self-Calibrated Illumination Network

Zayar Tun, Khant Kyawt Kyawt Theint
*University of Computer Studies, Yangon, Myanmar*
*mrzayartun303@gmail.com, khant2theint@gmail.com*

## Abstract

*Generally, low-light image enhancement techniques are mostly not just made to achieve both visual quality and computational efficiency but also commonly invalid in unknown complex scenarios. The system is focused on the image high quality displaying of low-light images using enhancement techniques. This system is used Self-Calibrated Illumination (SCI) module Network combination with Convolutional Neural Network (CNN) [2] based on ResNet architecture Network to enhance the low-light image. In this system, Low-Light (LOL) dataset is applied. The system will be used LOL testing dataset for performance evaluation of the model. This system is implemented with the software program as Python language code and Anaconda application for running. Moreover, this system uses the three types of low-light image dataset for testing as LOL images, captured images by Camera and Black & White dataset.*

**Keywords**: Low-Light Image Enhancement (LLIE), Convolutional Neural Network (CNN), ResNet Architecture and Self-Calibrated Illumination.

## 1. Introduction

People often get some low-light image in the process of image acquisition because of the complexness of the geographical environment and time. And then, low-light images usually get from two problems. First, they have low visibility that is small pixel values. Second, noise convert significant and reduce the image content, because of low signal to noise ratio. Therefore, low-light image techniques would be used.

The low-light image techniques described as followed:

- Low-Light Image Enhancement (LLIE)
- Using Dehazing Algorithm, Low-Light Image Enhancement (LLIE)
- Using Imreducehaze Optional Parameters, Improving Results Further
- Improving Low-Light Image of Another Example
- Using Different Color Space, Reducing Color Distortion
- Using Denoising, Improving Results
- Illumination Map Estimation
- Controlling

There are different methods proposed by various researchers until now which build Retinex [11] for image contrast enhancement. Low-light image enhancement usually formal images captured in low-light quality like night time, where the usual goal is to illumination and improve the contrast of the image for more visual quality and display details that were unseen in night. In addition, examples and methods of low-light image enhancement are Median filtering, Noise removal using Wiener, Unsharp mask filtering, and so on. The purpose of low-light image enhancement is to increase the light and clarity of image. Therefore, people can provide better useful. And then, building a deep convolutional neural network model is applied ResNet architecture and Self-Calibrated Illumination (SCI) Networks for fast, adjustable, and powerful better illumination images in real-world low-light scenarios.

## 2. Related Work

Though low-light image enhancement belongs to low-level image processing works, it differs a lot from super resolution and image denoising. Therefore, CNN architecture would be used to enhance low-light images. And then, Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee [5] proposed network as Low-light Convolutional Neural

Network (LLCNN) [2]. It trains to filter low-light images with various kernels and then addition multiscale feature maps with enhanced images, which seem to be captured lower normal light conditions, and opposite input features and textures.

### 2.1. Background Theory

Most deep learning methods apply neural network architectures, which is why deep learning models are known as deep neural networks. The term deep especially describes the number of hidden layers in the neural network. Traditional neural networks just contain 2-3 hidden layers, while deep networks can have like 150. Deep learning models are trained by using large sets of labeled data and neural network architectures that learn features directly from data no for manual feature extraction.

Deep learning is usually serval from conventional machine learning. With deep learning, all that is needed is to supply the system with a larger number of cat images, and the system can autonomously train the features that display a cat. As computer vision, speech recognition referred neural language processing, machine translation, and robotics, the performance of deep learning systems far exceeds that of conventional machine learning systems. Figure 1 describes the structure of Artificial Intelligence, Machine Learning and Deep Learning.

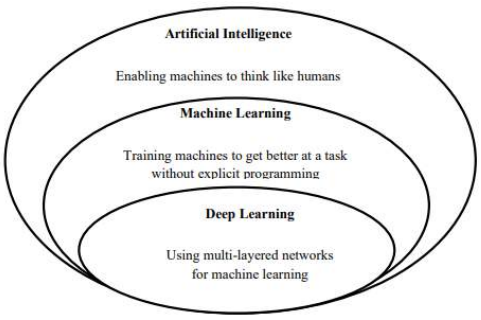



**Figure 1: The Structure of Artificial Intelligence, Machine Learning and Deep Learning**

### 2.2. Low-Light Image Enhancement

To addition image contrast and improve image light, different algorithms were proposed. Histogram equalization (HE) methods [4][11], repair pixel values to make obey uniform distribution. Retinex-theory-based methods applied a model which assumes an image is an interaction of illumination and reflectance.

All these methods as traditional methods. Recently, convolutional neural network (CNN) [2] achieves impressive progress in several computer vision applications. As to low-level image processing applications, CNN makes different breakthroughs in super resolution [5], image denoising [7][13], and so on.

### 2.3. Self-calibrated Convolutions

Self-calibrated convolution is smooth and generic, and can be applied to augment standard convolutional layers no extra parameters and complexity. The feature transformation process is performed in multiple parallel branches and the outputs from each branch are concatenated as the final output. Similar to grouped convolutions, the proposed self-calibrated convolutions also split the learnable convolutional filters into multiple portions, yet differently, each portion of filters is not equally treated but responsible for a special functionality. When applying self-calibrated convolutions into different backbones, our networks can increase the baseline models in a variety of vision tasks, including image recognition, object detection, segmentation, and key point detection, with no need to convert the network architectures.

## 3. Residual Networks (ResNets) of Deep Learning

Which was proposed in 2015 by researchers at Microsoft research, introduced a new architecture called Residual Network [6][9][13]. In this network, Figure 2 is the skip connection connects activations of layer to further layers by skipping some layers in between. This forms a residual block. ResNets are constructed by stacking these residual blocks together. Therefore, H(x), initial mapping, let the network fit:

$$F(x) = H(x) - x \text{ which gives } H(x) = F(x) + x \quad (2.1)$$

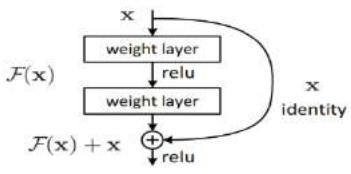



**Figure 2: ResNet of Skip Connection**

### 3.1. Residual Network (ResNet) Architecture

A residual neural network (ResNet) [6][9][13] refers an artificial neural network (ANN). It is the first working better deep feedforward neural with hundreds of layers, better previous neural networks. Skip connections are used to jump over some layers. Figure 3 is typical ResNet models are executed with double-or triple-layer skips, these models contain nonlinearities (ReLU) and batch in between.

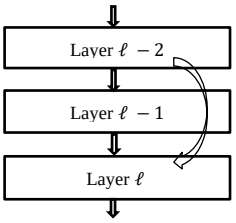



**Figure 3: Temporal form of a residual neural**

A convolutional layer consists 64 filters with a kernel size of 7x7 this is the first convolution, next followed by a max-pooling layer. And then, in conv2_x the pooling layer and the following convolution layers. The 2 layers are kernel_size of 3x3, num_filters are 128. This continues average pooling and the SoftMax activation function.

## 4. Proposed System

The proposed system flow of low-light image enhancement with CNN is shown in Figure 4.

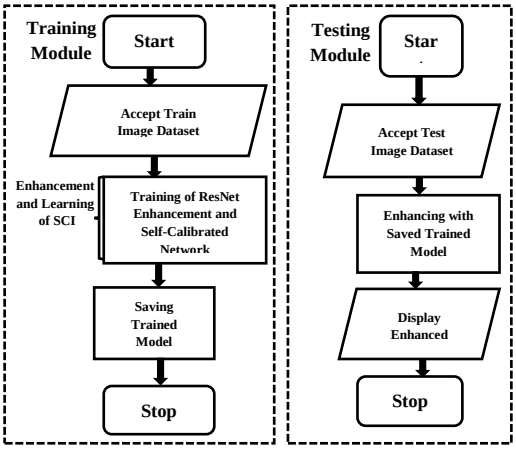



**Figure 4: System Flow Chart**

In this study, LOL image enhancement is proposed and the system is based on CNN by adding self-calibrated illumination system. There has been many CNN based on techniques. Among these techniques, this system uses the ResNet architecture enhancement network and learning of self-calibrated illumination (SCI) network. As the first step of the system, the user should be downloaded and installed necessary of applications. The networks are trained using the low-light dataset. After the training step, this system saved trained model. And the testing step, the system tested with the saved training model. Finally, this system displays the high resolution of low-light image enhancement.

### 4.1. Requirements of the System Processing

The following steps are requirements of the system processing in low-light image enhancement.

- Collect the low- light image
- Apply the system requirements of CNN networks
- Train the system enhance model
- Test the system enhance model
- Processor Intel core i3
- RAM 4GB
- Intel graphics 620

### 4.2. Low-Light Image of Dataset

These system uses low-light image dataset. Low-light images especially consist of low-light regions and images captured low-light conditions often display characteristics like low illumination, low contrast, and low gray range. LOL dataset includes dimensions (width 1080 pixels x height 720pixels) PNG file type low-light images. The LOL dataset contains 6,000 low light images captured such as parks, bridges, humans, streets, faces, things, the nighttime, and so on.

**Figure 5: Samples from LOL Dataset**

LOL dataset of 3,945 is used for training and 2,150 is used for testing. This system is used for testing with standard LOL dataset 2,150, Camera dataset 386 and Black&White dataset 150. Figure 5 shows the sample images of LOL dataset and Figure 6 shows the sample images of camera dataset. In addition, balck&white images are used for testing. Some images of black&white dataset are shown in Figure 7.

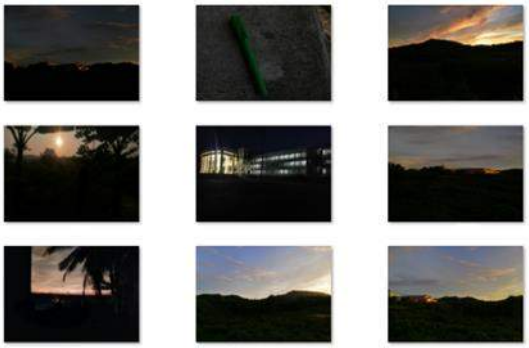


**Figure 6: Sample Images Captured by Camera**

**Figure 7: Sample from Black&White Dataset**

### 4.3. Implementation of the System

The system is implemented by using Python programming language and by using ADAM optimizer [19] with the parameters $\beta_1 = 0.0666$, $\beta_2 = 0.00009999$ and $\epsilon = 10^{-8}$, ADAM optimization is addition to stochastic gradient decent and can be used in place of classical stochastic gradient descent to increase addition network weights better efficiently. $H_\theta$ is the output of the hypothesis function.

**Table 1. Different settings for $H_\theta$ of blocks and channels on LOL dataset testing**

| Setting for $H_\theta$ | | Quality | Efficiency | |
|---|---|---|---|---|
| Blocks | Channels | PSNR | Delta | Time(s) |
| 1 | 3-3 | 54.9074 | 0.1000 | 0.0600 |
| 2 | 3-3-3 | 54.8809 | 0.1000 | 0.0680 |
| 3 | 3-3-3-3 | 54.7943 | 0.1000 | 0.0750 |
| 3 | 3-8-8-3 | 54.5779 | 0.0090 | 0.0870 |
| 3 | 3-16-16-3 | 54.5215 | 0.0090 | 0.0950 |

Table 1 describes the different values of setting for Blocks and Channels, quality of PSNR, and efficiency of Delta and Time(s) on LOL dataset testing. When using Block 1, Channel 3-3, and Delta 0.1000, the quality of PSNR 54.9074 at Time(s) 0.0600. When using Block 2, Channel 3-3-3, and Delta 0.1000, the quality of PSNR 54.8809 at Time(s) 0.0680. When using Block 3, Channel 3-3-3-3, and Delta 0.1000, the quality of PSNR 54.7943 at Time(s) 0.0750. When using Block 3, Channel 3-8-8-3, and Delta 0.0090, the quality of PSNR 54.5779 at Time(s) 0.0870. When using Block 3, Channel 3-16-16-3, and Delta 0.0090, the quality of PSNR 54.5215 at Time(s) 0.0950. Figure 8 displays the comparisons different cases on the LOL dataset using Table 1.

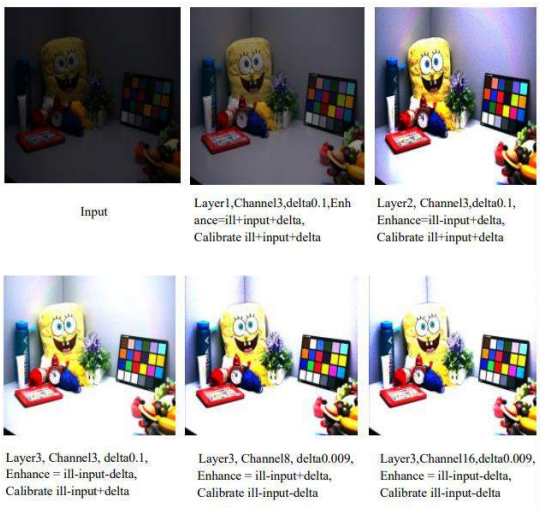



**Figure 8: Comparison Results of LOL datasets based on different parameters**

Figure 9 shows the comparisons results testing with the Camera dataset using Table 1. And Then, Figure 10 is the comparisons different cases on the Black&White dataset using Table 1.


**Figure 9: Comparison Results of Images Captured by Camera based on different parameters**

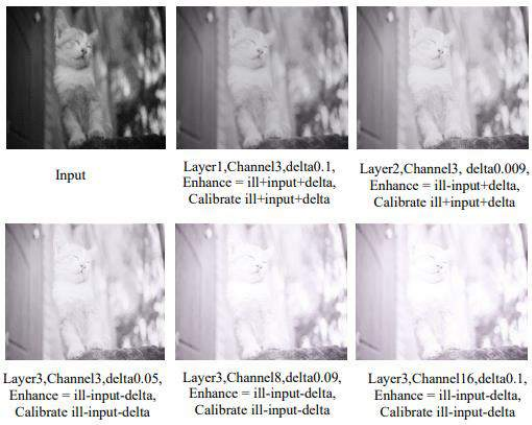



**Figure 10: Comparison Results of Images on the Black&White dataset based on different parameters**

## 4.4. Experimental Results

For the proposed system, the several experiments were conducted with enhanced model using three types of datasets (LOL, real camera, and black&white) on the Anaconda Prompt (Anaconda3). Figure 11 shows the main page of the proposed system.


**Figure 11: The main page of the proposed system**

In Figure 12 is after enhancement result using ResNet Enhancement on LOL dataset when the button of "ResNet" clicked.


**Figure 12: Using ResNet Enhancement on LOL dataset**

Figure 13 shows after enhancement using SCI Enhancement on LOL dataset when the button of "SCI" clicked.

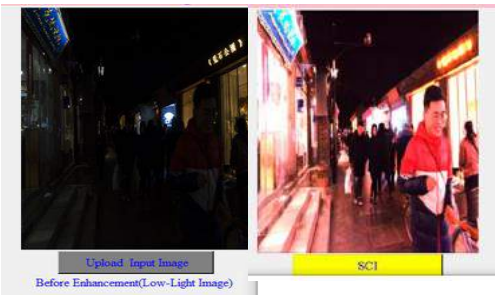



**Figure 13: Using SCI Enhancement on LOL dataset**

Figure 14 describes the final result changing processing after enhancement processing of ResNet and SCI using LOL dataset on the main window when above the label "After Enhancement (High Resolution Image)" of the button "Result Enhancement Image" clicked.


**Figure 14: After enhancement processing of ResNet and SCI using LOL dataset**


**Figure 15: After enhancement processing of ResNet and SCI using Camera dataset**

Figure 16 displays the final result changing processing after enhancement processing of ResNet and SCI using Black&White dataset on the main window when above the label "After Enhancement (High Resolution Image)" of the button "Result Enhancement Image" clicked.

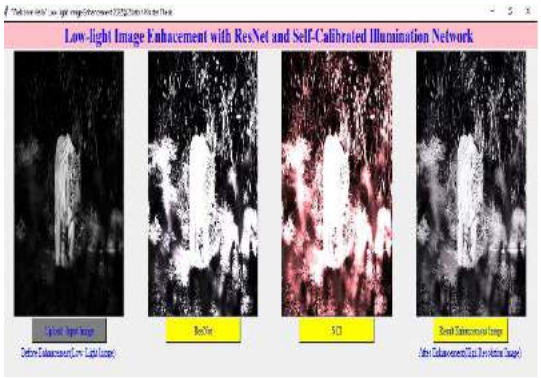



**Figure 16: After enhancement processing of ResNet and SCI using Black&White dataset**

### 4.4.1. Performance Evaluation

Performance evaluation of our SCI displayed competitive performance as shown in following Table 2.

**Table 2. Performance Evaluation results in terms of PSNR, SSIM and Accuracy on the LOL datasets**

| Dataset | Metrics, Accuracy % and Time | Unsupervised Learning Methods | | |
|---|---|---|---|---|
| | | ResNet Enhance-ment | SCI Enhance-ment | ResNet+SCI Enhance-ment |
| LOL | PSNR (dB) | 52.4207 | 51.6648 | 54.7943 |
| | SSIM | 0.9328 | 0.9216 | 0.9547 |
| | Accuracy % | 83% | 81% | 85% |
| | Time | 0.8317 | 0.8425 | 0.9715 |

Table 2 describes the performance evaluation results on the LOL dataset. Like metrics are PSNR and SSIM. Unsupervised learning methods are ResNet, SCI and ResNet + SCI Enhancement. When using ResNet, the value of PSNR is 52.4207, SSIM is 0.9328, Accuracy is 83% and Time is 0.8317. When using SCI, the value of PSNR is 51.6648, SSIM is 0.9216, Accuracy is 81% and Time is 0.8425. When using ResNet + SCI, the value of PSNR is 54.7943, SSIM is 0.9547, Accuracy is 85% and Time is 0.9714. The good value of PSNR is 30dB to 60dB. Therefore, the system of PSNR value is good. The SSIM values range between 0 to 1. The good value of SSIM is 0.8 to 1. The system of SSIM value is good. Accuracy better value is over 90%. Accuracy good value is between 70% and 90%. Thus, the system of Accuracy value is good.

## 5. Conclusion

Low-Light Image Enhancement (LLIE) defines at increasing the perception or interpretability of an image captured in an environment with low-light. The enhancement system is used to make it easier for visual interpretation and understanding of imagery. The low-light images are enhancement by integrating CNN based ResNet model and SCI framework. The result of the enhanced image of low-light image is more effective by integrating of the ResNet and SCI. In this study, proposed integrated system is achieved to get brighter image from low-light images. It improves in value the illumination and the details of the low-light images while preserving the serenity.

In the proposed system, CNN based on ResNet architecture enhancement network and the self-calibrated illumination network system has been successfully constructed combination. Like testing of three low-light dataset are used LOL, Camera and Black&White dataset. The system of LOL dataset is used standard images to train and test. Camera dataset is real environment nighttime and poor light captured image. Black&White dataset is the nearest low-light conditions image. The system is used Camera and Black&White data to test. Finally, the changing different of input low-light image compare displayed the output of high-resolution image.

## References


[1] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. Computational Visual Media, 5(2):117–150, 2019. 1

[2] CNN:https://cs231n.github.io/convolutional-networks/

[3] Fu X, Zeng D, Huang Y, et al. "A weighted variational model for simultaneous reflectance and illumination estimation," IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 2782-2790.

[4] https://datasets.activeloop.ai/docs/ml/datasets/lol-dataset/

[5] Guo X, Li Y, Ling H. "LIME: Low-Light Image Enhancement via Illumination Map Estimation, vol. 26, no.2, 2017.

[6] Jiwon Kim, Jung Kwon Lee and Kyoung Mu Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646-1654

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: "Deep Residual Learning for Image Recognition, Dec, 2015.

[8] Lim, J., Kim, J. H., Sim, J. Y., Kim, C. S., 2015. Robust contrast enhancement of noisy lowlight images: Denoising-enhancement-completion. In: Image Processing (ICIP), 2015 IEEE International Conference on. IEEE, pp. 4131–4135.

[9] Qingnan Fan, Jiaolong Yang, David Wipf, Baoquan Chen, and Xin Tong. Image smoothing via unsupervised learning. ACM Transactions on Graphics), 37(6):1–14, 2018.

[10] ResNet:https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/

[11] Shen, L., Yue, Z., Feng, F., Chen, Q., Liu, S., Ma, J., 2017. Msr-net: Low-light image enhancement using deep convolutional network. arXiv preprint arXiv: 1711.02488.

[12] Wei Chen, Wenjing Wang, Wenhan Yang, and Jiaying Liu.''Deep retinex decomposition for low-light enhancement''.

[13] Zhang, Wei (1988). "Shift-invariant pattern recognition neural network and its optical architecture". Proceediings of Annual Conference of the Japan Society of Applied Physics.

[14] Zhang K, Zuo W, Chen Y, et al. "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising", IEEE Transactions on Image Processing, 2017.

# OBJECT DETECTION AND DISTANCE ESTIMATION USING YOLO ARCHITECTURE

May Thu Aung, Khaing Khaing Wai
*University of Computer Studies, Yangon*
*maythu.aung@ucsy.edu.mm, khaingkhaingwai@ucsy.edu.mm*

## Abstract

*Detecting various classes of objects and measuring the distance between camera and objects are used in this system. The input of the system is an image or video which are captured by the camera. YOLO-v5 model detects the objects of input image and calculate the distance between camera and detected objects. If the object is detected, the output result will be the label with distance meter values.*

**Keywords**: Object detection, distance estimation, YOLOv5 model

## 1. Introduction

One of the computer vision tasks that detects things of a specific type within an image is object detection. One-stage methods and two-stage methods are the two primary categories under which it can be divided. One-stage techniques like YOLO, SSD, and RetinaNet focus on inference speed. Two-stage approaches prioritize accurate detection, and Faster R-CNN, Mask R-CNN, and Cascade R-CNN are three examples of such models [14]. The MSCOCO dataset is the most well-liked one. Usually, a Mean Average Precision metric is used to evaluate models. Distance estimation is employed as a crucial component of recognition and orientation in real time, as well as enabling various gadgets to move automatically in the actual world.

YOLOv5 is utilized in this system to detect objects, classify images, and determine the distance between the camera and the item. Rather than starting with image pixels and working its way down to bounding box coordinates and class probabilities, it reframes object detection as a single regression problem. This integrated model predicts multiple bounding boxes and class probabilities for items covered by boxes simultaneously [11].

## 2. Related Work

In the paper [2] uses the YOLOv3 to predict the absolute distance of objects using only information from a monocular camera and design the two ways of measuring the distance, class-agnostic and class-aware. Class-agnostic creates smaller prediction vectors than class-aware and achieves better results. In this paper, KITTI dataset is used and show the distance range within [0, 150] m. Uses cameras instead of LIDARs to present the possibility for distance estimation in the paper [15]. This paper is based on the YOLOv3 deep neural network and principles of stereoscopy. In this paper, uses two slightly moved cameras to get two pictures which goes through algorithm for stereoscopy-based measurement and estimate distance to detected objects. In the paper [1] uses the YOLOv5 model to measure the distance between objects for processing real-time images with OpenCV in order to restrict the distance between several people in the same space and also add Euclidean distance calculation method in DeepSORT and OpenCV to minimize occlusion. Detecting the distance between people and using the open-source COCO dataset for learning in this paper.

## 3. Background Theory

Image Processing based object detection and distance measuring between camera and object of an input from camera using one-stage detector is implemented with YOLOv5 model in this system.

### 3.1. You Only Look Once

The system generates prediction vectors corresponding to each object present in the input image after it has been passed through a single neural network of multiple convolutional networks. The YOLO system computes all the

features of the image and produces predictions for all objects at once, as opposed to iterating the process of classifying various parts on the image [11].

There are various versions of YOLO algorithm. One version of YOLO algorithm is related with another. In this system, YOLOv5 is used for object detection and distance measurement between camera and object of detected image or video.

### 3.2. Architecture of YOLOv5 Model

YOLOv5 is a popular real-time object detection tool that offers high accuracy and quick detection speeds. The overview architecture of YOLOv5 is shown in Figure 1.

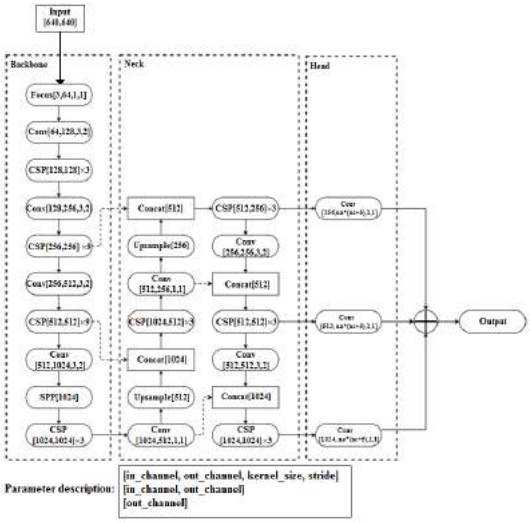



**Figure 1: Overview Architecture of YOLOv5 Model**

The architecture of the YOLO-v5 has three parts;

(i) Backbone: Focus structure and CSP network
(ii) Neck: SPP block and PANet
(iii) Head: Output using GIoU-Loss

First, focus structure is a layer that splits the image into layers and further divides the layers of the divided image [1]. Through multiple downsampling using CSP-Darknet53, Backbone can efficiently extract feature information from the input image. The cascade structure of FPN and bottom-up PANet are used by the neck to combine the image features extracted by the backbone. Three output branches in the head anticipate the bounding boxes and distinct types of item categories [16].

An S × S grid is used to divide up the input image. YOLOv5s predicts B boundary boxes for each grid cell. Each bounding box has three categories of parameters: object confidence C, prediction probabilities P of n classes, and the center coordinate (x, y), width, and height of the box. Consequently, as stated in Equation 1, the loss function is made up of the bounding box position loss, object confidence loss, and class probability loss. The cross-entropy loss function computes the object confidence loss and the class probability loss [16].

$$Loss = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} L_{GIoU} - \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [\hat{C}_j^i \log(C_j^i) (1 - \hat{C}_j^i) \log(1 - C_j^i)] - \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} [\hat{C}_j^i \log(C_j^i) + (1 - \hat{C}_j^i) \log(1 - C_j^i)] - \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \sum_{C \in classes} [\hat{P}_j^i \log(P_j^i) + (1 - \hat{P}_j^i) \log(1 - P_j^i)] \quad (1)$$

where $I_{ij}^{obj}$ is defined as 1 if object presents inside j-th predicted bounding box in i-th cell, and 0 for otherwise. $I_{ij}^{noobj}$ is the opposite. λcoord and λnoobj are the loss weights [16].

#### 3.2.1. CSP- Darknet53

The CSP-core Darknet53's design, known as DenseNet, utilizes the prior input and concatenates it with the current input before proceeding into the dense layer (CSP stands for Cross Stage Partial). In order to address vanishing gradient issues, DenseNet was created to connect layers in a very deep neural network (as ResNet) [8].

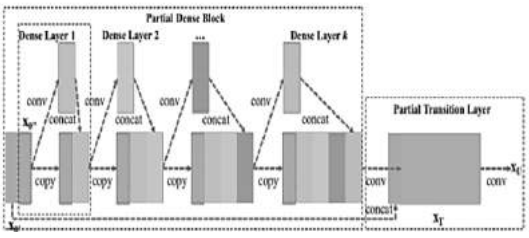



**Figure 2: Process of Input Processing in a CSP Dense Block**

### 3.2.2. Spatial Pyramid Pooling Block (SPP Block)

The Spatial Pyramid Pooling block (SPP block) concatenates these 3 feature maps pooled with the sizes of $size_{fmap} \times size_{fmap} \times 512$ and 30 including the input feature maps to avoid loss of significant features in the case that 3-scale max-pooling is insufficient. The feature maps were converted to a one-dimensional vector after conducting multi-scale max-pooling. As a result, the input retained the spatial dimension in addition to extracting the key aspects that facilitated training.

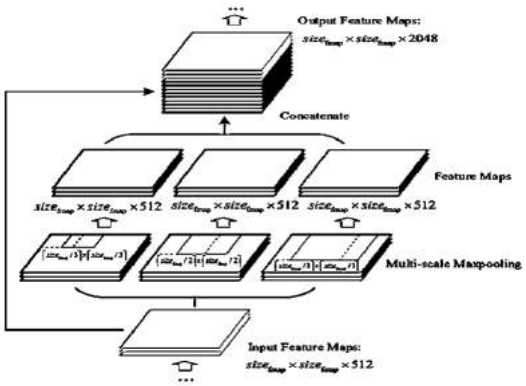



**Figure 3: SPP Block Adapted to YOLO**

### 3.2.3. Feature Pyramid Network (FPN)

The neck of YOLOv5 uses the Feature Pyramid Network (FPN) architecture to preserve these fine-grained features. FPN used top-down method to transmit semantical information (from the high-level layer) and concatenate them to fine-grained features (from the low-level layer in the backbone) for predicting small objects in the large-scale detector [4].

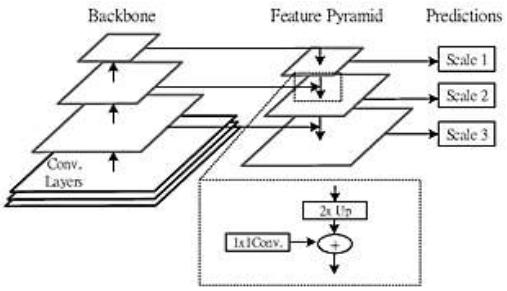



**Figure 4: FPN Architecture**

Due to the top-down flow in the FPN architecture, Path Aggregation Network (PAN) is a more sophisticated variant of FPN. The large-scale detector from low-level layers in FPN is able to concurrently receive the semantic data from high-level layers and fine-grained information from low-level layers. Small-scale detector of the FPN limits the application of object detection to semantic features. The idea of concatenating semantic features and fine-grained features at high-level layers was taken into consideration to enhance the performance for the small and medium-scale detector. There are many layers, possibly over 100, in the deep neural network's backbone. As a result, the fine-grained features in FPN must travel a lengthy distance to get from low-level to high-level layers. In addition to the top-down augmentation method utilized in FPN. A bottom-up augmentation path is proposed for PAN architecture. The direct connection of fine-grained features from lower-level layers to the top ones was thus made possible. This shortcut has fewer than ten layers, which facilitates easy information flow. Figure 5 depicts the PAN architecture's overall process.

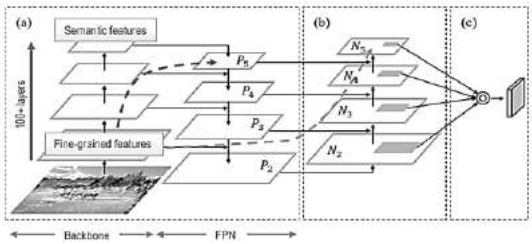



**Figure 5: PANet Architecture Including (a) FPN Backbone, (b) Bottom-up Path Augmentation, (c) Adaptive Feature Pooling**

## 3.3. Distance Estimation

At first, find the focal length of the camera to estimate the detected object distance from bounding box's width and height which are get from YOLOv5 object detector using triangle formula.

$$\text{Focal length} = \sqrt{w^2 + h^2} \quad (2)$$

where, w = bounding box's width
h = bounding box's height

Then, insert into torch library to get the distance between camera and detected objects layer by layer.

## 4. The Proposed System Flow and Experiment

### 4.1. System Flowchart

To perform object detection, the camera captures the image or video, focus structure of the YOLOv5 model divides the input into the layers and then extracts the feature information using CSP-Darknet53 in the backbone layer. In the neck layer, aggregates the image features which are coming from the backbone layer using FPN and bottom-up PANet. Then, apply anchor boxes on feature maps and combine the outputs as the final result in the head layer. For calculate distance, uses the triangle formula to get the focal length. The values of bounding box width (w) and height (h) are get from YOLOv5s model. After that compute the distance meter between camera and detected object using focal length value in Torch.tensor metrics. Then, the final detected result will be with class labels and distance meters. The flowchart of the system is given in Figure 6.

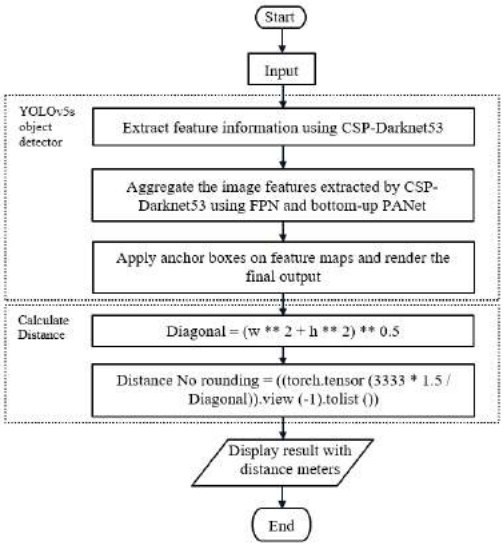



**Figure 6: Flowchart of the System**

### 4.2. Experimental Results of the System

To test the system, 800 images with 80 different object classes were used. Figure 7 shows the detection of two types of objects. They are a car and a motorcycle, respectively.

**Figure 7: Detected Image Result 1**

In Figure 8, bicycle, bench, potted plant and person are detected.

**Figure 8: Detected Image Result 2**

There are five class of objects are detected in Figure 9 (person, chair, laptop, cup and dining table).

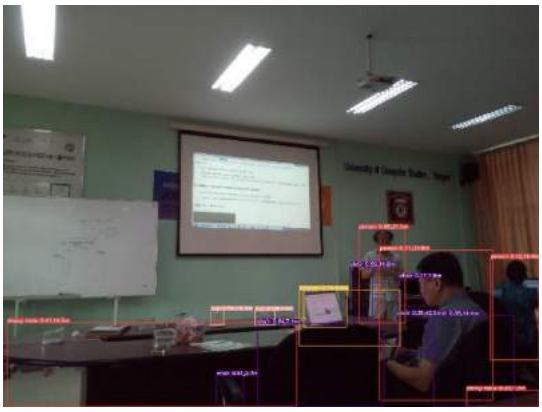


**Figure 8: Detected Image Result 3**

This system also detects the video file (.mp4) and the example of vehicles video result is described in Figure 6 (a), (b) and (c).


**Figure 10 (a): Example of Detected Video Result**

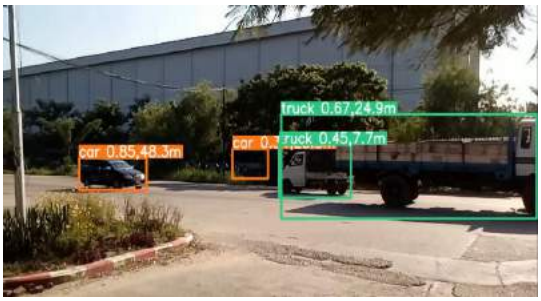



**Figure 10 (b): Example of Detected Video Result**


**Figure 10 (c): Example of Detected Video Result**

## 4.3. Performance Evaluation of the System

Precision is a measure of how many correct positive predictions are made (true positive). The formula is:

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{all\ detections} \quad (3)$$

Determine the precision values of the test set to learn about its performance. The precision values of the test set are displayed in the figures below. To better understand the precision values, a test set with 10 object classes was created. When the confidence score is 0.915 for test set 1, the precision values for all classes will be 1 in Figure 11.

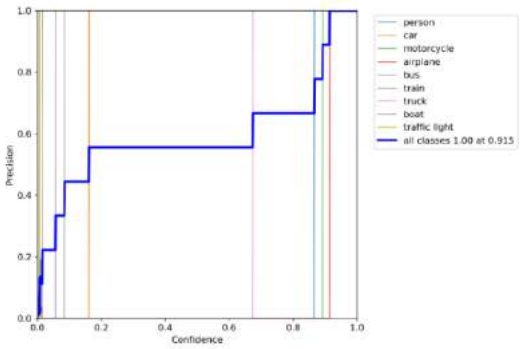



**Figure 11: Precision Curve of the Test Set 1**

In Figure 12, the precision values for all classes are 1 and the confidence score is 0.696 for the test set 2.

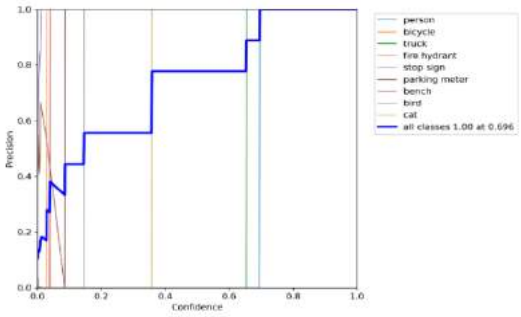



**Figure 12: Precision Curve of the Test Set 2**

When all precision values from all classes are added together, the result is illustrated in Figure 13.

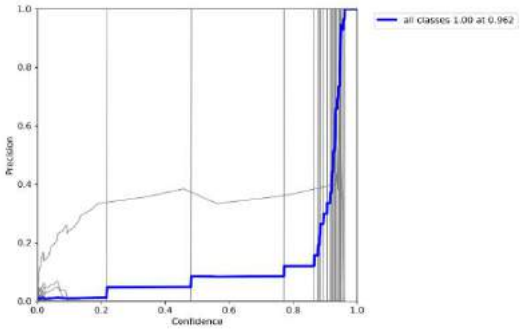



**Figure 13: Precision Curve of All Test Set**

Recall is a metric for how many out of all the positive cases in the data that the classifier correctly predicted.

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{all\ ground\ truths} \quad (4)$$

Precision Recall curves for the test set 1 and 2 are shown in Figure 14 and 15.

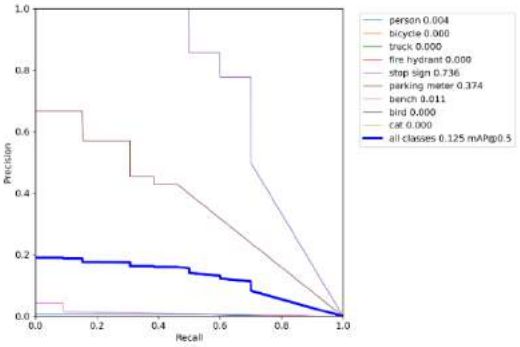



**Figure 14: Precision Recall Curve of the Test Set 1**

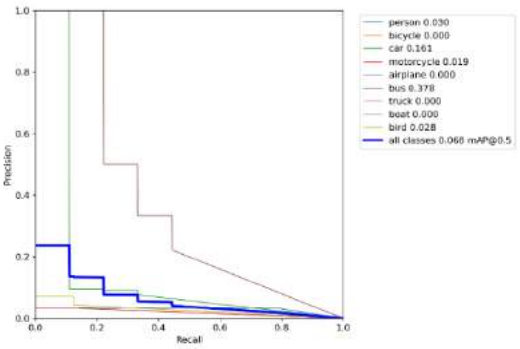



**Figure 15: Precision Recall Curve of the Test Set 2**

However, when all precision values from all classes are added together, the result is not clear and will be shown in figure 16.

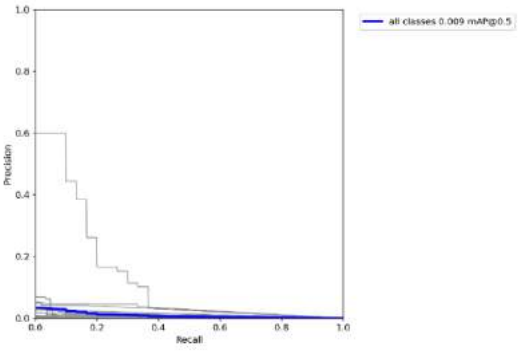



**Figure 16: Precision Recall Curve of All Test Set**

## 5. Conclusion

In summary, AI based computer vision tasks are become popular in nowadays. Object detection is one of the famous computer vision task to apply in various applications such as healthcare, security surveillance and self-driving cars. Distance estimation is combined together with object detection so that to improve driving system of AI based autonomous vehicles and remote controls. YOLOv5s is most suitable for real-time detecting objects and measure the distance between camera and its detected objects because of its higher performance and accuracy.

### 5.1. Advantages

This system is only base on the single camera of the device. Therefore, it can easily use in any operation system and no need to install other external devices such as LiDAR or other cameras. This system can detect various classes of objects.

### 5.2. Limitation

Distance meter values can vary according to the focal length because distance calculating formula is based on the device camera's focal length. The greater the number of cameras lenses, the more accurate the distance meter values.

### 5.3. Further Extension

The system functions can be added distance accuracy correction. When the distance is too close, a warning message will be issued. Another extension is not only can add emoticons but also vehicle speed per hour and target object speed per hour. If one of the speed per hour of the vehicle and the speed of the target exceeds the upper limit, an early warning will be given.

## References

[1] A study on object distance measurement using OpenCV-based YOLOv5, International Journal of Advanced Culture Technology, Vol.9 No.3 298-304 (2021), DOI.

[2] Dist-YOLO: Fast Object Detection with Distance Estimation, Appl. Sci. 2022, 12, 1354.https://doi.org/10.3390/app12031354 by Marek Vajgl, Petr Hurtik and Tomáš Nejezchleba

[3] Gochoo, M. (2020). ReseachGate. Search date 03.12.2020. researchgate.net: https://www.researchgate.net/figure/a-Feature-pyramid-network-FPN-b-YOLO3-c Proposed-concatenated-feature-pyramid_fig2_335538302

[4] https://blog.superannotate.com/introduction-to-computer-vision/

[5]https://docs.ultralytics.com/tutorials/architecture-summary/

[6]https://github.com/o920130130/YangSongbo/pulls
[7] https://iq.opengenus.org/yolov5/
[8] Huang, G., Liu, Z., & Maaten, L. v. (2018). Densely Connected Convolutional Networks. arXiv. Seach date 28.11.2020. https://arxiv.org/pdf/1608.06993.pdf
[9] Hui, J. (2020). YOLOv4. Medium. Seach date 27.11.2020. https://jonathan-hui.medium.com/yolov4-c9901eaa8e61
[10] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. arXiv. Seach date 03.12.2020. https://arxiv.org/pdf/1803.01534.pdf.
[11] Redmon, J., & Farhadi, A. (2016). YOLO9000: Better, Faster, Stronger. arXiv. Seach date 17.11.2020. https://arxiv.org/pdf/1612.08242.pdf
[12] Solawetz, J. (2020). Breaking Down YOLOv4. Roboflow. Seach date 27.11.2020. https://blog.roboflow.com/a-thorough-breakdown-of-yolov4/
[13] V Thatte, A. (2020). Evolution of YOLO — YOLO version 1. Medium. Seach date 14.11.2020. https://towardsdatascience.com/evolution-of-yolo-yolo-version-1-afb8af302bd2
[14] Wang, C.-Y., Mark Liao, H.-Y., Yeh, I.-H., Wu, Y.-H., Chen, P.-Y., & Hsieh, J.-W. (2019). CSPNET: A new backbone that can enhance learning capability of CNN. arXiv. Seach date 30.11.2020. https://arxiv.org/pdf/1911.11929.pdf
[15] YOLO Multi-Camera Object Detection and Distance Estimation, DOI: 10.1109/ZINC50678.2020.9161805 by Bojan Strbac, Marko Gostovic, Zeljko Lukac and Dragan Samardzija
[16] Zheng, G.; Zhao, J.; Li, S.; Feng, J. Zero-Shot Pipeline Detection for Sub-Bottom Profiler Data Based on Imaging Principles. Remote Sens. 2021, 13, 4401. https://doi.org/10.3390/rs13214401

# Face Mask Detection by Using Convolutional Neural Network

Ei Cherry Lwin, Myat Mon Kyaw
*University of Computer Studies, Yangon*
*eicherrylwin@ucsy.edu.mm, myatmonkyaw.fis@ucsy.edu.mm*

## Abstract

*The proposed system is designed to classify people who is wearing face masks or not. The model used in this system is MobilenetV2, a convolutional neural network (CNN). The image dataset contains 7553 images.3832 images used to train model and 3721 images are used for testing. Firstly, the input images are needed to be processed. Resizing, One-hot Encoding and data Augmentation are applied in preprocessing. The model is constructed with MobilenetV2 model. If a person in an image is wearing a mask, the system displays the face region with a green anchor box. If a person in an image is not wearing a mask, the face region is displayed with a red anchor box. This system can merge at airports, railway stations, workplaces, schools, and other public places for safety. The accuracy is 82 % for testing images in dataset.*

**Keywords**—Convolutional Neural Network, MobilenetV2.

## 1. Introduction

Coronavirus is continuously spreading until now everywhere on the earth, and causing a serious health problem. Therefore, masks should be worn by people. And people should live with social distancing to avoid serious spread of Coronavirus.

Face mask detection has been received more attention because of spreading of corona virus. Face mask detection is crucial for Covid-19 prevention. Face mask prevents infection from a person whether they have symptoms of disease.

## 2. Methodology

MobileNetV2 is an architecture of bottleneck depthseparable convolution building of basic blocks with residuals [7]. MobileNetV2 has two types of blocks. All blocks contain three layers. The first one is 1x1 convolutions with “ReLU6” and a onestride residual block. The second one contains depth-wise convolution, a residual block with stride 2 and is used for shrinking. The last one consists of a 1x1 “convolution” with no non-linearity. The methodology of object detection makes a classification to determine the input class and to adjust the bounding box. Most backbone networks for detection except the last completely connected layer are classification networks. The backbone network can be used as a simple feature extractor for object detection tasks to take input images and produce feature maps for each image.

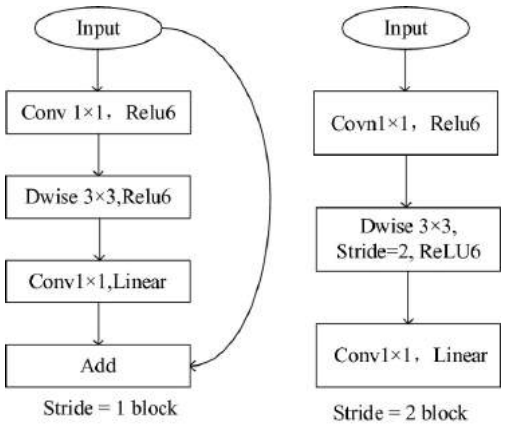



**Figure 1. Two Types of blocks of MobilenetV2**

The pretrained techniques are used to extort maps of the feature with high-quality classification problems. This part is called base model. The base model uses the “image net” weights. ImageNet is an image database. It has been trained on hundreds of thousands of images. It is very useful for categorization of images. During training, the evaluated “bounding boxes” are contrasted to the “ground truth boxes”. During backpropagation, parameters are modified as needed. The MobileNet contains two parts. These are a classifier and a base model [6].

## 3. Overview of the proposed system

The aim of the proposed system to detect a face mask in an image by using Convolutional Neural Network (MobilenetV2) method. This system can detect whether a person is wearing a face mask or not. After detecting the location of the face, the status whether face has a mask or not need to be determined.

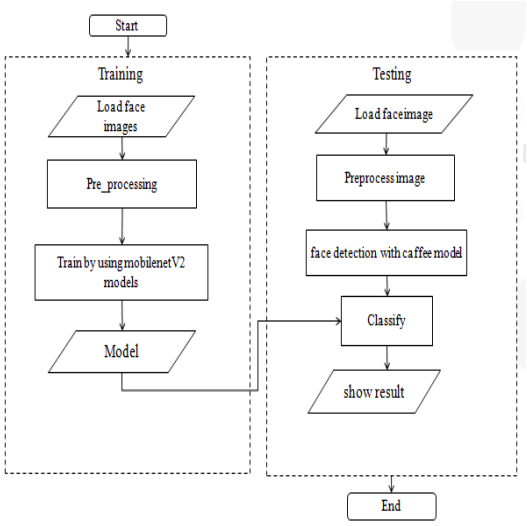



**Figure 2. Flow Diagram**

### 3.1. Dataset Description

The image dataset from Kaggle is applied in this system. It has two classes. The first one is image with people who wear masks. The second one is image with people who do not wear masks. This dataset consists of 7553 images. The first class is 3725 images and the second class is 3828 images [8].

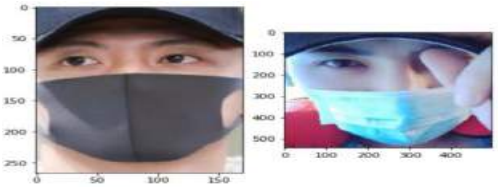


**Figure 3. Images In which People Wear Masks**

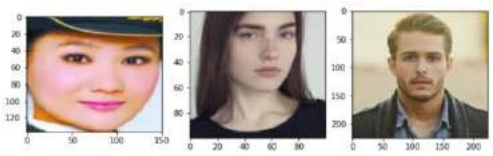


**Figure 4. Images In which People Do Not Wear Masks**

### 3.2. Data preprocessing

The images are resized into 224x224. Image resizing is n import and step in preprocessing. To perform One-Hot Encoding, Firstly, the image is transformed into a NumPy array to perform one-hot encoding. Encoding categorical data into vectors 0s and 1s is applied in one-hot encoding. Images with masks are encoded to 0 and without masks are encoded to 1. Random rotation augmentation rotate randomly the images from 0 to 360 degrees in clock wise direction. In data augmentation, firstly the image is rotated with random rotation augmentation. The zoom augmentation method is used for image zooming by making zooming in or adding some pixels around the image to enlarge the image. After rotation, the image is zoomed. Finally, the image is flipped horizontally. Horizontal flip augmentation is reversing the entire rows and columns of an image pixels in horizontally.

### 3.3. Training Model

Firstly, the base model MobilenetV2 is loaded with “ImageNet". Then the dimension is flattened and the dense activation value and dropout value are launched into. Then the dense and activation function value are entered. The final layer is fine tuned. The Model summary is obtained and configuration is preserved. The optimizer loss entropy and accuracy metric are configured and the trained model is assessed and maintained.

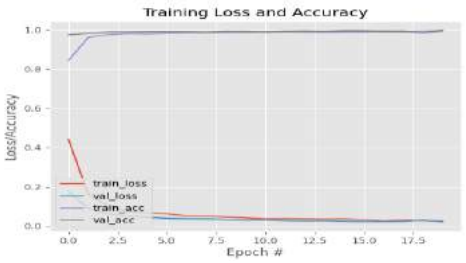



**Figure 5. Training Loss and Accuracy**

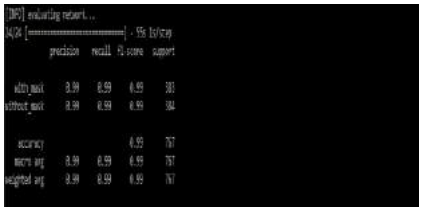


**Figure 6. Classification Report**

## 4. Implementation

The proposed system implemented a training model with MobilenetV2, Convolutional Neural Network. The testing images can be browsed from the dataset and from real world such as google, real time captured images. Pre-processing is carried out for resizing, One-Hot Encoding and data augmentation. The face region is detected with Caffe face detector model. And then an input image is classified with the trained model.

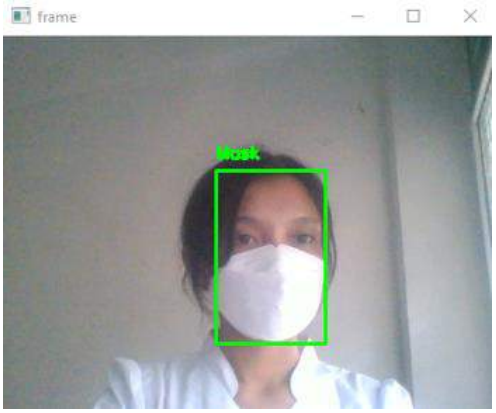



**Figure 7. Result of Predicting input data**

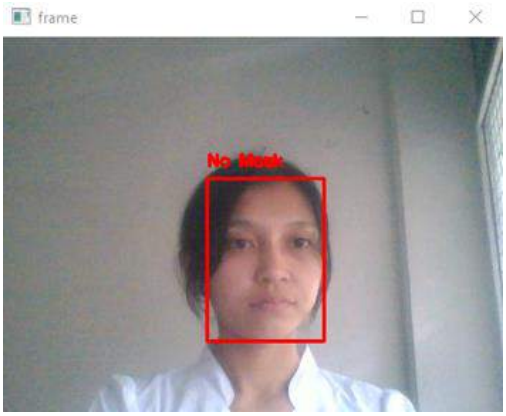



**Figure 8. Prediction result for image from camera without mask**


**Figure 9. Prediction result for image from dataset without mask**

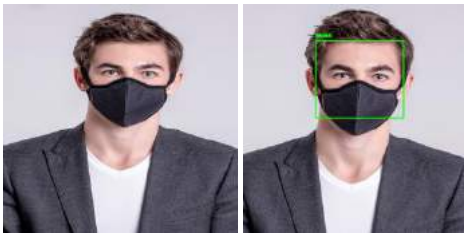


**Figure 10. Prediction result for image from dataset with mask**

### 4.1. Experimental results

Total images = 201 images
True Positive (TP) = 71,
True Negative (TN) = 93,
False Positive (FP) = 29,
False Negative (FN) = 8

Accuracy = (TP + TN)/ (TP + TN + FP + FN) = (71+93)/201 = 0.82
Precision = TP/(FP + TP) = 71/(29+71) = 0.71
Recall = TP/(FN+TP) = 71/(8+71) = 0.90

## 5. Conclusion

The proposed system aims to classify face mask wearing or not for COVID-19 precaution. The proposed system can classify a person in an image who is wearing mask or not and can be applied in crowded areas. The primary intention is to detect face region and then to classify face mask wearing or not. 3832 images are trained with training model and 201 images are applied for testing. The accuracy is 82 % for testing images in dataset.

### 5.1. Limitation of the System

The system cannot recognize precisely all face regions if an image has a large number of people that is large number of face regions.

### 5.2. Further Extensions

The system can be extended by adding messaging system with mobile phone for efficient and fast notification. Adding alarm signal for alerting the person in the real can be extended. Performing and adding new classification and detection model for getting better workflow of the system.

## References


[1] W.H.O., “Coronavirus disease 2019 (covid-19): situation report, 205”.

[2] “Coronavirus Disease 2019 (COVID-19) – Symptoms”, Centers for Disease Control and Prevention, 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/symptoms testing/symptoms.html. 2020.

[3] Ge S., Li J., Ye Q., Luo Z, "Detecting Masked Faces in the Wild with LLE-CNNs," IEEE Conference on Computer Vision and Pattern Recognition, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8099536.

[4] M. R. I. M. S. a. A. S. M. S. Ejaz, "Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition," 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019. [Online].Available: https://ieeexplore.ieee.org/document/8934543.

[5] M. Loey, G. Mangogaran, T. M.H.N. and K. N.E.M., "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," National Library of Medicine, 1 January 2021. [Online].Available:https://pubmed.ncbi.nlm.nih.gov/32834324/

[6] Haddad, J., 2020. How I Built A Face Mask Detector For COVID-19 Using Pytorch Lightning. [online] Medium. Available at: https://towardsdatascience.com/how-i-built-a-face-mask-detectorfor-covid-19-using-pytorch-lightning-67eb3752fd61.

[7] M. Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov,and Liang-Chieh Chen, “MobileNetV2: Inverted Residuals andLinear Bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR), 2018.

[8] O. Gurav, "Face Mask Detection Dataset" [Online]. Available: https://www.kaggle.com/datasets/omkargurav/face-mask-dataset.

# Mungbean Leaf Disease Detection Using K-Nearest Neighbor Algorithm

*Hnin Pwint Zaw, Khant Kyawt Kyawt Theint*
*University of Computer Studies, Yangon*
*hninpwintzaw@ucsy.edu.mm, khant2theint@gmai.com*

## Abstract

*Disease detection is very important part to protect loss of crop in agriculture. Symptoms of the plant diseases can be detected by using machine learning techniques. Machine learning technique can solve for classification and regression problems. This paper proposed mungbean leaf disease detection by using digital image processing and machine learning techniques. Image enhancement technique is used in the image preprocessing state to improve image quality. And also, mungbean leaf diseases are classified using the k-NN algorithm, grey level co-occurrence matrix (GLCM) for feature extraction, and k-means clustering to segment region of interest in leaf area. According to the results of the experiments, the system can successfully detect and classify healthy and unhealthy or infected leaf areas.*

**Keywords:** Image processing, feature extraction, k-Nearest Neighbor (K-NN), k-means Clustering.

## 1. Introduction

Myanmar is an agricultural country because agriculture is the primary source of income for the Burmese people. One factor influencing agricultural productivity is a disease outbreak. Farmers must therefore accurately identify the type of disease and treat it as soon as possible. The leaf is the most important part of the plant to inspect for plant diseases. It is critical to accurately detect and classify leaf diseases in order to prevent agricultural losses. Mungbean is also one of the most important plants and seeds in the world, whether dried or fresh. Mungbeans are a high-protein food with numerous health benefits. Furthermore, this is the most economically important bean on the planet, providing protein to millions of people as well as a variety of other products. Mungbean diseases, such as angular leaf spot and bean rust, stifle production. To solve the problem at an early stage, an accurate classification of leaf diseases is required. Using the Plant Village dataset of leaf images, a machine learning approach is proposed to identify and classify bean leaf diseases. In machine learning, there are numerous classification approaches. This system uses the k-nearest neighbors (KNN) algorithm to determine whether a leaf is healthy or unhealthy. This algorithm can accurately determine the suffer area of a leaf by analyzing the symptoms of the image.

In this paper, section 1 describes an introduction, section 2 related work, section 3 background theory, section 4 proposed system, section 5 experimental result and performance evaluation, section 6 result and discussion, and section 7 conclusion of the proposed system.

## 2. Related Work

In [1], Beef Image Classification Using K-Nearest Neighbor Algorithm for Identification Quality and Freshness was presented by the author. The experiment results in the ability of the system that detects meat quality based on color and texture to detect the type of beef. In [2], the author presented Plant Leaf Disease Classification and Detection System Using Machine Learning Technique, which focuses on developing an advanced and efficient system that makes the process of producing high yields of tomato much easier for farmers. In [3], Leaf Disease Detection and classification K-means segmentation and neural-network-based classification were demonstrated by the author. Because of the use of these techniques, which perform well in all types of leaf diseases sampled

and can detect and classify the examined diseases with 93% accuracy.

## 3. Background Theory

Machine learning algorithms are classified into four types: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. It is gaining popularity in a wide range of fields, including disease diagnosis in health care. Traditional diagnostic methods are costly, time-consuming, and frequently require human intervention.

Traditional diagnosis techniques are limited by the individual's ability, whereas machine learning-based systems are not, and machines do not exhaust themselves as humans do. As a result, in health care, a method for diagnosing disease in the presence of an unexpectedly large number of patients may be developed. The term "supervised learning" refers to the presence of a teacher in the form of a supervisor. Essentially, supervised learning is a type of learning in which the machine is taught or trained using data that has already been labeled with the correct answer. This algorithm recognizes the input pattern and produces the expected output.

The learning result from predicting new data should be an accurate prediction rule. Using historical data as input, machine learning algorithms predict new output values. Machine learning is similar to "programming by example." Because they are data-driven and can examine large amounts of data, machine learning algorithms are frequently more accurate than static programming. The k-nearest neighbors (KNN) algorithm is a simple supervised machine learning algorithm for classification and regression. It is an object classification method that employs nearby learning data as well as the number of nearest neighbors, also referred to as k values.

## 4. Proposed System

The goal of this system is to classify the type of mungbean leaf disease by using the k-nearest neighbors algorithm. The proposed system includes preprocessing, clustering, texture and color feature extraction, and classification of mungbean leaf diseases based on these extracted features using KNN as a classifier. Figure 1 depicts overview of the proposed system.

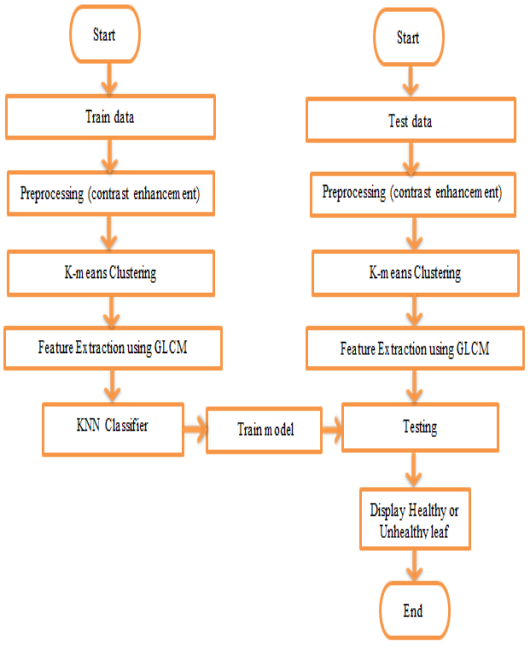



**Figure 1. Overview of the proposed system**

### 4.1. Image Preprocessing

Firstly, the input image is preprocessed using the contrast enhancement technique. This technique is used to enhance image quality by adjusting the relative brightness and darkness of objects in the scene to improve visibility. The following is the contrast enhancement equation:

$$Y = ax + b$$

Y= output value
a= contrast value
x= input value
b= brightness value

Secondly, the enhanced image is segmented by using k-means clustering techniques. K means clustering is used to create groups of observations with similar characteristics. In this proposed system, three parameters for k values are used to determine the region of interest in the leaf's affect area. The following equation describes K-means clustering:

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ \|x_n - m_n\|^2$$

Where,

N = total number of data points
K = number of clusters
$x_n$ = vector of measurement n
$m_k$ = mean for cluster k
$r_{nk}$ = an indicator variable that indicates whether to assign $x_n$ to k

### 4.2. Gray Level Co-occurrence Matrix

Gray Level Co-occurrence Matrix is used in step three to extract features from preprocessing and cluster images. The GLCM matrix will compute the probability value of the relationship between two pixels in the image with a given intensity at a given distance and orientation at a given angle. It is one of the techniques used to extract texture analysis feature and angled separate the two-pixel coordinates. Distances are represented by pixels, while angles are represented by degrees. The angular orientation will be divided into four directions with a one-pixel distance between pixels: 0°, 45°, 90°, and 135°.
Step:1 The initial GLCM matrix is made up of two-pixel pairs that line up at 0°, 45°, 90°, or 135°.
Step:2 Creating a matrix by combining the GLCM's initial matrix;
Step:3 Subtract the number of pixel pairs from each GLCM element's probability value.
Step:4 Determine the total number of extracted features for each formed direction and namely.

In this proposed system, Contrast, Correlation, Mean, Variance, Energy, Entropy and Homogeneity are used for texture feature and RMS, Smoothness, Kurtosis, S. D, IDM and Skewness are used for colors feature. The terms of the equations are by the following:

$$\text{RMS}= \sqrt{1/T \int f(x)^2 \, dx}$$

$$\text{Kurtosis}= n^* \frac{\sum_i^n (Yi-\bar{Y})^4}{\sum_i^n (Yi-\bar{y}^2)^2}$$

$$\text{Mean}= \frac{\sum_{i=1}^{N} xi}{N}$$

$$\text{Variance}= \frac{\sum (xi-\bar{x})^2}{N}$$

$$Contrast = \sum_{i,j=0}^{N-1} P_{ij} (i-j)^2$$

$$Energy = \sum_{i,j=0}^{n-1} (P_{ij})^2$$

$$Entropi = \sum_{i,j=0}^{N-1} -ln(P_{ij}) P_{ij}$$

$$Homogenity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(j-j)^2}$$

$$Correlation = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(j-j)^2}$$

### 4.3. K-Nearest Neighbor Algorithm

At final stage, k-nearest neighbor algorithm is applied based on extracted texture and color feature of the images. As a classification method, the supervised algorithm employs the K-Nearest Neighbor algorithm. K-NN is a simple algorithm that can be easily applied to machine learning algorithms to solve classification and regression problems. In classification, the KNN algorithm is used to find the value of group k on the object in the training data that is closest (similar) to the object in the testing data. The following algorithm can be used to explain how k-NN works:

Level 1: Count the number of neighbors (K).
Level 2: Calculate the distance in Euclidean terms each of the K neighbors.
Level 3: Determine the K closest neighbors using the calculated Euclidean distance.
Level 4: Count the number of information sources in each of these k categories.
Level 5: Assign the new information within the realm of the greatest number of neighbors.
Level 6: Our model is finished.

In broad terms, this algorithm is used to calculate the separation between two objects x and y using mathematical formula for Euclidean distance. Euclidean distance equation is by the following:

$$\text{d(x,y)}=\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Where,
d (x,y) : the separation between testing and training data;
x: testing data;

y: training data;
n: the number of features

## 5. Experimental result and Performance Evaluation

The accuracy of an algorithm is measured in this section by presenting experimental results images and the algorithm for determining research performance level.

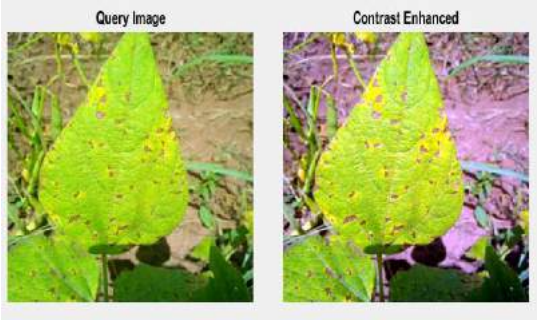



**Figure 2. Original image and Contrast Enhancement image**

Figure 2 depicted a contrast enhanced image created from an original image during the image preprocessing stage.

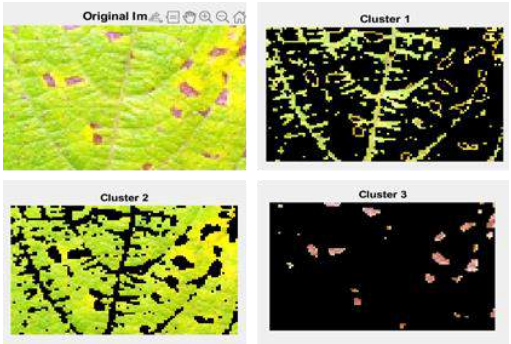



**Figure 3. Clustering Image using K-means Clustering**

Figure 3 depicts clustered images obtained by using k-means clustering to identify regions of interest in an affected area of the leaf.

Figure 4 presented that extracted feature values based on color and texture from preprocessed and clustered images.

Figure 5 presented that classification result image by using k-nearest neighbor algorithm based on extracted feature value.

As accuracy result, the confusion matrix is used to evaluate the KNN algorithm's performance. The Confusion Matrix is a table-based evaluation of a data mining classification. It contains information that compares the system's classification results with the actual classification results.

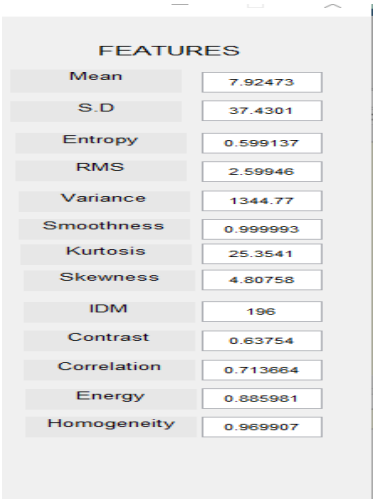



**Figure 4. Extracted texture and color feature values**

**Figure 5. Classification result image**

**Table 1. Calculation Confusion Matrix**

| Actual Class | Predicted Class | |
|---|---|---|
| | True | False |
| True | True Positive(TP) | False Negative(FN) |
| False | False Positive(FP) | True Negative(TN) |

Table 1 displays a confusion equation matrix with two labels, True and False, and four different predictive and actual value combinations.

Accuracy is a critical component in data mining and machine learning module performance because the success of a module is dependent on its accuracy because measurement accuracy shows how close it is to its true value and precision is defined as "the quality of being exact," and it refers to how close two or more measurements are to each other, regardless of

accuracy. Precision measurements can be accurate or inaccurate. Precision is defined mathematically as the number of true positives divided by the number of true positives plus the number of false positives. The following equation can be used to calculate an algorithm's accuracy and precision from the table:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

$$\text{Precision} = (TP)/(TP+FP)$$

Where,

• TP: Positive classification results from positive data.

• FN: Positive data is represented by negative classification results.

• FP: Is it possible to have negative data and still have a positive classification result?

•TN: Data that is negative but has a positive classification result.

Figure 6 depicts the accuracy values for KNN and SVM using barchart. According to the results of the tests, the KNN algorithm is more accurate than the SVM algorithm. In this system, the KNN algorithm can classify disease types with 96.7% accuracy and the SVM algorithm with 86.7%.

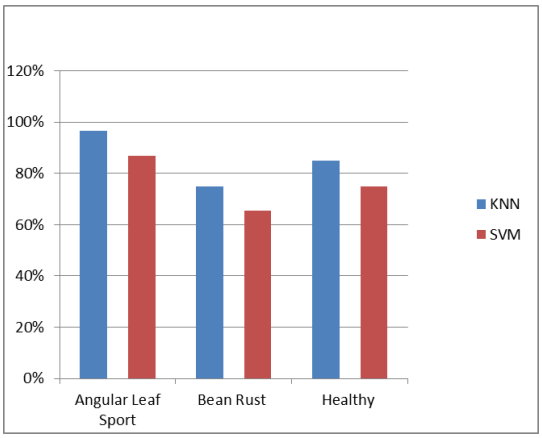



**Figure 6. Comparison Result KNN and SVM**

## 6. Result and Discussion

In this proposed system, the KNN classifier will classify diseases such as angular leaf spot, bean rust, and healthy mungbean leaf diseases. The proposed method detects and recognizes the selected diseases with 96.76% accuracy. However, there are numerous classification methods that can be used to detect leaf disease. As a result, any extended version of the algorithm can be used for classification. Other improved methods should be used to improve accuracy performance.

## 7. Conclusions

Many more modern technologies are used than traditional technologies, and agriculture is vital to the economy. Disease detection in this system is accomplished through the use of image processing and machine learning algorithms. The k-nearest neighbor algorithm (k-NN) can provide better classification results for mungbean while being easier to implement. The image classification of mungbean leaf disease using KNN and GLMC for feature extraction is extremely accurate. To achieve high accuracy, this method can be improved using the attribute selection techniques for bagging or optimization methods, and it can be applied to applications that assist farmers in preventing crop failure and providing automatic detection of mungbean leaf disease, making it efficient for the agricultural sector. As a result, this system can detect and recognize the selected diseases with 96.76% accuracy. Other plant species can also be classified using this classifier.

## References

[1] Dhanachandra, Nameirakpam, Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm /India/December 2015.

[2] Dheeb Albashish, Detection and Classification of Leaf Disease using K-means-based Segmentation and Neural-networks-based Classification /India/February 2011.

[3] D. Al Bashish, M. Braik, and S. Bani-Ahmad, "Detection and classification of leaf diseases using K-means-based segmentation and", Information Technology Journal, vol. 10, no. 2, pp.267275,2011.http://dx.doi.org/10.3923/itj.201.

[4] G.Geetha ,S.Samundeswari , G.Saranya ,K.Meenakshi and M. Nithya,Plant Leaf Disease Classification and Detection System Using Machine Learning, ICCPET 2020.

[5] H. Al-Hiary, "Fast and accurate detection and classification of plant diseases", Int. J.Comput. Appl., vol. 17, no. 1, pp. 31-38, 2011.

[6] Malti K. Singh, Detection and Classification of Plant Leaf Diseases in Image Processing using MATLAB /India/ December 2017.

[7] P.M. Mainkar, S. Ghorpade, and M. Adawadkar, "Plant leaf disease detection and classification using image processing techniques", International Journal of Innovative and Emerging Research in Engineering, vol. 2, no. 4, pp. 139-144,2015.

[8] Vagisha Sharma, Amandeep Verma, Neelam Goel Classification Techniques for Plant Disease Detection, IJRTE, 2020.

[9] Weizheng S, Yachun W, Zhanliang C, Hongda3 W (2008) Grading Method of Leaf Spot Disease Based on Image Processing. Int Conf on Comp Sc and Soft Eng, IEEE 491-494.

[10] XE.Pantazi , D.Moshou and AA.Tamouridou "Automated leaf disease detection in different crop species through image features analysis and One Class Classifiers." Comput Electron Agric v.156, pp.96–104, 2019.

# Natural Language and Speech Processing

# Sentiment Analysis of Product Reviews Using Hybrid Approach

Khin Kyawt Kyawt, Dr. Zar Chi Su Su Hlaing
*University of Computer Studies, Yangon, Myanmar,*
*Myanmar Institute of Information Technology Mandalay*
*khinkyawtkyawtcumagway@gmail.com, Zarchissh@gmail.com*

## Abstract

*During the past decades, online market places have been popular. Most the sellers request the customers to express the reviews of the products. Nowadays individual and grouping depend heavily on website for consumers' reviews in their agreement on buying the product. Product manufacturer also need to take time for analyzing the huge amount of opinions. This paper expresses about the sentiment analysis, the process of mining the texts, in order to distinguish the extract written by the user. So, the thesis proposes the classification of product reviews whether they are positive, negative or neutral by the use of hybrid approach. This paper proposes a framework for reviews data using hybrid approach. hybrid approach used in lexicon and machine learning approach. Machine learning approach is Naïve Bayes classifier. Positive reviews are found the more and negative reviews are found the least in this paper. This paper describes a guideline for training data using Vader lexicon and testing data using machine learning algorithm. Positive reviews are found the more and negative reviews are found the least in this paper.*

**Keywords**: *sentiment analysis, Vader lexicon, text classification, opining mining, supervised learning*

## 1. Introduction

Nowadays, Twitter, Facebook, Blog and Amazon have become a vital role of human living in recent day. People can share their opinion (sentiment analysis) on each and every happening of the life. Thus, sentiment analysis is useful to extract and understand their attitudes on certain topics. The combination of lexicon and machine learning approach is hybrid approach. In this paper hybrid-based approach is used to analyze the sentiment analysis. The objectives are to classify the polarity of reviews dataset, to distinguish the attitude of customers they commented, to analyze all the reviews of product buyers' opinions within a short time, to support customers and save their time and long effort to reviews that may lead to erroneous checking for the human.

Lexicon-based approach is a straightforward and factual approach to Sentiment Analysis. Lexicon-based signify the numbers of positive and negative words in the text and the most frequently used will be counted. Lexicon give off recommendations for eliminating the negative aspects of individuals. Sentiment analysis is opinion mining. It is also a natural language processing problem and used to extract subjective information from texts and to analyze people's reputations, sentiments, evaluation towards entities such as materials, servicing, organization.

The aims of sentiment classification to classify positive, negative and neutral words in the reviews. It applies a machine learning approach, a lexicon-based approach in the current paper. Sentiment analysis contains three levels: Document level, Sentence level and Aspect level. A classification model means machine learning approach, which is trained using the pre-labeled dataset (positive, negative and neutral) before it can be applied to actual classification task. Machine learning uses classifiers such as Naïve Bayes, SVM, etc.

This study used both lexicon and machine learning approach. This approach may provide compliments in order to bind the tough and dispose of the flaws of the individual to techniques. In this paper, we collected product review data from Amazon product review dataset.

The more the size of training data, the higher accuracy. This paper expresses the following facts. In section 2, we express background theory and in section 3, we express related research. In section 4, we explain the proposed technique. In section 5, steps of proposed approach are discussed.

## 2. Background Theory

Lexicon based methods and machine learning methods are two main approaches that are usually

used for sentiment classification. Sentiment analysis is opinion mining. It is also a natural language processing problem and used to extract subjective information from texts and to analyze people's opinion, sentiments, evaluation towards entities such as materials, servicing, organization. The aims of sentiment classification to classify positive, negative and neutral words in the reviews. It applies a machine learning approach, a lexicon-based approach and hybrid approach in this paper.

**VADER** (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon. It uses a combination of a sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled according to their semantic orientation as Positive, Negative or Neutral. The approach maps words to *sentiment* by using a lexicon or a 'dictionary of sentiment. It is a rule-based sentiment analysis used for text sentiment analysis. The system use this dictionary to assess the sentiment of product review text. Naïve Bayes is a classification technique based on Bayes' Theorem with as assumption of independence among predictors. Naïve Bayes classifier assumes that the presence of a particular feature in class is unrelated to the presence of any other feature. Ever if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naïve'.

Naïve Bayes model is easy to build and particularly useful for very large data sets. Along with has been simplicity, Naïve Bayes is known to outperform ever highly sophisticated classification methods. The simplest solutions are usually the most powerful one and Naïve Bayes is a good example. It has been successfully used for many purposes but it works particularly well with natural language processing problem. Sentiment analysis contains three levels: Document level, Sentence level and Aspect level.

Document level is used to classify a whole page which expresses a positive or negative sentiment. Sentence Level is used to classify each sentence that expresses a positive, negative or neutral opinion. Aspect Level is used to classify based on the facts. The aim of sentiment classification is to analyze the feedbacks of user and distinguish them positive, negative and neutral opinion.

Classification is the process of assigning labels to the reviews. In proposed work, data collection, pre-processing, feature extraction using multinomial Naïve Bayes and sentiment classifier using Naïve Bayes. Naive Bayes is independent of the number of features in the feature space. It is a statistical classification technique based on Bayes Theorem. Naïve Bayes classifier currently experiencing a renaissance in machine learning, has long been a core technique in information retrieval. Naïve Bayes is one of the simplest supervised learning algorithms and its classifier is the fast, high accuracy, speed on large datasets, thus Naïve Bayes algorithm is reliable algorithm. Further, it assigns the label of nearest neighbor to the unlabeled reviews.

## 3. Related Work

The purpose of this paper is to analyze amazon reviews as positive, negative and neutral. It uses two approaches such as lexicon-based approach and machine learning approach.

Lexicon based approach perform sentiment analysis at document level by calculation the sentiment score of the product reviews that are already defined in lexicon to obtain training set. SEPIDE PAKNEJAD describe Sentiment classification on Amazon reviews using machine learning approaches. Author explains classifying reviews positive or negative from beauty products from Amazon. Finally, two different supervised machine learning techniques such as SVM and Naiıve Bayes. In this paper, accuracy the SVM approach achieves better results than the Naive Bayes approach [1].

Thakare Ketan Lalji and Sachi N. Deshmukh (2016) discuss the literature survey related to this paper and the methodologies for twitter sentiment detection in the work "Twitter Sentiment Analysis using Hybrid Approach". They show the results of experiment they done on different twitter datasets [2].

Amlan Chakrabarti and Paramita Ray (2017), propose a framework for sentiment analysis using R software which can analyze sentiment of users on Twitter data using Twitter API in the research "Twitter Sentiment Analysis for Product Review Using Lexicon Method". They have done both document level and aspect level analysis based on the proposed methodology, which helped in decision making [3].

Bo Yan (2017) discusses about the specialized lexicon for the petroleum domain in the study

"Sentiment Analytics: Lexicons Construction and Analysis". It is hypothesized that coupling a specialized lexicon to a general lexicon, such as Senti WordNet, will produce better results. The results suggest that this hypothesis is supported [4].

Sepideh Paknejad (2018) showed that in terms of accuracy the SVM approach achieves better results than Naïve Bayes approach when the whole data set was used as training and testing data set in the study "Sentiment Classification on Amazon reviews using machine learning approaches". However, both algorithms reached promising accuracies of at least 80% [5].

H.M. Keerthi Kumar, B.S. harish, H.K.Darshan (2018) clearly differentiate between a positive review and negative review in the work "Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method". Thus, results obtained are highly promising both in terms of space complexity and classification accuracy [6].

Vipul Kumar Chauhsssar, Ashish Bansal, Dr. Amita Goel (2018) describes sentiment analysis along with the new evaluating tool VADER in the paper, "Twitter Sentiment Analysis Using Vader". VADER distinguishes itself from others in terms that it is more sensitive to sentiment expressions in social media contexts while also generalizing more favorably to other domain [7].

## 4. The Proposed System

The proposed system approach to sentiment analysis as follow:

First, it is labeling processing by calculating the sentiment polarity scores of the raw Amazon product review data using Vader to obtain training set. It supports in calculating as it contains word and sentiment score built-in together inside. Important to calculate the normalization score of each sentence.

Naïve Bayes algorithm includes supervised learning algorithm which requires training data that has been labeled or known in advance.

Then, Naïve Bayes performs review data with unknown labels in accordance with the training data labels for classification process. The performance of sentiment analysis to improve we use the following figure 1.

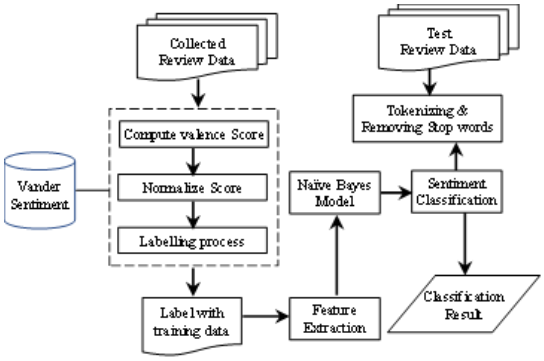



**Figure 1. System Flow Diagram of Sentiment Analysis of Product Review**

## 5. Step of Proposed Approach

The Sentiment Analysis of reviews includes following steps.

Step One - Reviews Data Collection
Step Two - Data Cleaning
Step Three - Labeling process **using Vader**
Step Four - Classification of Navies Bayes Machine Learning Algorithm
Step Five - Sentiment result output.

### 5.1. Data Collection

The data is obtained by scraping with google chrome web scraper and manually to store required data into review data collection. The following table is sample reviews for training.

**Table 1. Product Review Content**

| Reviews ID | Content |
|---|---|
| Review 1 | The shoe is pretty good. |
| Review 2 | It wastes your money buy this faulty shoe. |
| Review 3 | I received faulty piece. |
| Review 4 | This is a really smart shoe; the quality of the leather is amazing. |
| Review 5 | Expensive, but is a comfortable shoe with advance method. |
| Review 6 | It is a product. |
| Review 7 | It is simple one. |

### 5.2. Data Cleaning

This stage comprises of removing the punctuations, emoji and transformation of upper

case to lower case. Unnecessary data will need to be cleaned from the reviews data. Then, we use the method word_tokenize () to split a sentence into words. A stop word is a commonly used word (such as "the", "a", "an", "in"). After noise removal, we obtain the training set by using Vader sentiment analysis tool. The process of stemming reduces the inflection words in the language.

### 5.3. Labelling process using Vader

Positive values are positive valence and negative values are negative valence.

**Table 2. Sentiment Score**

| Sentiment Word | Valence Score |
|---|---|
| good | 1.9 |
| smart | 1.7 |
| pretty | 2.2 |
| bad | -2.5 |
| terrible | -2.1 |
| faulty | -1.3 |
| comfortable | 2.3 |
| amazing | 2.8 |
| poor | -2.1 |

**Example: Input Review Text: ("The shoe is pretty good.")**
Two emotional words**:** *super and cool* Lexicon ratings (sentiment score) for *pretty* and *good* are: 2.2 and 1.9.

**Table 3. sentiment score value**

| **Review** | **Sentiment Metric** | **Score** | **label** |
|---|---|---|---|
| **The shoe is pretty good.** | Positive | 0.670 | pos |
| | Neutral | 0.329 | |
| | Negative | 0.0 | |
| | Compound_ score | 0.726 (0.726>0.05) | |

**Explanation of Double Negatives Sentences**

A **double negative** is using two negative words or phrases in a sentence. A double negative is produced when we use two negative words in a clause so that it would create a positive effect. This is not a simple sentiment problem, it's total natural language understanding (negation, context, etc.). Some examples of double negatives:

She is not incorrect.
Time is not unlimited.
The phone is not poor.

Negation Handling is important part while sentiment analysis. Many sentences include the negation word that shifts the polarity of the sentence. This approach solves the problem when the system find any negation term in sentence. And the system multiplies the negation scalar value to valence scores of sentiment word. (Neg-scalar= -0.74)
Example 1: Input Review Text: “I do not mean that this product is not good." Lexicon ratings (sentiment score) for good is 1.9.
Example 2: Input Review Text**:** "I never recommend to buy this bad quality product. "Lexicon ratings (sentiment score) for recommend and bad are 1.5 and -2.1.

**Table 4. Sentiment Score values for Double Negative**

| No | Review | Sentiment Metric | Score | label |
|---|---|---|---|---|
| Example 1 | I don’t mean that this product is not good. | Positive | 0.203 | pos |
| | | Neutral | 0.796 | |
| | | Negative | 0 | |
| | | Compound __score | 0.259 (0.259> 0.05) | |
| Example 2 | I never recommend to buy this bad quality product. | Positive | 0.419 | pos |
| | | Neutral | 0.580 | |
| | | Negative | 0.580 | |
| | | Compound _ score | 0.619 (0.5>0. 05) | |

**Table 5. Labeling process using Vader**

| Review Text | Pos-Score | Neg-Score | Neu-Score | Norm-Score | Compound-Score | Label |
|---|---|---|---|---|---|---|
| Review-1 | 0.67 | 0 | 0.33 | 0.769 | 0.7269 | pos |
| Review-2 | 0 | 0.325 | 0.675 | -0.5228 | -0.5228 | neg |

| | | | | | | |
|---|---|---|---|---|---|---|
| Review-3 | 0 | 0.535 | 0.465 | -0.3182 | -0.3182 | neg |
| Review-4 | 0.43 | 0 | 0.57 | 0.7778 | 0.7778 | pos |
| Review-5 | 0.389 | 0 | 0.611 | 0.6652 | 0.6652 | pos |
| Review-6 | 0 | 0 | 1 | 0 | 0 | neu |
| Review-7 | 0 | 0 | 1 | 0 | 0 | neu |

Feature extraction have two stages namely – lexicon-based feature extraction and Machine learning based features extraction.

Rule-based sentiment analysis tool used for text sentiment analysis **is VADER** (Valence Aware Dictionary and sEntiment Reasoner) and is also a lexicon. The approach maps words to sentiment by building not only a lexicon but also a dictionary of sentiment. The system uses this dictionary to analyze the sentiment of product review text.

**Algorithm for Positive, Negative, Neural Scores**

```
for sentiment_score in sentiments:
If sentiment_score > 0 then
 pos_sum += sentiment_score + 1
Else if sentiment_score < 0 then
        neg_sum += sentiment_score -1
Else if sentiment_score ==0 then
        neu_count += 1
End if;
End for
total= pos_sum+neg_sum+neu_count
pos_score= pos_sum/total
neg_score= neg_sum/total
neu_score= neu_count/total
```

The total score (sentiment score) is calculated by summing the sentiment(valence)scores of each word in the lexicon,

Calculate total score of all positive valence scores and all negative valence scores: total score, x

$$x = \sum_{i=1}^{n} pos\_valence\_score + neg\_valence\_score$$

The total score is modified according to the rules, and then normalized to be between -1 and +1. Normalize the score to get **compound score**:

$$norm\_score = \frac{x}{\sqrt{x^2+\alpha}}$$

where x = sum of valence scores and α = Normalization constant value is 15. Typical threshold values are:
(compound score >=0.05) **positive sentiment**:
(compound score>-0.05) and (compound score <0.05) **neutral sentiment**:
(compound score<=-0.05) **negative sentiment**:

The pos, neu, and neg sentiment scores are texts that fall in each category.

Machine learning based feature extraction is used in computing term weight using Multinomial Naïve Bayes algorithm. Finally, Naïve Bayes classifier classifies the data based on the training data into predicted sentiment (positive, negative or neutral).

## 6. Training Data

The data is obtained by scraping with google chrome web scraper and manually to store required data into review data collection. To get raw data, you may enter at Chrome, firstly and then https://www.scrapehero.com/amazon-review-scraper.
Training data is data which can be labeled as positive, negative and neutral by using rule-based sentiment analysis tool.

## 7. Accuracy Result

### 7.1. Classification with Naïve Bayes Machine Learning Algorithm

We calculate term weight values using Multinomial Naïve Bayes Classification process use Naïve Bayes. It sets standardized thresholds for classifying text as either positive, neutral or negative. Typically, threshold values are:

compound score values greater than 0.05 is positive, compound score values between 0.05 and 0.05 is neutral compound score values less than 0.05 is negative.

### 7.2. Input Review Text:

The following sentences are used for testing.

(a) It is simple one.

The input review text results are 0.00026875 for positive, 0.000261 for negative and 0.0006 for neutral. Neutral result value is maximum thus the input reviews text result is neutral.

(b) faulty product

The input review text results are 0.000268 for positive, 0.000783 for negative and 0.000609 for

neutral. Negative result value is maximum thus the input reviews text result is negative.

(c) Smart shoe perfect one.

The input review text results are 0.00000100781 for positive, 0.0000004698 for negative and 0.0000002349 for neutral. Positive result value is maximum thus the input reviews text result is positive.

## 8. Performance Evaluation

There are so many comments reviews in Amazon. Among them, the comments reviews data for shoes 1000 obtained. We collected reviews from 1000, 700 reviews use for training data and 300 reviews use for testing data. We used F-measure, Precision, Recall, and Accuracy to analyze the performance of reviews. F-measure is a value between the balances of Precision and Recall. Accuracy focuses on the aggregates of exact prophecies. Precision and Recall are effective ways to evaluate the correction of classes.

PR represents precision,
RE is Recall,
FM is F-measure,
True Positive is TP,
False Positive is FP,
True Negative is TN,
False Negative is FN,

Accuracy is ACC Respectively,

$$PR=\frac{TP}{TP+FP}$$

$$RE=\frac{TP}{FN+FP}$$

$$FM=2\frac{PR.RE}{TP+FP}$$

$$ACC=\frac{TP+TN}{TP+FP+TN+FN}$$

**Table 6. Four Matrix Calculation Results**

| | Positive | Negative | Neutral |
|---|---|---|---|
| Accuracy | 86% | 84% | 84% |
| Precision | 82% | 61% | 100% |
| Recall | 83% | 97% | 56% |
| F-Measure | 83% | 75% | 72% |

These evaluation results for accuracy are positive 86 %, negative 84% and neutral 84% in sentiment calculation.

## 9. Conclusion

This paper has explored the sentiment analysis of the product reviews from Amazon use Hybrid Based Approach. We can realize the importance of the Sentiment Analysis. Thus, this paper is aiming to conduct Sentiment Analysis of product reviews by classifying reviews into positive, negative and neutral sentiment. Based on results of this paper, positive sentiment is mostly found and the least one is negative sentiment. Moreover, Sentiment Analysis enables the company to find out the customer's opinion about its product. Sentiment Analysis helps to obtain the solutions and satisfactions from customer.

## 10. Acknowledgement

Firstly, I would like to describe my gratitude for the permission to submit this paper. I would like to acknowledge my sincere gratitude and appreciation to my supervisor her effort time in reading and patience to help me in accomplishing this paper. It was a great privilege and honor to work and study under her guidance. Furthermore, I am extremely grateful to my companions who have given me their precious ideas and invaluable knowledge throughout the paper. Finally, I am extending my heartfelt thanks to my family for their encouragements and support to accomplish this paper.

## References

[1] SEPIDE PAKNEJAD, Sentiment classification on Amazon reviews using machine learning approaches. DEGREE PROJECT IN TECHNOLOGY, FIRST CYCLE, 15 CREDITS STOKCKHOLM, SWEDEN 2018.

[2] Kalpana Algotar and Ajay Bansal, Detecting Truthful and Useful consumer Reviews for Products using Opinion Mining. Arizona State University, Mesa AZ 85212 USA [ Kalgotar, ajay.banssal] @ asu.edu

[3] H. M. Keerthi Kumar1, B. S. Harish2, H. K. Darshan3, Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method.JSSRF, JSS TI Campus, Mysuru, Karnataka (India)2 Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, Karnataka (India)3 Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, Karnataka (India)

[4] Thakare Ketan Lalji1, Sachin N. Deshmukh, Twitter Sentiment Analysis Using Hybrid Approach.1, 2Department of Computer Science and IT,Dr. Babasaheb Ambedkar MarathwadaUniversity, Aurangabad

[5] Bo Yuan, Sentiment analytics: Lexicons construction and analysis. Yuan, Bo, "Sentiment analytics: Lexicons construction and analysis" (2017). Masters Theses. 7668. https://scholarsmine.mst.edu/ masters_theses/7668.

[6] Paramita Ray1and Amlan Chakrabarti2,Twitter Se ntiment Analysis for Product Review Using Lexicon Method.2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Pune, India, Feb 24-26, 2017.

[7] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. 2011. Lexicon- based methods for sentiment analysis.Comput Linguist. 37, (2): 267--307.

[8] A.Ortigosa, J. M. Martín, and R. M. Carro.“Sentiment analysis in Facebook and its application to e- learning.” Computers in Human Behavior Vol. 31, pp.527-541. 2014.

[9] Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram. (2004). Design

[10] Science in Information Systems Research.DIS Quarterly, 75-105.

[11] S, ChandraKala1 and C. Sindhu2, “OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY,”.Vol .3(1),Oct 2012,420-427

[12] Richard A Berk.Statistical learning from a regression perspective. Springer, 2016.

[13] Multi Perspective Question Answering (MPQA). OnlineLexicon“http://www.cs.pitt.edu/mpqa/subj_lex icon.html”.

[14] Ding, X., Liu, B., & Yu, P.S. 2008. A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08). ACM, New York, NY, USA. pp. 231-240

[15] Maharani, W. (2013). Microblogging sentiment analysis with lexical based and machine learning approaches. Information and Communication Technology (ICoICT), 2013 International Conference (pp. 439-443). Bandung: IEEE.

[16] Kyomin Jung, Byoung-Tak Zhang, Prasenjit Mitra. (2015). Deep Learning for the Web. the 24th International Conference on World Wide Web (pp. 1525-1526), International World Wide Web Conferences Steering Committee.

# Automatic Speech Recognition for Rakhine Language

Hnin Thi Dar Kyaw, Aye Nyein Mon
*University of Computer Studies, Yangon, Myanmar*
*hninthidar.kyaw@ucsy.edu.mm, ayenyeinmon.ucsy.edu.mm*

## Abstract

*Speech Recognition is the process of converting speech signal into text. Automatic Speech Recognition (ASR) has been carried out by many researchers for their languages to improve in language technologies. This paper developed automatic speech recognition for Rakhine language by utilizing the Gaussian Mixture Model based Hidden Markov Model (HMM-GMM). Rakhine Language is one of the low-resourced languages and speech corpora are no freely available. Therefore, in this work, speech corpus is created on two domains: broadcasts news and daily conversations. The speech corpus size is 6 hours 20 mins 3 sec and includes 10 female speakers and 4 male speakers. The experimental results achieved the lowest word error rate (WER) of 20.50% on Testset1 (recorded conversational data) and 17.77% on Testset2 (broadcast news) using HMM-GMM.*

**Keywords**: Automatic Speech Recognition (ASR), Hidden Markov Model and Gaussian Mixture Model (HMM-GMM), Rakhine language

## 1. Introduction

Nowadays, speech recognition is one of the modern technologies for human computer interaction and speech signal is prominent feature to interact with natural language. Speech Recognition is the process of converting speech signal into a sequence of words without typing by hand. Automatic Speech Recognition (ASR) is a challenging task because of human speech signals variability. In recent years, every nation tries for developing ASR with own language to show their nations language technology. Many researchers have been done for ASR both of well-resourced and low-resourced. ASR research has been conducted on the Myanmar Language by using different techniques to improve Myanmar ASR. There are some recent works ASR for Myanmar language using different techniques [1] [2] [3] [4].

Rakhine (Arakanese) Language is the native language in Rakhine spoken by Arakanese people. ASR related to Rakhine Language has not been done yet. In this work, Rakhine ASR is implemented for the first time. A speaker independent and continuous Rakhine Language speech recognition is built by using HMM-GMM.

This paper is structured as follows. Section 2 is about the nature of Rakhine Language. Section 3 presents building a speech corpus and pronunciation lexicon for Rakhine ASR. The overview of ASR architecture and about its components are described in Section 4. The experimental setup and results are shown in Section 5. Section 6 describes the summarization of the research work and future research on Rakhine ASR.

## 2. The nature of Rakhine Language

Rakhine Language is a tonal language which is closely similar to Myanmar Language. It is mainly spoken by Arakanese people in Rakhine state. Arakanese speak in the tongue and they also have many vocal changes. Rakhine Language can be divided into two dialects Sittwe (North) and Thandwe (South). Although Rakhine (Arakanese) is very similar to Myanmar (Burmese), there are many distinct differences between the two languages. There are significant vocabulary differences and also differs pronunciation with Myanmar language Example, အာကာ (a g a) in Rakhine, ကောင်းကင် (g aun g in) in Myanmar). Some words are same words, same pronunciations but difference meanings eg. ပုဆိုး (p a- hs ou:) means စောင် in Rakhine. လုံချည် (l oun gy i) is called ဒယော (d a- j o:) in Rakhine. Some are same words but difference pronunciation eg, ကြိုတင် pronounce (k r ou t en) in Rakhine but ကြိုတင် pronounce (k j ou t in) in

Myanmar. Others are all the same both meaning and pronunciation eg, ပန်းသီး (pan: thi:). There are many native words such as အဘုသွေ့ (a- b u. ch ei), ပဒါကာသီး (b a- d a g a th i:), ချောဒေါင့်သီး (ch o: d aun. th i:), ငသတိုက် (ng a- dh a- d ai'), အမင် (a- m en), ဘားဘာ (b a: b a) which are not found in Myanmar. Others are foreign words such as ကိုဗိုက် (k ou b ai'), စမ်တမ်ဘာ (s a- t an b a). Most of the foreign words are pronounced one or more sounds in Rakhine. Example, ကိုဗိုက် (k ou b ai'), ကိုဗိုက် (k ou b ain), ကိုဗိုက် (kh ou b i.), စမ်တမ်ဘာ (s a- t en b a). စမ်တမ်ဘာ (s an t en b a), စမ်တမ်ဘာ (s en t en b a).

## 3. Building a Speech Corpus and Pronunciation Lexicon for Rakhine Language

Rakhine language is one of the low-resourced languages and there is no pre-defined data to build speech recognition system. For low-resourced languages, building a speech corpus is crucial steps for training any speech recognition system. Speech corpus building is also the first step in developing Rakhine ASR system.

Generally, speech data can be built in two methods. One method is designing the text corpus first. And, recording the speech by uttering the collected texts. The second method is collecting the speech data which is already been recorded and they are manually transcribed into texts. In this task, a Rakhine speech corpus is constructed using the two methods and it is built on two types of domains: daily conversations and broadcasts news [5].

### 3.1. Recording Daily Conversations

The first approach (designing the text corpus first and then recording it) is used for daily conversations data. These sentences are collected from Rakhine guidance book [8] and daily conversations. After that these collected Rakhine sentences are recorded with the help of Tascam recorder at UCSY NLP lab, lab1 and digital library. It involves Rakhine digits and daily conversations (telephone, hotel, market, street) etc. The range of the segmented files is between 1 sec to 3 sec.

### 3.2. Collecting data from APM Broadcast news

Nowadays, a lot of speech data are available on the internet and they can be collected from the web. Although Rakhine news is found on websites, news in Rakhine Language is very rare. Rakhine language has lack of online resources to construct Rakhine speech corpus. Arakan Princess Media (APM) is Rakhine Broadcast news, it was established on Facebook page, website and telegram. This channel broadcast international news and Myanmar news and Arakan News. It is spoken in Rakhine language. Therefore, the speech data is collected from the site of APM Broadcast news [6]. It includes both local news (political, health, social) and foreign news (political, health, sport, weather, social). The speech file segmentation is made with the help of audacity1. And, the speech file format is converted WAV file format with single channel mono type and sampling rate is 16 KHz. Moreover, silence and background noise portions are removed in segmenting the speech files. The range of the speech file is between 2 sec to 18 sec. The detailed statistics of the speech corpus is described in Table 1, (N means north and S means south).

**Table 1: Statistics of Rakhine Speech Corpus**

| Data | Size | Speakers | | | | | Utteraces |
|---|---|---|---|---|---|---|---|
| | | Female | | Male | | Total | |
| | | N | S | N | S | | |
| APM Broadcast | 3 hrs 4 mins 3 sec | 3 | 4 | 2 | 1 | 10 | 1439 |
| Daily Conversations | 3 hrs 16 mins | 3 | - | - | 1 | 4 | 6848 |
| Total | 6 hrs 20 mins 3 sec | 7 | 4 | 2 | 1 | 14 | 8287 |

### 3.3. Building a Rakhine Text Corpus

Rakhine text corpus are manually constructed which contains 8287 sentences from APM

[1]http://audacity.sourceforge.net/help/documentation

broadcasts news and daily conversations. These sentences are manually segmented into word level and manually check the spelling of the words. It involves 1 word to 55 words on average in one utterance. The example sentences of daily conversational data and Broadcast news are as shown in Figure 1 and Figure 2.

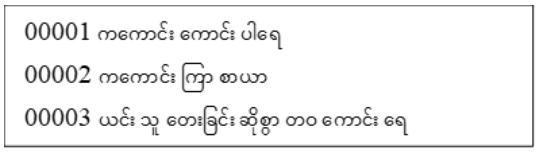



**Figure 1: Sample texts of daily conversations**

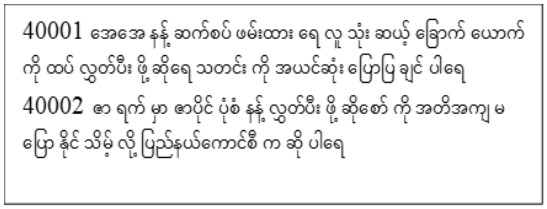



**Figure 2: Sample texts of APM broadcast news**

## 3.4. Rakhine Pronunciation Lexicon

Pronunciation lexicon is one of the components of ASR system. It is a list of words which is expressed phoneme for each word.

### 3.4.1. Rakhine Consonants Phonemes

Rakhine has 33 consonants (က to အ) as standard Myanmar language and 23 phonemes for 33 consonants scripts.

### 3.4.2. Rakhine vowel phonemes

Rakhine language uses 44 vowels. Most of the vowel phonemes are similar with Myanmar language. Example, အိ (i.), အီ (i), အီး (i:), အေ (ei), အေ့ (ei.), အေး (ei:), အယ် (e), အယ့် (e.), အဲ (e:), အာ. (a), အာ (a), အား (a:), အော် (o), အော့် (o.), အော (o:), အူ (u), အု (u.), အူး (u:). However, some vowels are not used in Rakhine. Some phonemes are the same in writing script but they have different pronunciations. For instance, အစ် (i') and အင် (in) pronunciation are not used in Rakhine. For example, 'စစ်တွေ' is as pronounced (s i' t we) in Myanmar, however စစ်တွေ is pronounced as (s ai' t we) in Rakhine. In this example, writing script 'အစ်' is used however not pronounced as အစ် (i'). Instead of အစ် (i'), အိုက် (ai') is pronounced in Rakhine. The next one is that ခင် is not pronounced as (k in) in Rakhine instead of ခင် (k in), ခင် (k en) is pronounced. In this example, although writing script အင် is used, pronunciation for အင် is (en) in Rakhine. Other examples are that ပွတ် is pronounced as (p u') in Myanmar however (p wa') in Rakhine and မွန် (m un) is pronounced as (m wan).

**Table 2: Example of Phonetic Mapping**

| Rakhine | Proposed mapping |
|---|---|
| ယ | j/r |
| ကျ | ky |
| ဂျ | gy |
| ချ | ch/kh r |
| ပြ | p r a./ p j a. |
| ရှ/သျ | sh/hr |

In this work, Rakhine Pronunciation lexicon is created to develop Rakhine ASR. It collected 2134 words from ရခိုင်ဘာသာစကားလမ်းညွှန် [7], 2100 words from ရခိုင်ဂန္ထဝင်ဝေါဟာရ [8] and 2950 words from Arakan princess media (broadcast news) for Rakhine Pronunciation lexicon. The pronunciation for each word is manually annotated. The total vocabulary of Rakhine lexicon contains 7184 words and 68 phonetic units to represent the pronunciation of words.

**Table 3: Sample of Rakhine Phonetic dictionary**

| Rakhine Words | Phonetic |
|---|---|
| ကြိုတင် | k r ou t en |
| ညမချေ | nj a- m a. ch ei |
| ကိုဗိုက် | k ou b ai' |
| ရှိဆိုင် | hr i. hs a in |
| ရီတံခွန် | r i t a- kh w an |
| ချောင်းဆိုး | kh r a un: hs ou: |
| အီးယောင် | i: j a un |

## 4. Automatic Speech Recognition (ASR)

The Automatic Speech Recognition system mainly involves five stages: feature extraction, acoustic model, language model, pronunciation lexicon and decoding [9].

### 4.1. Feature Extraction

Feature extraction is one of the integral parts of speech recognition process. It transforms speech signal into sequence acoustic feature vectors, each vector representing the information in a small-time window of the signal. In this work, Mel Frequency Cepstral Coefficient (MFCC) feature extraction technique is used.

### 4.2. Acoustic Modeling

Acoustic model is the main component of ASR system. Acoustic models are used to map the observed features of speech signal with the expected phonetic of hypothesis sentence. There are many approaches for acoustic modeling such as Hidden Markov Model and Gaussian Mixture Model (HMM-GMM), Deep Neural Network (DNN), Convolutional Neural Network (CNN), Time Delay Neural Network (TDNN) and so on. In this work HMM-GMM based acoustic model is used. HMM is used to represent the transition probabilities between states. And, the transition between phones and corresponding observable can be modeled with the Hidden Markov Model (HMM) [10]. Hidden Markov Models represent each unit of speech in the acoustic model. Each state of an HMM is represented by a set of Gaussian mixture density functions. GMM is the observed probability distribution of the feature vectors given a phone and provides a principled method to measure distance between a phone and our observed audio frame. A Gaussian distribution is a function parameterized by mean or average value and a variance which characterizes the average spread or dispersal from the mean.

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (1)$$

$$\mu(mean) = {}^{1}/_{N} \sum_{i=1}^{N} x_i \quad (2)$$

$$\sigma^2(\text{variance}) = {}^{1}/_{N} \sum_{i=1}^{N} (x_i - u)^2 \quad (3)$$

A Gaussian mixture model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [11]. M mixtures of Gaussian,

$$f(x|\mu,\Sigma) = \sum_{k=1}^{M} c_k \frac{1}{\sqrt{2\pi|\Sigma_k|}} \exp[(x-u_k)^T \Sigma^{-1}(x-\mu_k)]$$

x = observations O for phone likelihood computation
μ = mean
Σ = covariance matrix

### 4.3. Language Modeling

ASR systems apply n-gram language model to find correct word sequence. In this work, 3-gram language model is created by using SRILM language modeling toolkit [12]. It includes 6848 sentences from daily conversation data, 1439 sentences from Arakan Princess Media broadcast news, 1854 sentences from the corpus of Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese) [14]. Rakhine sentences are manually segmented into words to create language model and manually check the spelling of the words. The following sentence is an example of word segmented Rakhine sentence.

Rakhine : စျိုး က အချာ တစ် ဟပ် ဝယ် ခ ပလင်
Myanmar : ဈေး က ပိတ်စ တစ် ထည် ဝယ် ခဲ့ ပါလား

## 5. Experimental Setup and Evaluation Result

The experiments are done using Kaldi [13] toolkit. The detailed training data and test data used in the experiment is shown in Table 4.

### 5.1. Experiment Setup

The speech corpus has two types of domains: web news (APM Rakhine broadcasts news and daily conversations). The type of speech data is both read speech. Rakhine Language can be divided into two dialects: Sittwe (North) and Thandwe (South). Therefore, this corpus consists

of speakers from North (N) and South (S) regions. For the baseline GMM-based acoustic model training, the standard Mel-Frequency Cepstral Coefficients (MFCC) with its first and second derivates without energy features are used. Then, cepstral mean and variance normalization (CMVN) is applied on MFCC features. After that, splicing 9 frames of MFCCs together and linear discriminant analysis (LDA) is used to project down to 40 dimensions. A maximum likelihood linear transform (MLLT) is applied for estimation on the LDA features. After that, speaker adaptive training is conducted with feature-space Maximum Likelihood Linear Regression (fMLLR) on the top of LDA and MLLT model. There are an average of 44 Gaussian components per state with 2050 context dependent (CD) triphones in GMM-HMM model.

**Table 4: Training data and test data used in the experiments**

| Data | Size | Speakers | | | | | Utterances |
|---|---|---|---|---|---|---|---|
| | | Female | | Male | | Total | |
| | | N | S | N | S | | |
| **Train Set** | 5 hrs 46 mins | 5 | 4 | 3 | 2 | 14 | 7723 |
| **Test Set 1** | 15 mins | 1 | - | - | - | - | 418 |
| **Test Set 2** | 15 mins | 2 | 1 | 1 | 1 | 5 | 146 |

In Broadcasts news, the speech corpus size is 3 hours 4 mins 3 secs spoken by 10 speakers (3 females and 2 males in North and 4 females and 1 male in South) with 1439 utterances. For daily conversational data, the duration of the recorded speech is 2 hours 16 mins. Hence, the total size of training size is about 6 hours. TestSet1 includes 1 speaker with recorded conversational data. TestSet2 involves with native 5 speakers. Both TestSet1 and TestSet2 are open Testset.

## 5.2. Evaluation Metrics

The evaluation of speech recognition system is measured by word error rate (WER). The following formula is used to compute (WER),

$$\text{WER} = \frac{Insertions(I)+Substitutions(S)+Deletions(D)}{Total\ Words} \quad (5)$$

**Reference Text**

*တစ်ဆက်တည်း မှာ ပင် အာဖကန် အရှိုး ကူညီ ဆောင်ရွက် ဖို့ မူဆလင် နိုင်ငံ တိ တွိဆုံ ဆွေးနွေး ဖို့ ဆိုရေ သတင်း ကို လေ့ ပြောပြ ချင် ပါရေ*

**Hypothesis Text**

S S I

*တစ်ဆက်တည်း မှာ ပင်း အာဖကန် အရှိုး ကူညီ ဆောင်ရွက် ဖို့ မှာ မူဆလင် နိုင်ငံ တိ တွိဆုံ ဆွေးနွေး ဖို့ ဆိုရေ သတင်း ကို လေ့ ပြောပြ ချင် ပါရေ*

Insertions (I) = 1, Substitutions (S) = 2, Deletions (D) = 0

WER = (1+2+0)/21*100 = 14.29%

Rakhine ASR performance is evaluated using HMM-GMM acoustic modeling technique. The result of different experiment as shown in Table 5.

**Table 5: Evaluation of Rakhine ASR performance in terms of WER**

| HMM-GMM Model | WER% | |
|---|---|---|
| | TestSet1 | TestSet2 |
| **Mono (Δ+ΔΔ)** | 27.35 | 28.81 |
| **Tri (Δ+ΔΔ)** | 24.11 | 19.89 |
| **Tri (LDA+MLLT)** | 22.24 | 18.68 |
| **Tri (LDA+MLLT+SAT)** | 20.56 | 17.77 |

For context independent (CI) monophone training, word error rates (WERs) of 27.35 % on TestSet1 and 28.81% on TestSet2 are achieved. When context dependent triphone model with MFCC+Δ+ΔΔ features are applied, it can be reduced 3.24% on TestSet1 and 8.92 % on TestSet2 than the baseline monophone model. When the triphone model with speaker independent transformation (MFCC+ Linear Discriminant Analysis (LDA) + Maximum Linear Likelihood Transform (MLLT)) are used, WERs of 22.24% on TestSet1 and 18.68% on TestSet2 are obtained. With speaker adaptive training (MFCC + LDA + MLLT + SAT), it can be decreased 6.79% WER on TestSet1 and 11.04% on TestSet2 in comparison with the baseline monophone model. Therefore, the lowest WERs 20.56% on Testset1 (Conversational data) and 17.77% on Testset2 (APM Broadcast news) are attained with the speaker adaptive training.

Although the training data set is a small data set, the ASR performance for Rakhine language

gets promising result because of the Rakhine lexicon containing words which covers the most frequent words of the web text and daily conversations data. When comparing the evaluation result of Testset1 (Conversational data) and Testset2 (Broadcasts news data), Testset2 has lower error rate than TestSet1 because broadcast news has clear voice and less noisy than the recording data.

## 6. Conclusion

In this paper, Rakhine continuous speech recognition is developed by using the classical HMM-GMM approach. It is the first time of ASR development for Rakhine language. Since Rakhine language is a low-resourced language, a speech corpus for Rakhine language is built on two types of domains: daily conversations and APM broadcast news data. In this work, the ASR performance is evaluated 6 hours of Rakhine speech corpus and it obtained 20.56% of WER on Test set1 (Conversational data) and 17.77% of WER on Test set2 (APM broadcasts news).

This study focuses on the classical HMM-GMM approaches. There are still many interesting avenues of discriminative acoustic modeling approach such as Maximum Mutual Information (MMI), boosted Maximum Mutual Information (BMMI), etc. Therefore, sequence discriminative training will be performed to improve Rakhine ASR performance. Moreover, the size of lexicon and language model will be extended for future work.

## References


[1] K.M.M.Chit, L.L.Lin, "*Exploring CTC Based End-To-End Techniques for Myanmar Speech Recognition*", International Conference on Intelligent Computing & Optimization ICO 2020, pp 1038-10 46.

[2] M.A.A.Aung, W.P.Pa,*"Time Delay Neural network for Myanmar Speech Recognition*", In proceeding of the IEEE 18th International Conference on Computer Applications ,27th - 28th February,2020

[3] H.M.S.Naing, W.P.Pa, "*Automatic Speech Recognition on Spontaneous Interview Speech*", 16th International Conference on Computer Application 2018, Yangon, Myanmar

[4] A.N.Mon, W.P.Pa, "*Exploring the Effect of Tones for Myanmar Language Speech Recognition using Convolutional Neural Network (CNN)*", in the 15th International Conference of the Pacific Association for Computational linguistics (PACLING), August 16-18, 2017, Yangon, Myanmar.

[5] A.N.Mon, W.P.Pa, Y.K.Thu, and Y.Sagisaka, "Developing A Speech Corpus From Webs News for Myanmar (Burmese) Language", In Proceedings of the 20th Conference of the Oriential Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (OCOCOSDA 2017), Seoul, R.O.Korea, pp. 1-6, November 1-3,2017.

[6] Rakhine Broadcast news http://www.facebook.com/ArakanPrincessMedia

[7] ရခိုင်ဘာသာစကားလမ်းညွှန်, author အသျှင်စက္ကိန္ဒ, published in 1994, October. (Rakhine Guidance Book)

[8]http://drivegoogle.com/file/d/1A23zRnYjzoBBFnratpM06Dj3xkTRHAA/view?usp=drivesdk (ရခိုင်ဂန္ဓဝင်ဝေါဟာရ)

[9]http://medium.com/@jonathanhui/speechrecognition-gmm-hmm-8bb5eff8b196

[10] P. Bansal, A. Kant, S. Kumar, A. Sharda, S. Gupta, "IMPROVED HYBRID MODEL OF HMM/GMM FOR SPEECH RECOGNITION," International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008.

[11] N.Singh, A.Agrawal, R.A.Khan ''Gaussian Mixture Model: A Modeling Technique for Speaker Recognition and its Component'' Advanced Computing and Communication Techniques for High Performance Applications (ICACCTHPA-2014).

[12] A.Stolcke, "Srilm-An Extensible Language Modeling Toolkit", pp. 901--904 (2002).

[13] D.Povey, et al., "The Kaldi Speech Recognition Toolkit," Idiap, 2011.

[14] T.M.Oo, Y.K.Thu and K.M.Soe, "Neural Machine Translation Between Myanmar (Burmese) and Rakhine (Arakanese)", In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 80–88, Ann Arbor, Michigan. Association for Computational Linguistics.

# English to Pa-O Translation System for Village Development Plan (VDP) Using Rule-Based Method

Wah Wah Soe[1st], Yin Nyein Aye[2nd]
*University of Computer Studies (Taunggyi)*
*wahwahsoe@ucstgi.edu.mm, yinnyeinaye@ucstgi.edu.mm*

## Abstract

*Information and Communication Technology (ICT) helps for its great potential to increase in agriculture and similar sectors. A village development plan (VDP) is a written document that can be recognized the issues of the village. Using the VDP, the opportunities and weaknesses of a village are well-defined and will be resulted in an improved situation for the village. For arranging VDP, a lot of information has been collected for villages with the support of ICT. Nevertheless, it's very difficult to get the exact information due to the communication, especially the language barrier. This paper will be very helpful for proper decision-making, and effective implementation of village development especially in Southern Shan State. In this paper, the classification will be used from the rule-based method for collecting data and, English to Pa-O translation for the Pa-O local people. It can be applied not only to widen the ICT but also to develop the social sector.*

**Keywords:** Pa-O Language, ICT, Rule-based, VDP

## 1. Introduction

The Department of Rural Development (DRD) in the Ministry of Cooperative and Rural Development, Myanmar has been implementing Green Village Project (MSY), Village Development Plan (VDP), National Community Driven Development Project (NCDDP), Cash for Work (CfW), Village Revolving Fund Project (VRFP) and Enhancing Rural Livelihood and Income Project (ERLIP). In collecting data to implement those projects, the correct data cannot be filled due to the language barrier and it can be faced with being impossible to be able to fill the data through own decision. Pa-O nationals are the second largest population after Shan nationals who are residing in Shan State. That is why, out of the projects from DRD that are being implemented in all the townships of Shan State, English words will be translated to Pa-O words dealing with all the VDP data. Following the adoption of the RDSF in 2014, DRD developed the Comprehensive Capacity Development Plan (2014-2018), and Village Development Project (VDP) is one of the activities under the Plan. The government in its budgeting policy has emphasized a system of top-down budgeting and bottom-up planning. But much work in terms of institutional mechanism and technical know-how needs to be organized and developed at all levels for operating a participatory Bottom-up Planning Process. With the key objective to contribute to operating the National participatory bottom-up planning and budgeting process, the Village Development Project is implemented. The objective of VDP is to support and facilitate the process of people-centered participatory village development planning, identification of village development priorities, and mobilization of financing for meeting funding needs for the assessed priorities. VDP project is the one to be implemented in its way after finding out the need of the people in the village. In doing so, this system has to be used to be able to fill correct data through own decision. In the process of using this system, the classification will be used in the respective sectors from the rule-based method for all the data collected from the project.

This paper structure is as follows: Section II presents related works. Section III describes the proposed system and the theoretical background. Section IV discusses the implementation and Section V provides the results of the experiment.

## 2. Related Work

It was described as part of research on the development of a competency-based assessment system for mathematics in an Indonesian primary

school environment. It used five types of rule-based algorithms such as One R, RIPPER, PART, FURIA, and J48. The study of this paper was applied to a dataset containing 9454 real mathematics test questions collected from elementary schools in Indonesia. To fully measure the performance of the system, it was created that 10 expert teachers were used in the question classification phase and the results established that the ability and difficulty level of a question met the stated objectives of automatic classification [7].

It was stated that using the Rule-Based and Natural Language Toolkit to do Sentiment Analysis of texts, is learning humankind's written opinions, feelings, attitudes, and emotions. This paper displays the sentimental analysis process by Natural Language Toolkit and Python Libraries to find the hidden meaning in the unstructured data [6].

It was described that a new rule-based classification and prediction algorithm called uRule is used to classify uncertain data. It was stated that the uRule algorithm can handle uncertainty in both numerical and categorical data. According to test results, it was displayed that uRule has good performance even with highly uncertain data. Classification and prediction based on new measurements have been observed that the optimal splitting attribute and splitting value can be identified and used [2].

It was described that information and communication technology (ICT) acts as a catalyst for village development. It was illustrated that the village-related problems and opportunities for a village, advantages Threats, and vulnerabilities can be identified. A dynamic web-based data entry and monitoring system is a plan for this project. This system was observed that very helpful for proper decision making, implementation, and effective monitoring of different development [3].

## 3. System Design and Methodology

In this section, classified data are stored in the Microsoft Access database by using a rule-based method. In this system, using the IF-THEN rule can be classified as VDP data in the Department of Rural Development (DRD).

### 3.1. System Design

When the system starts, check admin and user. If the system user is not an admin, the system user can be seen the information wanted to know. If the system user is admin input data to the system using a rule-based method. Village Development Plan (VDP) data of English and Pa-O languages are stored in Microsoft Access Database. And then input data of English word type in the given text box and output data of translated Pa-O word are shown at the end of the system.

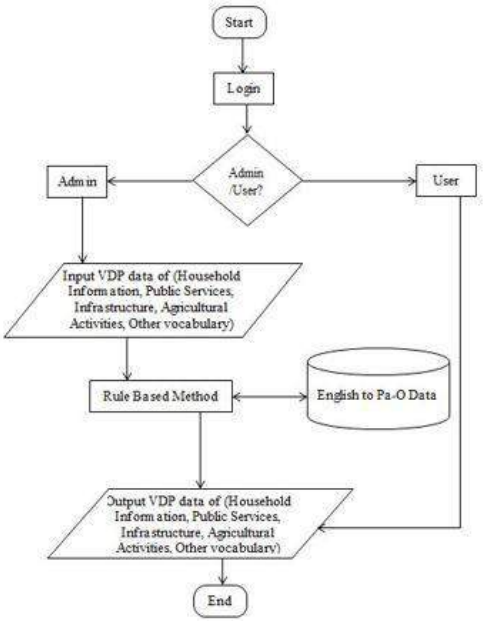



**Figure1. System Design for English to Pa-O information translation using Rule-Based Method**

### 3.2. Linguistic Architecture

In linguistic architecture, there are three basic approaches as Direct approach, the Transfer-based approach, and the Interlingua approach being used for developing machine translation (MT) systems that differ in their complexity and sophistication.

#### 3.2.1. Direct approach

In direct translation, translation is direct from the source text to the target text. It was founded that the vocabularies of source language (SL) texts were analyzed as required for the resolution of SL ambiguities, for the correct identification of

target language (TL) expressions as well as for the specification of word order in TL. Components of this system were studied that a large bilingual dictionary and a program for lexically and morphologically analyzing and generating texts [4].

#### 3.2.2. Transfer-based approach

In the Transfer approach, translation has three stages. The first stage consists in converting source language (SL) texts into an intermediate representation. The second stage consists in converting these representations into equivalent ones in the target language. In the third stage, it was founded that one of the generations of the final target text [4].

#### 3.2.3. Interlingua approach

The Interlingua approach is the most suitable approach for multilingual systems. It has two stages. Analysis (from the source language (SL) to the Interlingua) and Generation (from the Interlingua to the target language (TL)). In the analysis phase, a sentence in the source language was analyzed. It was observed that the semantic content is extracted and represented in the Interlingua form representation, where Interlingua is an entirely new language that is independent of any source or target language and is planned to be used as an intermediary internal representation of the source text. The generation of the target sentences from the Interlingua representation was followed by the analysis phase. Furthermore, the generation program for a particular TL can be used again for translation from every SL to this particular TL since it is TL-specific and not designed for input from a particular SL [4].

### 3.3. Rule-based classification method

The rule-based classification method represents a classifier utilizing a set of IF-THEN rules. The "IF" part is recognized as the rule antecedent or precondition. It consists of one or more attribute tests or conditions with AND relationships between them. The "THEN" part is the rule consequent that consists of class prediction. A rule-based classifier is a technique for classifying records using a collection of "if ... then ..." rules.

The main categories of Household Information include six types of sub-categories. There are Person information, Religious, Location, Occupation, Over-Sensitive and Financial.

IF Household Information = "yes" then Person Information = "yes" and Religious = "yes" and Location = "yes" and Occupation = "yes" and Over-Sensitive = "yes" and Financial = "yes".

The main categories of Public Services include six types of subcategories. There are Health vocabulary, Symptoms, Vaccines, Current Education, School Building, and Graduation.

IF Public Services = "yes" then Health vocabulary = "yes" and Symptom = "yes" and Vaccine = "yes" and Current Education = "yes" and School Building = "yes" and Graduation = "yes".

The main categories of Infrastructure include six types of subcategories. There are Road and Communication, Transportation, Availability of Water, Get Electricity Situation, Healthy Hygiene, and Information.

IF Infrastructure = "yes" then Road and Communication = "yes" and Transportation = "yes" and Availability of Water = "yes" and Get Electricity Situation = "yes" and Healthy Hygiene= "yes" and Information = "yes".

The main categories of Agricultural activities include five types of subcategories. There are Agriculture, Livestock, Kind of Resources, Applicable Resources, and Able of Task.

IF Agricultural activities = "yes" then Agriculture = "yes" and Livestock = "yes" and Kind of Resources = "yes" and Applicable Resources = "yes" and Able of Task = "yes".

The main categories of Other Vocabulary include six types of subcategories. There are Organization, Department, Able Man, Material Expression, Standard of Measurement, and Possible Conflict.

IF Other Vocabulary ="yes" then Organization = "yes" and Department = "yes" and Able Man = "yes" and Material Expression="yes" and Standard of Measurement = "yes" and Possible Conflict = "yes".

In the first part, the collected VDP data are classified by using the rule-based method. In the second part, the English words of VDP data are translated by using the direct approach.

## 4. Implementation

In this section, the English words of VDP's data (600) words are collected and translated to Pa-O words. Included English words (600) are for all subcategories. This system included words that are mostly used in data for VDP. In this system, five parts of main categories such as household information, public services, infrastructure, agricultural activities, and other vocabulary are included. And twenty-nine of subcategories are also included in this system. Generally, mostly used of VDP data are collected and classified to develop this system. The collected data was stored in the Microsoft Access database. The translated Pa-O words are collected from the Pa-O National Literature and Culture and also the VDP data are collected from the VDP Forms.

If the system user is an admin, it can view the translated data of Pa-O words for their relative English word, and new data of English and Pa-O words can insert, update and delete.

If the system user is a user, it can view the translated data of the Pa-O word for their relative English word.


**Figure2. English to Pa-O information translation System for User**

Figure (3) describes the five main categories of household information, public services, infrastructure, agricultural activities, and other vocabularies are included.


**Figure3. Five main categories for User**

"Household Information" can be classified into six categories such as Person Information, Religious, Location, Occupation, Over-Sensitive and Financial. First, data is chosen in the given box and translated into Pa-O language data which can be seen in the following figure (4).

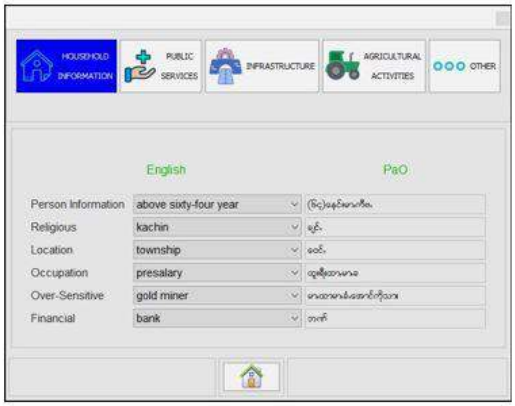



**Figure4. Choose Sub Categories of Household Information for User**

Figure (5) describes inserting new data, firstly, select whether it is household information, public services, infrastructure, agricultural activities, and other vocabularies. Secondly, an admin must select the further subcategories. After that, fill in the data to insert and press "Insert". And then, the "Are you Sure?" message alert box will appear.

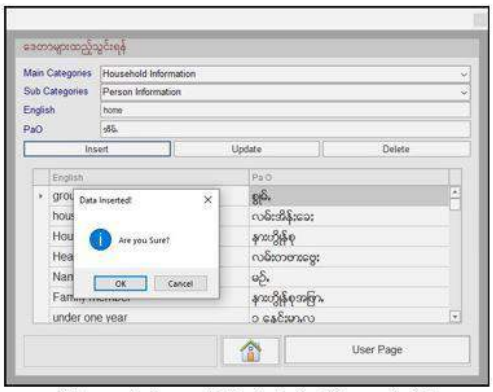



**Figure5. Insert Data into Household Information**

Figure (6) describes the inserted data in the Household Information form and can be seen in the new data added.

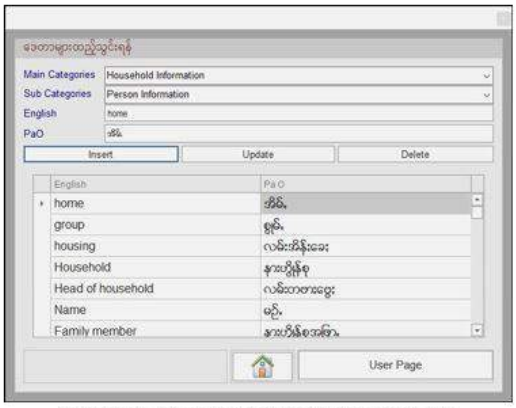



**Figure6. Insert Data into Household Information**

Figure (7) describes the data inserted in the Household Information form. If the insert data are existing in the form, the "data exist" message alert will appear.

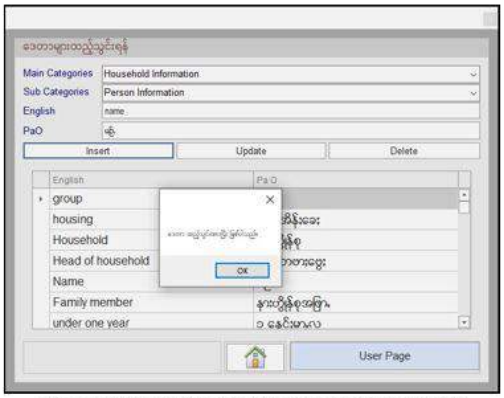



**Figure7. Duplicate Alert for Household Information**

Figure (8) describes the updated data to update. Firstly, select the data to want to update. Secondly, enter the updated data and then click update. Finally, the updated data can be seen.

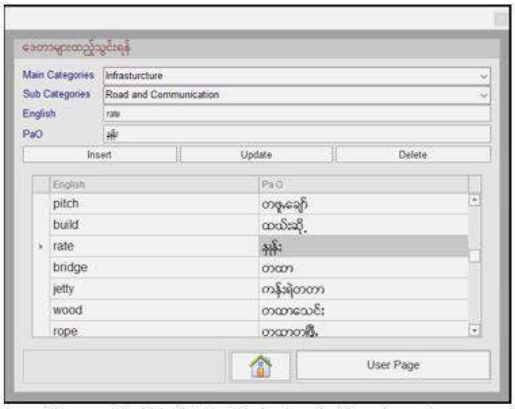



**Figure8. Update Data in Infrastructure**

Figure (9) describes the data to delete, select the data want to delete, and click delete. After deleting data can be seen in the following figure.

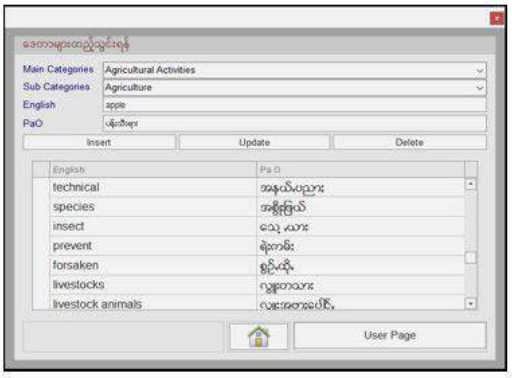



**Figure9 .Delete Data of Agricultural Activities**

## 5. Conclusion

This system is intended to know the right information for the local people's needs and wants. If this system fully utilizes, it can be very effective for local people. The self-realization of the Village Development Plan (VDP) for the local people can be created better surroundings. The local people may be entering the data correctly in the project's form. By using this system, Pa-O people will not only be able to understand the information provided by the department and fill correctly the data. Also, the organizations such as (World Bank (WB) and Asia Development Bank (ADB), etc.,) will know the actual needs and wants of the local people.

And then, they help to connect to the local people and their respective organizations.

In the future, translated from Pa-O words to English words can also be available in English and Pa-O language words and vice versa and also will be implemented with audio. It will be more effective for the local people. Currently, the rule-based method is used for VDP data according to the time limitation. In approaching machine learning, if one of the best machine translation methods is chosen and applied for developing the system, it will be more effective in future.

## References

[1] B. Bringmann, S. Nijssen, A. Zimmermann, "Pattern-Based Classification: A Unifying Perspective" proceedings of the ECML/PKDD-09 workshop (LeGo-09), Bled, Slovenia, pp36–50, (2009).

[2] B. Qin, Y. Xia, S. Prabhakar, Y. Tu, "A Rule-Based Classification Algorithm for Uncertain Data", IEEE International Conference on Data Engineering, 1084-4627/09 $25.00 © 2009 IEEE DOI 10.1109/ ICDE.2009.164.

[3] J. Furnkranz, "Rule-based Methods", Encyclopedia of Systems Biology,pp1883–1888, 2013.

[4] J. Hutchins, "Machine Translation: A Brief History, Concise History of the Language Sciences: From the Sumerians to the Cognitivists." E. F. K. Koerner and R. E. Asher (ed.). Oxford: Pergamon Press, pp. 431- 445, (1995).

[5] K. C. Sahu, "Use of ICT on village development plan (VDP)", IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), 2017.

[6] M. S. Solanki, "Sentiment Analysis of Text using Rule Based and Natural Language Toolkit", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12S, October 2019.

[7] U. L. Yuhana1,S. Rochimah , E. M. Yuniarno, A. Rysbekova, A.Tormasi, L. T. Koczy and M. H. Purnomo, "A Rule-Based Expert system for automatic question classification in mathematics adaptive assessment on indonesian elementary school environment", ICIC International 2019 ISSN 1349-4198,Volume 15,Number 1, February 2019 pp. 143–161.

# Grapheme-to-Phoneme Conversion for Foreign Words in Myanmar Language

Swe Zin Aung, Aye Mya Hlaing
*University of Computer Studies, Yangon*
*swezin.aung@ucsy.edu.mm, ayemyahlaing@ucsy.edu.mm*

## Abstract

*Grapheme to Phoneme Conversion (G2P) is the task of generation the pronunciation on a given input word. Pronunciation lexicon is one of the most important things for building automatic speech recognition (ASR), Text-to-Speech Systems (TTS).G2P conversion model is implemented for foreign words in Myanmar language using n-gram language modeling and Weighted Finite State Transducer (WFST) based approach. Firstly, we build Pronunciation Dictionary for foreign words in Myanmar language. After that, we generate the alignment of the corresponding grapheme and phoneme sequence pairs on that dictionary. A joint n-gram model was trained based on joint Grapheme ↔ Phoneme chunks aligned during the training process. Finally, the joint n-gram model is converted to an equivalent Weighted Finite State Transducer (WFST). The performance of the model has been evaluated based on Phoneme Error Rate (PER). To ensure the validity of manually prepared pronunciation dictionary and the consistency of the performance of the G2P model, we applied 10-fold cross validation and 2.36%.in average phoneme error rate (PER) was obtained for a test set.*

**Keywords:** Grapheme to Phoneme Conversion, Myanmar, G2P, N-grams, WFST

## 1. Introduction

There is no pronunciation dictionary for foreign words in Myanmar language. So, we needed to extend to build G2P conversion model for foreign words. Grapheme-to-Phoneme (G2P) Conversion is the process of automatically generated phoneme symbols form of unseen words of its pronunciation. For example, given a foreign word in Myanmar Language "ကေဘး", the process is to generate its pronunciation "**k a- hp ei**:''. It is an important in Nature Language Processing such as automatic speech recognition (ASR) and text-to-speech (TTS) development for Myanmar language. G2P conversion model is intent to get the correct pronunciation of a given input sequence.

The G2P conversion problem is composed to three process such as (1) G2P alignment for input sequence that align the grapheme and phoneme chunks pairs in a training dictionary. (2) Model training by WFST based on N-gram language model that process is to generate new pronunciations for foreign words in Myanmar Language (3) Decoding is to find the best pronunciation given the model.

## 2. Related Work

Soky et al. [8] researched Khmer G2P conversion based on weighted finite state transducer (WFST). In this paper, they presented performance of G2P on Khmer language pronunciation dictionary by comparing Rule-based and WFST techniques. The performance of the WFST based G2P is much accuracy than rule-based G2P technique. The result was obtained 3.49% in phoneme error rate (PER) or 2.98% in word error rate (WER) for test set.

A. M. Hlaing et al. [1] analyzes sequence to sequence models in G2P conversion for Myanmar language. In this paper, Myanmar pronunciation dictionary was built that is applied on sequence to sequence models such as joint sequence model, Transformer, simple encoder-decoder, and enabled encoder-decoder models that were measured by in terms of phoneme error rate (PER) and word error rate(WER). The PER and WER were gained 1.7% and 1.0 % respectively.

Y. K. Thu et al. [11] investigated Myanmar G2P conversion by using four Myanmar syllable pronunciation patterns as features that can be used in a Conditional Random Field (CRF) approach. The results were shown with the Myanmar Language Commission (MLC) test data and the Basic Travel Expression Corpus (BTEC)

test data. In this paper, all four features gave rise to the highest performance. The word accuracy and phoneme (87.90% and 95.04 %) were achieved by Feature-1234 based on (MLC).

Y. K. Thu et al. [12] examine G2P conversion approaches such as Adaptive Regularization of Weight Vectors (AROW) based structured learning (S-AROW), CRF, Joint-sequence model (JSM), PBSMT, RNN, Support Vector Machine (SVM) based point-wise classification, Weighted Finite State Transducer (WFST) based on manually tagged Myanmar dictionary.G2P conversion models such as CRF, PBSMT and WFST approaches are the best performing methods. The result of phoneme error rate (PER) was obtained approximately 13% for a testing.

## 3. Building Pronunciation Dictionary for Foreign Words

The following steps were applied for building pronunciation dictionary for G2P conversion model based on training. (1) Firstly, to build pronunciation dictionary the transliterated Myanmar Words were extracted from English-Myanmar Transliteration [13]. (2) In second step, the transliterated Foreign Words were tagged by using Sequitur-G2P Model to generate pronunciation dictionary. (3) Finally, we had been manually checked and repaired phonemes generated by Sequitur-G2P Model [1]. This step is tough and time consuming Figure1 will be explain about the process flow diagram of building pronunciation dictionary.

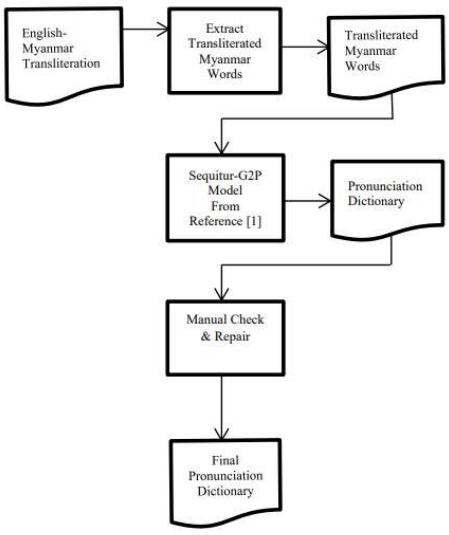



**Figure 1. Process Flow of Building Pronunciation Dictionary**

### 3.1. Final Pronunciation Dictionary for Foreign Words

The final pronunciation dictionary contains 34,000 words to cover foreign words in Myanmar language with their pronunciation. The phoneme symbols in Myanmar Language Commission (MLC) [4] and final dictionary was used to train our G2P conversion model.

In Table 1, it can be seen that pronunciations of foreign words in Myanmar language.

**Table 1 Foreign Words for Pronunciation Dictionary**

| Foreign Words | Pronunciation |
|---|---|
| ဒိုင်နာ | d a- z ain n a |
| မိုက်ကရိုဖုန်း | m ai' kh a- r o hp oun: |
| ခရစ်ယာန် | kh a- r i' j an |
| နျူးကလီးယား | n j u: k a- l i: j a: |
| ဘလော့ဂ် | b a- l o. G |

## 4. Building Grapheme to Phoneme Conversion Model

Define Myanmar G2P conversion approach in this work was implemented with Phonetisaurus [6] based on Open FST frameworks. The training procedure is then,

(1) Convert aligned sequence pairs to sequences of aligned joint label pairs, (g1:p1,g2:p2, ...,$g_n$:$p_n$);
(2) Train aligned sequence pair by using an N-gram model which MITLM [5] language modeling toolkit is applied.
(3) Convert the ARPA with N-gram model to a WFST.

Figure 2 describes about how to implement the system of G2P conversion model for training and testing with system flow diagram.

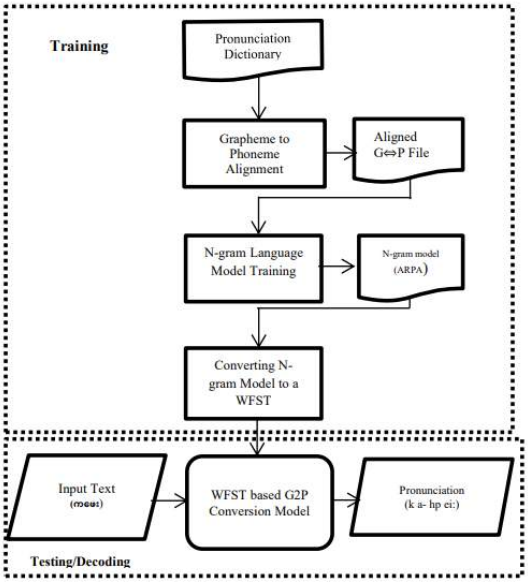



**Figure 2 System Flow for Proposed System**

### 4.1. Grapheme-to-Phoneme Alignment

The approach consists of a three-step process. First the target pronunciation dictionary is automatically aligned in [9] and detail in [10], the result of which is a chunk of aligned, joint sequences. As an example, we used four foreign words with their pronunciations shown in Table 2 and Table 3 shows the results of aligned chunks of Grapheme ↔Phoneme.

**Table 2 Sample Foreign words**

| Foreign Words | Pronunciation |
|---|---|
| ကာလာ | k a l a |
| ကာဗာ | k a b a |
| ကာဗွန် | k a b un |
| ကော်တွန် | k o t un |

**Table 3 aligned G↔P sequences. Where symbol “,” indicates a one-to-many or many-to-one relationship.**

| Aligned Entry |
|---|
| က,ာ\|\|\| k လ\|\|\| a,l ာ\|\|\| a |
| က,ာ\|\|\| k ဗ\|\|\| a,b ာ\|\|\|a |
| က\|\|\| k ာ\|\|\| a ဗ,ွ\|\|\| b န,်\|\|\|un |
| က,ေ\|\|\|k ာ,်\|\|\|o တ,ွ\|\|\|t န,်\|\|\|un |

### 4.2. Joint Sequence N-gram Model.

A n-gram is the number of sequence pair and count of a word based on the occurrences of a given trained input data with aligned G↔P pairs in a corpus. A model cannot consider at previous chunk depend on a word occurs is called unigram**.** If a model considers only from the previous chunk to predict the current word, then it is called bigram and also if two previous words are considered situation then it is a trigram model.

**Bigram:** $$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

**N-gram:** $$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

$W_n$ is current state, $W_n - 1$ refers to the previous words from current state

Figure 3 shows the generated ARPA file of the 2-gram probability. Total counts of 2-grams and 3-grams trained on our pronunciation dictionary are 8637 and 38095.

```
data\n gram 2=15
-0.651243      က|ာ}k ဗ}a|b      1.499812629
-0.651243      က|ာ}k လ}a|l      1.499812629
 -0.409048     က|ေ}k ာ|ိ}o     0.942037544
-0.425871      က}k ာ}a          0.980780913
\end\
```

**Figure 3 ARPA Format with 2 Grams Probabilities Calculation**

Joint-sequence model [7] is the most popular among many different approaches that have been proposed.

#### 4.2.1. Kneser-Ney Smoothing

Aligned G2P pair of the dataset was trained by MIT LM toolkit that is used to build n-gram language model with Kneser-Ney Smoothing. It's based on the concept of absolute discounting that is removed from all non-zero counts that word does not exist in the vocabulary. This situation occurred in test data that could never see during training process. It was applied calculate some probability value from 4-grams or 3-grams to simpler unigram models. The formula for applied to a bigram language model is presented below:

$$P_{KN}(w_i \mid w_{i-1}) = \begin{cases} \dfrac{C(w_{i-1}\, w_i) - D}{C(w_{i-1})}, & \text{if } C(w_{i-1}\, w_i) > 0 \\ \propto (w_i) \dfrac{|\{w_{i-1} : C(w_{i-1}\, w_i) > 0\}|}{\sum_{w_i} |\{w_{i-1} : C(w_{i-1}\, w_i) > 0\}|}, & \text{otherwise} \end{cases}$$

D=assume that the discount is a constant D

## 4.3. Converting ARPA To WFST

The final preparatory step is then to convert the resulting n-gram model to an equivalent WFST [3]. The WFST-based model was utilized to produce pronunciation hypotheses for chunk words by first transforming the target word into an equivalent finite-state machine and composing it with the model. When converting ARPA with n-gram probability that have $\log_{10}$ into WFST with $\log_e$, the $\log_{10}$ value must be needed to product the value with 2.303 according to the formula $\log_e x = 2.303 \log_{10}$.

### 4.3.1. Weighted Finite State Transducer

Weighted Finite-State Transducer (WFST) is the similar Finite State Transducer system relies on OpenFst [2]. Finite State transducer is a finite automaton whose state transitions are regarded as labeled with both input and output symbols. Transition in WFST machine is intent to encode a weight, and both the input labels in the FSA and output label. If an arc does not have nothing input or output, it will be marked with <eps>, ε or "-".

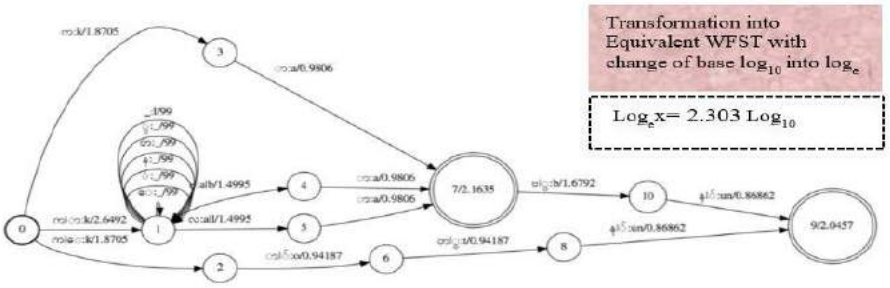



**Figure 4 Example of Weighted Finite State Transducer**

According to an example in Figure 4, FST are composed five features set (Q, q, F, Σ, Δ).
Q = {0,1,2,3,4,5,6,7,8,9,10}
q = {0}
F = {7, 9}
Where Q= State, q=Start State, F=Final State, Σ =input graphemes, Δ=output phonemes.

### 4.3.2. Decoding

Decoding is the process to predict the input grapheme to its pronunciation. Decoding is created a composition phoneme lattice with a single shortest path through the input word and the decoding process is summarized in Equation1.

$$H_{best} = \text{ShortestPath}(\text{Min}(\text{Det}(\text{Project}_o(w \circ M))) \quad (1)$$

Where, $H_{best}$ refers to the best pronunciation hypothesis given the model. 'w' is a linear FSA state with the target word and M refers to the WFST constructed from the joint G2P N-gram model. 'M' is a WFST-based representation of the joint n-gram model**.** The '∘' operator refers to weighted composition that combines transducers at different levels. $\text{Project}_o(\cdot)$ refers to projecting the output labels.

## 5. Experiments

In this section, we will report experimental results of G2P performance on foreign words pronunciation dictionary described in Section 3.

## 5.1. Data Set

The Foreign pronunciation dictionary is divided training data into 10 sets (3,400 words for each). Nine sets are set for training the G2P model, and the left one set is utilized for testing. This method is called 10-fold cross validation.

## 5.3. Evaluation

To evaluate the performance of the G2P approaches based on phoneme error rate (PER) by using SCLITE [14] (score speech recognition system output) program from the NIST scoring toolkit SCTK. The formula for PER is as follows:

$$PER = (I + D + S) * 100/N$$

where,

I =the number of insertions,

D =the number of deletions,

S =the number of substitutions,

$N_P$=the number of phonemes in the reference

The detail results of each model were shown in Tables 5 and we got average 2.36% of PER according to 10 folds cross validation.

**Table 4. Performance (PER) of G2P Conversion Model**

| Models | Phoneme Error Rate (%) |
|---|---|
| 1 | 2.8% |
| 2 | 1.6% |
| 3 | 2.1% |
| 4 | 1.6% |
| 5 | 1.9% |
| 6 | 2.8% |
| 7 | 2.5% |
| 8 | 2.7% |
| 9 | 2.8% |
| 10 | 2.8% |
| **Average** | **2.36%** |

As we can see on table 4, the performance of ten models gets similar accuracy which can prove the validity of manually prepared pronunciation dictionary and the consistency of the performance of the model.

## 6. Conclusion and Future Work

Building a G2P engine to generate the grapheme to phoneme of foreign words is an important step for Myanmar ASR and TTS development. We developed a WFST based G2P model for automatic G2P conversion of foreign words in Myanmar language. We built a pronunciation dictionary of foreign words and it consists of 34,000 entries. Joint N-gram language modeling and Weighted Finite State Transducer (WFST) based approach are applied in modeling G2P conversion. 10-fold cross validation is done on our pronunciation dictionary.

In our future work, we will extend our pronunciation dictionary for wide coverage and different Grapheme-to-Phoneme Conversion techniques will be applied on the dictionary.

## References

[1] A. M. Hlaing, W. P. Pa "Sequence-to-Sequence Models for Grapheme to Phoneme Conversion on Large Myanmar Pronunciation Dictionary" (IEEE)*O-COCOSDA 2019: 1-5* [2]Allauzen and M. Riley and J. Schalkwyk and W. Skut and M. Mohri. OpenFST: A General and Efficient Weighted Finite-State Transducer Library, Proc. CIAA 2007, pp. 11-23.

[3] D. Caseiro, I. Trancoso, L. Oliveira, and C. Viana, "Grapheme to-phone using finite-state transducers," in In: Proc. 2002 IEEE Workshop on Speech Synthesis. Volume, 2002, pp. 1349–1360.

[4] Department of the Myanmar Language Commission Myanmar-English Dictionary, Yangon, Ministry of Education, 1993.

[5] Hsu and J. Glass, "Iterative Language Model Estimation: Efficient Data Structure& Algorithms", Proc. Interspeech 2008.

[6]J.Novak,"Phonetisaurusg2p,"2012.[Online]. http://code.google.com/p/phonetisaurus

[7] M. Bisani, H. Ney," Joint-sequence models for grapheme-to-phoneme conversion", Speech Communication50, 2008, p p. 434-451.

[8] Soky, Kak, X. Lu, P. Shen, H. Kato, H. Kawai, C. Vanna, and V. Chea "Building WFST based Grapheme to Phoneme Conversion for Khmer", in Proceedings of KNLP2016.

[9] S. Jiampojamarn, et.al, "Applying Many-to-Many Alignments and Hidden Markov Models to Letter to-Phoneme Conversion", NAACL HLT, 2007,pp.372379. [10] S. Jiampojamarn, G. Kondrak, "Letter-to-Phoneme Alignment: an Exploration", Proc. ACL, pp.780788, 2010.

[11] Y. K. Thu, W. P. Pa, Finch Andrew, A. M Hlaing, H. M. S.Naing, S. Eiichiro, and H. Chiori, "Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion," in the13th International Conference on Computer Applications(ICCA),Yangon,Myanmar,February,2015, pp.161-167.) [12]Y.K. Thu, W. P. Pa, Y. Sagisaka, and N. Iwahashi, "Comparison of Grapheme-to-Phoneme Conversion Methods on a Myanmar Pronunciation Dictionary," Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing ,Osaka, Japan, December, 2016,pp.11-22.

[13]https://www2.nict.go.jp/astrecatt/member/mutiyama/ALT/

[14]https://github.com/usnistgov/SCTK/

# Statistical Machine Translation System between Karen and English Language using PBSMT Model

[1]Sharo Paw, [2]Hmway Hmway Tar
[1]*Information Technology Supporting and Maintenance Department,*
[2]*University of Computer Studies Hinthada, Faculty of Computer Science Department,*
*University of Information Technology, Yangon*
*sharopaw1417@gmail.com, hmwaytar34@gmail.com*

## Abstract

*Nowadays, there are top performance of machine translation systems for some foreign languages (high resource languages). Machine Translation (MT) is the automatic translation mechanism from one natural language into another language by means of a computerized system. There are many researches using machine translation systems in not only foreign languages but also Myanmar Ethnic languages (lower source languages) such as English-Myanmar, Myanmar-Rakhine, Myanmar-Dawei and Kachin-Rawang and so on. In this system, about 10K Karen-English parallel corpus sentences are collected, the system is proposed by using the phrase-based statistical machine translation model. The purpose of the system is to translate from source to target language: from English to Karen (E-K) and from Karen to English (K-E) using phrased-based statistical machine translation (PBSMT) model. Finally, the performance of the system is measured in terms of BLEU scores.*

**Keywords**: Karen-English Parallel Corpus, PBSMT, Moses Decoder, GIZA++, SRILM, BLUE.

## 1. Introduction

Machine Translation (MT), which is also known as Computer Aided Translation, is the task of specifically designing to translate both verbal and written texts between natural languages by a computer system. The development of Statistics, Statistical Machine Translation (SMT) is becoming a popular research area in the late 1980s. SMT also translates based on the parallel corpora. This method achieved the better performance than other method such as RBMT (Rule-based Machine Translation). Word-based, phrase-based, syntax-based and hierarchical phrase-based are the approaches based on SMT. The most prevalent version of SMT is Phrase-based SMT (PBSMT), which in general includes pre-processing, sentence alignment, word alignment, phrase extraction, phrase feature preparation, and language model training. The key component of a PBSMT model is a parallel corpus. For Myanmar language, the automatic machine translation systems began in 2010. Myanmar language is an under-resourced language (known as low-resource language) and there were not many parallel corpus or monolingual corpus.

In the proposed system, phrase-based statistical machine translation (PBSMT) model is used to translate from the source language to target language (Karen-English or English-Karen). About 10K parallel corpus is collected manually via the internet, English language books published by Cambridge University and Karen language books. As preprocessing step, Moses's tokenization scripts are used to segment for English sentences and Karen sentences are segmented manually. And Moses's cleaning scripts are used for both languages. To get the translation model for each language pair, the word segmented source language was aligned with the word segmented target language using GIZA+. The alignment was symmetrized by grow-diag-final and heuristic. The lexicalized reordering model was trained with the msd-bidirectional-fe option. For language model, KenLM and SRILM are used for each language pair. Finally, the decoder describes a simple phrase-based translation model consisting of phrase-pair probabilities of translation model and language model. There are six experiments for each language pairs. Finally, the performance of

the system is measured in terms of BLEU scores. For Karen to English PBSMT model, the experimental result of KenLM with 5-gram language model is the best. And the experimental result of KenLM with 3-gram language model is the best for English to Karen PBSMT model.

## 2. Karen-English Parallel Corpus

There are main eight ethnic groups with respective languages in Myanmar Nations: Kachin, Kayah, Kayin, Chin, Mon, Burma, Rakhine and Shan.

The Karen alphabet was derived from the Burmese script as created by the help of the American missionary Jonathan Wade around the 1830s. The Karen alphabet was created for the purpose of translating the Bible into the Karen language. They are unusual among the Sino-Tibetan languages in having a subject–verb–object word order. The Karen alphabet consist of 25 consonants, 9 vowels, 5 tones and 5 medials.

**Grouped consonants**

| | | | | |
|---|---|---|---|---|
| က<br>k(kaˀ) | ခ<br>kh($k^haˀ$) | ဂ<br>gh(ɣ) | ဃ<br>x(x) | င<br>ng(ŋ) |
| စ<br>s(s) | ဆ<br>hs($s^h$) | ၐ<br>sh(ʃ) | ည<br>ny(ɲ) | တ<br>t(t) |
| ထ<br>ht($t^h$) | ဒ<br>d(d) | န<br>n(n) | ပ<br>p(p) | ဖ<br>hp($p^h$) |
| ဘ<br>b(b) | မ<br>m(m) | ယ<br>y(j) | ရ<br>r(r) | လ<br>l(l) |
| ဝ<br>w(w) | သ<br>th(θ) | ဟ<br>h(h) | ဿ<br>vowel holder(ʔ) | ဧ<br>ahh |

**Figure 1. Grouped consonants of Karen Language**

**Vowels**

| ◌ါ | ◌ီ | ◌ၢ | ◌ိ | ◌ု | ◌ူ | ◌ဲ | ◌ိဳ | ◌ီဳ |
|---|---|---|---|---|---|---|---|---|
| ah (a) | ee (i) | uh (ɣ) | u (ɯ) | oo (u) | ae or ay (e) | eh (æ) | oh (o) | aw (ɔ) |

**Figure 2. Vowels of Karen Language**

**Tones**

| ◌ၢ်<br>(အၢသံ) | ◌ာ်<br>(အးသံ) | ◌း<br>(ဖျၣ်ဆံး) | ◌ၣ်<br>(ဟးသံ) | ◌ၤ<br>(ကၡဳ်ဖိ) |
|---|---|---|---|---|

**Figure 3. Tones and their description of Karen Language**

**Medials**

| ◌ျ(hg) | ◌ွ(y) | ◌ြ(r) | ◌ျ(l) | ◌ွ(w) |
|---|---|---|---|---|

**Figure 4. Medials of Karen Language**

Karen language is regarded as a low resource language, so there are some difficulties to build a parallel corpus. Building a Karen-English parallel corpus for Karen language is conditioned by various factors like the availability of the texts in that language. Some parallel sentences are collected from Karen-English published books via internet. Some parallel sentences are collected by translating manually from English to Karen language or Karen to English languages. Therefore, Karen-English parallel corpus is a general domain. The corpus consists of over 10K parallel sentences collected from different domains.

## 3. Machine Translation (MT)

MT's concept is to translate from source language to target language by using computer software. MT can translate a huge amount of source language to target language rapidly and save less costs and more accurately than using human translator. There are four types of MT. They are: Statistical Machine Translation (SMT), Rule-based Machine Translation (RBMT), Hybrid Machine Translation or HMT, Neural Machine Translation or NMT.

### 3.1. Statistical Machine Translation(SMT)

The first ideas of SMT were introduced by Warren Weaver as far back as 1947. He explained that language had an inherent logic that could be treated in the same way as any logical mathematical challenge and identified the target (untranslated) language based on what already existed in the source (translated) language. With the availability and effectively of using theory, SMT became a practical option.

SMT, a machine translation paradigm, where the translation from source text of translated material to target text of untranslated material by deriving parameters from the bilingual corpora analysis based on statistical models. SMT paradigm is based on something occurrence's probabilistic mathematical theory. The architecture of statistical machine translation is shown in Figure 5.

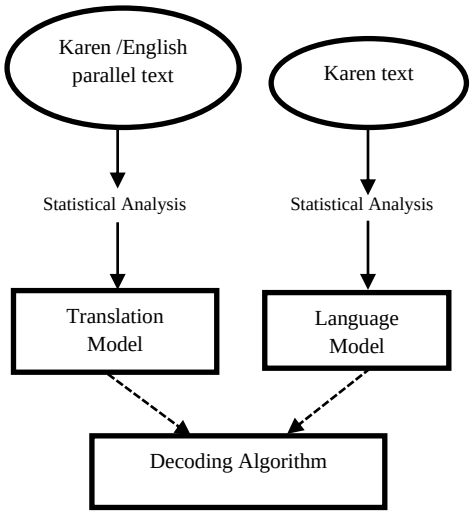



**Figure 5. Sample Architecture of Statistical Machine Translation**

### 3.1.1. Phrase-Based Statistical Machine Translation Model (PBSMT)

In phrase-based translation, the aim is to reduce the restrictions of word-based translation by translating whole sequences of words, where the lengths may differ. The sequences of words are called blocks or phrases, but typically are not linguistic phrases, but phrases found using statistical methods from corpora. It has been shown that restricting the phrases to linguistic phrases decreases the quality of translation. The chosen phrases are further mapped one-to-one based on a phrase translation table, and may be reordered. This table can be learnt based on word-alignment, or directly from a parallel corpus. The second model is trained using the expectation maximization algorithm, similarly to the word-based IBM model.

In PBSMT, the input sentence is segmented into a number of sequences of consecutive words (so-called phrases). Each phrase is translated into a target phrase. Target phrases in the output may be reordered. Bayes rule is applied to reformulate the translation.

$$\text{argmaxt } p(t|s) = \text{argmaxt } p(s|t)\ p(t) \quad (1)$$

where:

s = source sentence
t = target sentence
translation model - p(s|t)
language model - p(t)

Decomposition of the translation model:

$$p(S_1^i|t_1^i) = \prod_1^i \emptyset\,(S_{i|}t_i)d(start_i - end_{i-1} - 1) \quad (2)$$

Phrase translation is modeled by a probability distribution. Reordering of the source output phrases is modeled by a relative distortion probability distribution.

where:

- $start_i$ denotes the start position of the source phrase that was translated into the ith target phrase
- $end_{i-1}$ denotes the end position of the source phrase that was translated into the (i-1)th target phrase

In order to calibrate the output length, introduce a factor ω (called word cost) in addition to the trigram language model pLM. Usually, this factor is larger than 1, biasing toward longer output. The best output sentence ebest given an input sentence f according to our model is:

$$e_{best} = argmax_t p(t|s)$$

$$= argmax_t\ p_{TM}(s|t)p_{LM}(t)\omega^{length(e)} \quad (3)$$

$$= argmax_t \prod_{i=1}^{i} \emptyset\ (s_i|t_i)\ d\ (start_i - end_{i-1} - 1)\ \prod_{i=1}^{i} p_{LM}(t)\ \omega^{length(e)} \quad (4)$$

## 4. System Flow Diagram and Experimental Results

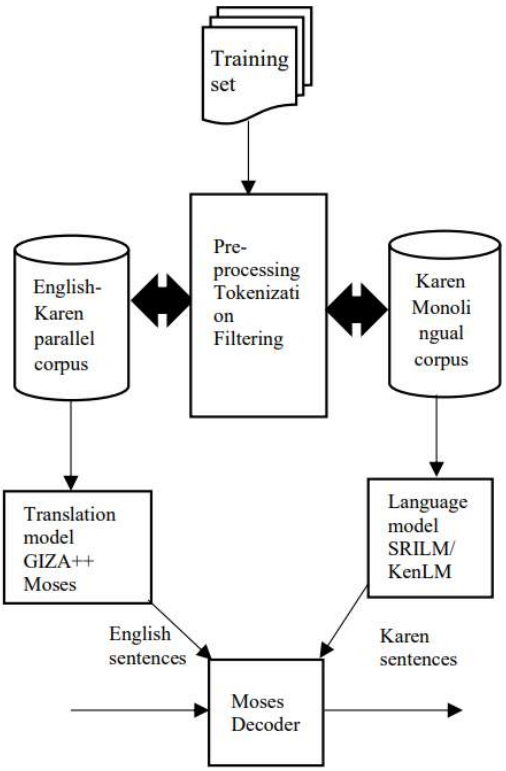



**Figure 6. System Flow Diagram of English to Karen Languages**

This section describes flow diagram of the proposed system, the dataset, preprocessing steps and the Phrase-based Statistical machine translation models.

### 4.1. Dataset

In proposed system, firstly, 10K Karen-English parallel corpus sentences are collected manually English sentences from three books: a course in spoken English, dialogue activities: Exploring spoken interaction in the language class and U Aung Hein Kyaw: English speaking designed for self-study. And translated Karen sentences are collected from Karen language books. The corpus is randomly divided into training data, development data and test data. This corpus is general corpus covering difference domains. Table 1 shows data statistics used for the experiments.

**Table 1. Data Statistics of the Corpus**

| | Parallel Sentences |
|---|---|
| Total Number of Sentences | 11500 (over 11k) |
| Training File | 10000 |
| Development File | 1000 |
| Testing File | 500 |

### 4.2. Preprocessing Step

Like Myanmar Language, Karen is an unsegmented language and there is no clear definition of word boundaries. It does not contain white space to delimit the words like English. Tokenization, called word segmentation, is not a trivial task for Karen text, same as other Asian languages. For Karen language, the proper text segmentation is lack. Therefore, Karen sentences are manually segmented in the system. For English language, Moses's tokenization script is used to segment English sentence. And as the preprocessing step, Moses's clean-corpus script is also used for both languages.

### 4.3. Model

Karen-English Phrase-based Statistical Machine Translation system is implemented by using Moses's Statistical Machine Translation System. Moses is a statistical machine translation system that allows automatically train translation models for any language pair.

The word segmented source language was aligned with the word segmented target language using GIZA++. The alignment was symmetrized by grow-diag-final and heuristic. GIZA++ is an extension of the program GIZA which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU). GIZA++ includes a lot of additional features. The extensions of GIZA++ were designed and written by Franz Josef Och. GIZA++ is used to train IBM Models 1-5 and an HMM word alignment model. Alignment models depending on word classes.

The lexicalized reordering model was trained with the msd-bidirectional-fe option. Lexicalized reordering models play a crucial role in phrase-based translation systems. They are usually learned from the word-aligned bilingual corpus by examining the reordering relations of adjacent phrases. The system experiments with KenLM and SRILM for training the 2-gram, 3-gram and 5-gram language models. Therefore. there are six model for each direction, namely, 2gramKenLM, 2gramSRILM, 3gramKenLM, 3gramSRILM, 5gramKenLM and 5gramSRILM. Minimum error rate training (MERT) was used to tune the decoder parameters and the decoding was done using the Moses decoder.

### 4.4. Experimental Results and Discussion

For the evaluation result of the translation output, the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) is used. In Karen to English phrase-based statistical machine translation, it is clear that the evaluation results of 5gramKenLM model are much better than those of the others. Table 2 shows the evaluation results between Karen and English phrase-based statistical machine translation models.

It is observed that 5gramKenLM model is effective when translating into English, obtaining a BLEU score of 22.50. In English to Karen phrase-based statistical machine translation system, 3gramKenLM model is the best.

**Table 2. Evaluation Results (BLEU) of Karen-English SMT system**

| Model | Kr-En (BLEU) | En-Kr (BLEU) |
|---|---|---|
| 2gramKenLM | 21.96 | 19.77 |
| 2gramSRILM | 21.47 | 18.58 |
| 3gramKenLM | 22.18 | 20.12 |
| 3gramSRILM | 21.53 | 19.68 |
| 5gramKenLM | 22.50 | 20.05 |
| 5gramSRILM | 21.64 | 19.45 |

## 5. Conclusion

This paper contributes the Karen-English parallel corpus translation by using PBSMT from Karen to English and from English to Karen. 10K Karen-English parallel corpus are translated and then evaluated with BLEU scores results.

## References

[1] Arun Babhulgaonkar, Shefali Sonavane, "Empirical Analysis of Phrase-Based Statistical Machine Translation System for English to Hindi Language", Vietnam Journal of Computer Science, 2022.
Empirical Analysis of Phrase-Based Statistical Machine Translation System for English to Hindi Language | Vietnam Journal of Computer Science (worldscientific.com)
[2] Aye Thida Win, "Phrase Reordering Translation System in Myanmar-English", May 2011.
Phrase Reordering Translation System in Myanmar-English - CORE
[3] Dojun Park, Youngjin Jang, Harksoo Kim, "Korean-English Machine Translation with Multiple Tokenization Strategy", English translation of the original Korean thesis in KCC2021 Undergraduate/Junior Thesis Competition, 2021.
Korean-English Machine Translation with Multiple Tokenization Strategy | Papers With Code
[4] Honey Htun, Ye Kyaw Thu, Nyein Nyein Oo, Thepchai Supnithi, "English-Myanmar (Burmese) Phrase-Based SMT with One-to-One and One-to-Multiple Translations Corpora", 2020.
(PDF) English-Myanmar (Burmese) Phrase-Based SMT with One-to-One and One-to-Multiple Translations Corpora | Honey Htun - Academia.edu
[5] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, Thepchai Supnithi, "Statistical Machine Translation between Myanmar and Myeik", Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020) Yangon (Rangoon), Myanmar (Burma), 2020, pp. 36-45.
Statistical Machine Translation between Myanmar and Myeik - WCSE 2020 Spring - WCSE
[6] Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, "Statistical Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese)".
https://onlinesource.ucsy.edu.mm/bitstream/handle/123456789/359/304-311.pdf
[7] Yi Mon Shwe Sin, Khin Mar Soe, "Attention-Based Syllable Level Neural Machine Translation System for Myanmar To English Language Pair", International Journal on Natural Language Computing (IJNLC) Vol.8, No.2, 2019.
Yi Mon Shwe Sin.pdf (ucsy.edu.mm)
[8] Wei yang, Hanfei Shen, Y Ves Lepage, "Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese- Japanese Machine Translation", Journal of Information Processing, Vol.25, pp. 88-99, 2017.
(PDF) Inflating a Small Parallel Corpus into a Large Quasi-parallel Corpus Using Monolingual Data for Chinese-Japanese Machine Translation (researchgate.net)
[9] Zar Zar Linn, Ye Kyaw Thu, Pushpa B. Patil, "Statistical Machine Translation between Myanmar (Burmese) and Kayah", Journal of Intelligent Informatics and Smart Technology, Vol. 4, April 2020, pp.62-68.
http://docslib.org/doc/4010071/statistical-machine-translation-between-myanmar-burmese-and-kayah
[10] Zun Hlaing Moe, Thida San, Ei Thandar Phyu, Hlaing Myat Nwe, Hnin Aye Thant, Naw Naw, Htet Ne Oo, Thepchai Supnithi and Ye Kyaw Thu, "Myanmar Text (Burmese) and Braille (Mu-Thit) Machine Translation Applying IBM Model 1 and 2", Journal of Intelligent Informatics and Smart Technology, Vol. 5, April 2021, pp. 18-26.
https://jiist.aiat.or.th/assets/uploads/16195380940305LtWrJIIST-44-FinalVersion.pdf

# Myanmar Spelling Error Detection and Correction

Yee Mon Kyaw, Phyo Phyo Wai
*University of Computer Studies (Magway), Myanmar*
*yeemonkyaw13@ucsmgy.edu.mm, phyophyowai81@gmail.com*

## Abstract

*Natural language processing (NLP) is a branch of AI (artificial intelligence) within computer science that helps computers to understand human languages. Spell checking and correction systems are important for many NLP applications, such as Machine Translation, Text Summarization, Text to Speech, and Information Retrieval, etc. Spell checking means to detect and correct the error. The Myanmar language is the official language of our country, Myanmar. This system intends to check for Typographic, Phonetic, and Context Errors in the Myanmar Language. Syllable Dictionary Lookup approach is used for Typographic Error Detection. Dictionary and Corpus Lookup approaches are used for Phonetic Error Detection. The Levenshtein Distance Algorithm is applied for giving a suggestion list of Typographic and Phonetic errors. For Context Errors, confusion sets approach is used in error detection and Naïve Bayes Classifier is used in suggestion generation. If there is an error in the incoming sentence, a suggestion list will be given and the correct sentence will be generated. The experimental results such as error detection rate, error correction rate and accuracy results on error correction are evaluated for performance.*

**Keywords:** Spell Checker, Edit Distance, Dictionary Lookup, Confusion Sets, Naive Bayes Classifier.

## 1. Introduction

In line with the rapid development of digital content related to the Myanmar (Burmese) language, applications such as machine learning, machine translation, and information retrieval have become popular, and it requires obtaining effective Natural Language Processing (NLP) studies. Spell checking remains a significant challenge in the Myanmar language NLP field. Natural Language Processing has been around for decades and has become an everyday part of our lives. Spell checking and correction systems are important for many NLP applications like Web Search Engines, Machine Translation (MT), Text to Speech (TTS), Text Summarization (TS), Information Retrieval (IR), etc. Spell checkers typically highlight or underline the words that they find to be misspelled. Using a spell checker reduces the number of typos in user documents significantly.

Until now, NLP applications such as MT, TTS, and IR in the Myanmar language have not yet achieved accurate results. This is because they do not follow the rules of a language and use it easily. This system intends to solve the challenges of MT, TTS, and information retrieval. If only spell checker, which is one of NLP's preprocessing steps, could be performed, better results could be achieved in MT, TTS, and IR. It can support ongoing research areas of NLP related to the Myanmar Language.

This system can support typographic and phonetic, and context errors in Myanmar words. And this system can provide paragraph levels. The main parts of this system are spelling error detection, relevant suggestion generation, and error correction. By using a Multinomial Naive Bayes classifier, this system can support classifying context error words.

The rest of this paper is presented as follows. First, we describe the related word. We present theoretical background in section 3. Section 4 describes error detection process and section 5 presents error correction process. Corpus creation is presented in section 6 and explanation of Naive Bayes Classifier is presented in section 7. Section 8 describes the experimental results. Finally, we present the conclusion.

## 2. Related Work

A1-Jefri, M.A. and Mahmoud, S.A. [1] proposed a method for addressing real-word

errors in Arabic language. Context words method and n-gram model are used in their system. They used 28 confusion sets and these sets are created based on one letter variation and the same pronunciation. If a word contains in confusion sets, the system disambiguates error word using the n-gram model and four words surrounding it are considered to predict the correct word. They considered three tri-grams model. The system defines the word with highest probability as the correct word.

Fossati, D. and Eugenio, B.D. [3] described a method for context sensitive spell checking. Mixed trigram model with POS tagging can address the problem of real-word errors. With the use of data from the Penn Treebank corpus, the model has been trained and tested. Based on hit rates, false positive rates, and coverage, the performances of the system are measured. The results show that hit rates on error detection and error correction was potential results, despite a high false positive rate.

Golding, A.R. [4] established a hybrid method for context-sensitive spelling correction. Lexical disambiguation task is addressed by using confusion sets. 18 confusion sets are considered and words are collected from Random House Unabridged Dictionary. Decision lists and Bayesian hybrid methods are applied for spelling correction. Golding performed similar experiments with Bayesian classifiers and achieved slight improvement over decision lists.

Oo, N.Z. and Htwe, T.M. [6] developed a spell checker for the Myanmar language using the Levenshtein distance algorithm and dictionary lookup approach. This system emphasized word-level checking and mainly checked the spelling of Myanmar words, consulting with Animals and Plants. Typographic, cognitive, and sequence errors can be solved. The dynamic threshold algorithm is used for suggestion generation. The user can easily know how the input word is transformed into the destination word by using the transformation algorithm.

## 3. Theoretical Background

The Myanmar language is an official language of the Republic of the Union of Myanmar. It is also known as Burmese. It is spoken by 33 million people and also the 38th most spoken language in the world. Myanmar has a syllabic writing system and also a very rich language. The Myanmar language is written horizontally from left to right. Myanmar's basic set of symbols consists of 33 consonants and 14 vowels. Myanmar writing is different from other languages because its writing does not use white spaces between words or syllables. So, it is difficult to tokenize words. Although each Myanmar word can be identified by word boundary correctly, if these words are not in the dictionary, they are called "error words". The most common causes of spelling mistakes are typing errors in characters and phonetic similarity [17].

Spell checking is a popular task in Natural Language Processing that deals with error detection and correction of spelling errors. Spell checker may be stand-alone, capable of operating on a block of text, or as part of a larger application, such as a word processor or search engine. There are two main issues related to the spell checker: error detection and error correction. While the "Error Detection" function identifies the misspell words in the text, the "Error Correction" suggests corrections for the misspelled word [11]. Spelling errors techniques were developed based on various spelling error tendencies, often known as error patterns. Techniques for spelling error detection were designed on the basis of different spelling error trends which are also called error patterns. According to these investigations, spelling errors are classified as Typographic and Cognitive errors [8]. Next section describes the types of spelling errors in this system.

### 3.1. Types of Spelling Errors

There are three types of errors can be handled in this system. They are typographic errors, phonetic errors, and context errors.

**Typographic Errors**: These errors occur when the correct spelling of a word is known but the word is typed mistakenly. These errors are mostly related to typing when a word is written incorrectly because a finger was placed on the wrong key of the keyboard. These errors are called non-word errors. Typographic errors are classified into four types: insertion error, deletion error, substitution error, and transposition error

[2]. Examples of Typographic errors are shown in Table 1.

**Table 1. Examples of Typographic Errors**

| Misspelled Words | Possible Words | Error Reasons |
|---|---|---|
| ကျောင် | ကျောင်း | Letter "ား" missing |
| ကြေါင် | ကြောင် | Error of letter "ါ" |
| ကြောင်းး | ကြောင် | Extra of letter "ား" |

**Phonetic Errors**: These errors occur when a writer does not know or has forgotten the correct spelling of a word. These errors are called cognitive errors. Phonetic error (cognitive error), which is pronounced the same as the intended word but the spelling is wrong. Cognitive errors can be understood from the following sentence; "နေ့လည်မှာအစည်းအဝေးရှိတယ်။". In this sentence, user types နေ့လည် as one word. When we combined the two words (နေ့ and လည်), they transformed cognitive errors (phonetic errors) because နေ့ and လည် have no combination. Taking the context in to consideration the correct sentences is "နေ့လယ်မှာအစည်းအဝေးရှိတယ်။". Examples of phonetic error words and correct words are shown in Table 2.

**Table 2. Examples of Error Words and Correct Words**

| Error words | Correct words |
|---|---|
| နေ့လည် | နေ့လယ် |
| ကြမ်းတန်း | ကြမ်းတမ်း |
| စိမ်းလမ်း | စိမ်းလန်း |
| ဂိုထောင် | ဂိုဒေါင် |
| ထိမ်းသိမ်း | ထိန်းသိမ်း |
| စျေးနှုံး | စျေးနှုန်း |
| ထုတ်နှုတ် | ထုတ်နုတ် |

**Context Errors**: There errors occur where the word is pronounced the same as the intended word, but the word is ambiguous for the input sentence. In the Myanmar language, when the syllables are combined, it becomes a word. A single syllable can also be a word. Context ambiguous words are best described as words that sound the same but have a different meaning [9]. Some ambiguous words and their meanings are described in Table 3.

**Table 3. Examples of Context Ambiguous Words**

| No. | Context Words | Descriptions |
|---|---|---|
| 1 | ကဏ်း | ပုစွန်လုံးတစ်မျိုး။ |
| | ဂဏ်း | ကိန်းအရေအတွက်။ |
| 2 | ကတိ | သဘောတူညီချက်။ |
| | ဂတိ | သတ္တဝါတို့လားရာဘဝ။ |
| 3 | ယဉ် | သိမ်မွေ့သောအမူအရာရှိသော။ပြေပြစ်သော။ နှစ်သက်ဖွယ်ဖြစ်သော။ ထိတွေ့မှုများ၍ရိုးနေပြီဖြစ်သော။ |
| | ယာဉ် | စီးနင်းသွားလာရန်၊ကုန်ပစ္စည်းသယ်ယူပို့ဆောင်ရန်ဆွဲအား၊တွန်းအား၊စက်အားဖြင့် ရွေ့လျားသောအရာမျိုး။ |

## 3.2. Segmentation

Text segmentation is the preprocessing step of the spell checker. Text segmentation is the process of breaking down a large text corpus into its individual words and sentences. Word segmentation is the process of splitting a large sample of text into words by locating the word boundaries, the points where one word ends and another begins. The identified words are usually referred to as tokens in computational linguistics, and word segmentation is also known as tokenization [14].

There are two types of segmentation in this system. They are syllable segmentation and word segmentation. Syllable segmentation process is used for Typographic error detection and word segmentation is used for Phonetic and Context error detection. For syllable segmentation and word segmentation, regular expression method and Maximum Matching (Longest Matching) algorithm are applied in this paper.

In this system, syllable segmentation is the preprocessing step of typographic error detection.

Myanmar is a low-resource language and therefore it is difficult to develop a good word segmentation tool based on machine learning techniques. In this paper, Regular Expression approach [15] is used for syllable segmentation.

For example: Syllable Segmentation with regular expression.

| Unsegmented Input: | Segmented Output: |
|---|---|
| ကားဖြင့်သွားသည်။ | ကား ဖြင့် သွား သည် ။ |
| မင်္ဂလာပါမိတ်ဆွေ။ | မင်္ဂ လာ ပါ မိတ် ဆွေ ။ |
| ဝီရိယမရှိဘူး။ | ဝီ ရိ ယ မ ရှိ ဘူး ။ |

In the Myanmar language, words composed of single or multiple syllables. Word segmentation is the preprocessing step of phonetic and context errors detection. Syllable level longest matching algorithm and Myanmar dictionary files are used for word segmentation. This algorithm is proposed by Htay, H.H. and Murthy, K. N. [5]. For example,

**Input Sentence:**
မမသည်ကဏ္ဍန်းဟင်းကိုအလွန်ကြိုက်သည်။

**After word segmentation:**
မမ သည် ကဏ္ဍန်း ဟင်း ကို အလွန် ကြိုက် သည် ။

Error detection process and error correction process are explained in the next section.

## 4. Error Detection Process

In general, spelling errors can be classified into two categories: non-word and real-words errors [8]. Many techniques are available for non-word errors. The error detection procedure typically consists of identifying if an input string is a valid index or dictionary word. Efficient techniques for detecting such errors have been developed. Dictionary lookup and n-gram analysis are two of the most well-known techniques [7].

There are three types of errors, typographic, phonetic, and context errors, which can be detected by the system. Dictionary lookup technique is used in typographic and phonetic error detection. Dictionary and corpus files are used in error detection and error correction process. The detail of corpus creation is explained in the next section.

**Typographic error detection**: Typographic errors are also called non-word errors [9]. Non-word errors occur when words do not exist in the dictionary. Dictionary lookup approach is used for typographic error detection. Typographic errors occur at the syllable level. After syllable segmentation, the system checks each syllable in the syllable-dictionary. If a syllable contains in a syllable-dictionary, pass it to the next word and otherwise, it is said to be an error.

**Phonetic error detection**: In the Myanmar language, words composed of single or multiple syllables [9]. Phonetic errors occur when syllables that sound the same are used incorrectly. To detect these types of errors, the system uses general knowledge of phonetic errors. Common phonetic error words, always typing phonetic errors as ones corrected by people, are collected from common misused words. Common phonetic error words are created as a corpus file. Myanmar dictionary and error file are used in word segmentation for the input sentence. For example,

Input Sentence:
နေရောင်တွင်စပါးခင်းများစိမ်းလမ်းနေသည်။
After word segmentation:
နေရောင် တွင် စပါး ခင်း များ စိမ်းလမ်း နေ သည်။
In the example sentence, user types စိမ်းလမ်း as one word. စိမ်းလမ်း is not included in Myanmar dictionary and it has no meaning. The correct word is စိမ်းလန်း and its definition is လန်းဆန်းသောအစိမ်းရောင်ရှိသည်။

After word segmentation, the system checks each word in the word-level dictionary. If a word is in a dictionary, pass it to the next word and otherwise, it is said to be an error.

**Context error detection**: Myanmar sentence is accepted as the input sentence in context error detection. Input sentence is segmented into words using longest matching approach. After word segmentation, each word from the input sentence is checked in confusion set. A common approach to the problem is to refer to it as a disambiguation task, when the ambiguity among words is modeled by confusion sets. A confusion set is a

set of words that are likely to be confused with each other. In this paper, 35 confusion sets are used to detect context error. Context ambiguous words are collected from books: မြန်မာ စာလုံးပေါင်း သတ်ပုံကျမ်းနှင့်ခွဲထား, ငယ်ပေါင်းကြီး ဖော်မြန်မာစာနှင့် အရေးအသားပြဿနာများ, and Myanmar Dictionary, published in August, 2008. If the word contains in confusion set, this word is called context error words (ambiguous words). Correction of ambiguous words using Naive Bayes Classifier is explained in the next section.

## 5. Error Correction Process

Error correction means just to replace the incorrect word with most likely corrected one. After detecting the misspelled words, the system recommends a list of appropriate suggestions. Minimum edit distance technique can also be used to measure the similarity of two words: the shorter the distance, the greater the similarity. The minimum edit distance between two strings is defined as the number of editing operations (such as insertion, deletion, and substitution) required to transform one string into another. Minimum edit distance methods include Hamming, Levenshtein, and the longest common subsequence. Levenshtein distance algorithm is applied for typographic and phonetic errors in this system.

**Typographic error correction**: If there is a typographic error in the incoming sentence, the system calculates the edit distance between the error syllable and the syllable from the dictionary using Levenshtein distance. After calculating the edit distance, those who have edit distance 1 will be shown as suggestion. After giving a suggestion list, the user chooses the correct word for misspelled word. And then, error word substitutes with the corrected word and generates the corrected sentence.

**Phonetic error correction**: If there is a phonetic error in the incoming sentence, the system calculates the edit distance between the error word and word from the dictionary using Levenshtein distance algorithm. The error words and correct words will have a Levenshtein distance less than or equal to 3 which are considered to get more similar Myanmar words. After calculating the edit distance, the system ranks suggestion list with the edit distance. And the system chooses the best similar word and generates the corrected sentence.

**Context error correction**: If the word is in confusion set, this word is called context ambiguous word. In this system, Naive Bayes method is used in addressing of context ambiguous. The about of Naive Bayes classifier is explained in the next section.

Context error correction process consists of four main parts. They are:

1. Preprocessing of Input sentence
2. Retrieving related sentence on Corpus using Confusion set and Text preprocessing
3. Calculating Probability based on the Bayes Theorem
4. Choosing the correct sense (Calculating maximum scores using the Bayes Theorem).

Firstly, the system accepted the input sentence, including context ambiguous words. In the preprocessing stage, Tokenization, Stopwords removal and Context word removal are performed for input sentence. Stopwords are a set of common or general terms in a language. Stopwords include pronouns, conjunctions, prepositions, interjections, etc. After the preprocessing step, the system uses the remaining words in the input sentence as features for predictive text.

Secondly, the system also retrieved the related sentences from the corpus using the Confusion set and text preprocessing is performed as the training data.

Thirdly, the system calculated prior probability and likelihood based on the Bayes Theorem.

Finally, the system computes the score of each ambiguous word on a given feature and decides the most appropriate one for a given ambiguous word in the test sentence.

## 6. Corpus Creation

A corpus is a collection of machine-readable authentic texts that have been sampled to be representative of a particular natural language or language variety, though "representativeness" is a relative concept. Corpora are vital in NLP research as well as a wide range of linguistic studies. They provide a material basis and a test

bed for building NLP systems [10]. Corpus and Dictionary files are mainly used in the system.

In this system, data are collected for three types of errors. The syllable-level dictionary file which consists of 2653 syllables and is collected from github, is used for checking typographic errors and calculating edit distance [12].

For phonetic error, word-level dictionary and corpus files are used. Dictionary file is used for checking phonetic errors and it also used for segmented words for the input text. 32917 words dictionary file is used and it is also collected from github [13]. Corpus file consists of 650 phonetic errors words. It is manually created based on the most common phonetic errors and this is collected from books: မြန်မာစာလုံးပေါင်း သတ်ပုံကျမ်း နှင့်ခွဲထား, ဃ်ပေါင်းကြီးဖော် မြန်မာစာနှင့် အရေးအသား ပြဿနာများ, and Myanmar Dictionary, published in August, 2008.

Predefined 35 confusion sets are used in context-error detection. These words were collected from the most common confused words. Training sentences are used in context error correction using Naive Bayes Classifier. These sentences are collected from Middle School Myanmar Textbooks.

## 7. Naive Bayes Classifier

The Naïve Bayes algorithm is a supervised learning technique that uses the Bayes theorem to solve classification issues. It is mainly used in text classification that includes a high-dimensional training dataset. The Nave Bayes Classifier is a simple and effective Classification algorithm that aids in the development of fast machine learning models capable of making quick predictions. It is a probabilistic classifier, which means it predicts based on an object's likelihood.

Multinomial Naive Bayes, Bernoulli Naive Bayes, and Gaussian Naive Bayes are the three types of Naive Bayes models. In this approach, Multinomial Naive Bayes is utilized in context errors correction. Because of its simplicity, speed, and outstanding performance, Multinomial Naive Bayes (MNB) is a popular text classification classifier. When the data is multinomial distributed, the Multinomial Nave Bayes classifier is utilized. It is primarily used to solve document classification problems; it indicates which category a specific document belongs to, such as Sports, Politics, Education, and so on. Multinomial Naïve Bayes consider a feature vector where a given term represents the number of times it appears or very often i.e., frequency.

Naive Bayes algorithms are commonly used in spam filtering, sentiment analysis, and recommendation systems, and so on. They are fast and simple to implement, but their main disadvantage is the requirement for predictors to be independent [16].

## 8. Evaluation Results

In this system, the experimental results are shown with error detection rates, error correction rates and accuracy on error correction. Error detection rates are calculated on real errors in the testing and error correction rates are calculated on the detected errors by the system.
Error detection rate is defined as

$$Detection\ rate = \frac{the\ number\ of\ error\ detected}{actual\ number\ of\ error\ words\ in\ the\ sample}$$

Error correction rate is defined as

$$Correction\ rate = \frac{number\ of\ error\ corrected}{number\ of\ error\ detected\ by\ the\ system}$$

Error detection and correction rate on typographic error is calculated on syllable level and phonetic and contexts errors are calculated on word level.

There are two types of testing data for experimental results. These are partially seen data and unseen data. For partially seen data, testing sentences are collected from middle school Myanmar textbooks. For unseen data, other texts are collected from the Internet websites such as agriculture, beauty and health.

To know the error detection rates and error correction rate of typographic errors, results are calculated on testing sentences containing only typographic errors. Typographic errors have been tested using a total number of about 2000 syllables. Error rate and correction rates for typographic errors are calculated on syllable count.

This system can solve sentences containing both phonetic error and context error. Therefore, to know the error detection rates and error

correction rate of phonetic and context errors, results are calculated on testing sentences containing both phonetic and context errors and are calculated on word count. About 500 sentences are used to test the result. For seen data, about 2500 words are tested and for unseen data, about 2000 words are tested to get error detection and correction rates of phonetic and context errors.

For both seen and unseen data, average error detection rates on typographic errors, phonetic errors and context errors are 100%, 88% and 96% respectively. Average error correction rates on typographic errors, phonetic errors and context errors for two types of testing are 96%, 96% and 86% respectively. Accuracy results are calculated on two types of testing data for typographic, phonetic and context errors. These results are described in Table 4.

**Table 4. Accuracy of Typographic, Phonetic and Context Errors**

| Testing Data | Typographic Errors (%) | Phonetic Errors (%) | Context Errors (%) |
|---|---|---|---|
| Partially Seen Data | 96 | 96 | 86 |
| Unseen Data | 77 | 95 | 94 |

In Figure 1, accuracy of typographic, phonetic and context errors are shown with chart. Accuracy of typographic error result on partially seen data is 96% and unseen data is 77%. The accuracy result on unseen data is less because of the adoptive words, for example: ဆဲလ်, ယမ်, ဂျယ်.

Accuracy result of phonetic error on seen data is 96% and unseen data is 95%. Accuracy result of context error on seen data is 86% and unseen data is 94%. Result on seen data is less than unseen data, because the rate of context errors including in the unseen data (internet websites such as agriculture, beauty and health) is less than the partially seen data.

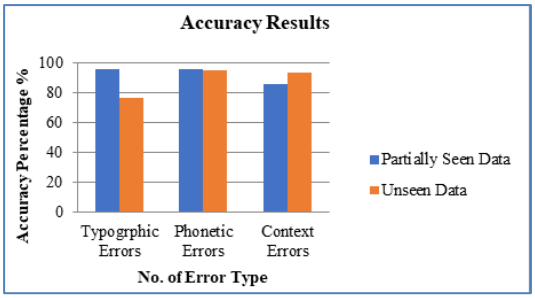



**Figure 1. Accuracy of Typographic, Phonetic and Context Errors**

## 9. Conclusion

In the field of NLP, spell checking is a popular type of research, and it is still challenging to use it effectively in the Myanmar language. In the future, we aim to build a large corpus with large amount of data and to achieve high accuracy. We implemented a Myanmar spelling error detection and correction which can handle typographic, phonetic and context errors. In this paper, corpus and dictionary lookup approaches are applied for typographic and phonetic errors detection and correction.

Finally, we conclude by showing the performance and evaluation of our system. The average accuracy results of error correction on Typographic errors, Phonetic errors and Context errors are 87%, 96% and 90%, respectively. These results show that the system can provide promising accuracy. This system can provide Myanmar NLP applications likes Machine Translation, Text to Speech, and Information Retrieval.

## References

[1] A1-Jefri, M.M. and Mahmoud, S.A. (2013), “Context-Sensitive Arabic Spell Checker using Context Words and N-gram Language Models”, Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences.

[2] Damerau, F.J. (1964), “A technique of computer detection and correction of spelling errors”, Communication of ACM, Volume7, Number3, March.

[3] Fossati, D. and Eugenio, B.D. (2007), "A Mixed Trigrams Approach for Context Sensitive Spell Checking", Proceeding of the 8th International Conference on Computational Linguistics and Intelligence Text.

[4] Golding, A.R. (1995), “A Bayesian Hybrid Method for Context-Sensitive Spelling Correction,” In Proceedings of the Third Workshop on Very Large Corpora, Cambridge, MA, pages 39-53.

[5] Htay, H.H. and Murthy, K. N. (2008), "Myanmar Word Segmentation using Syllable Level Longest Matching", Proceeding of the 6th Workshop on Asian Language Resources.

[6] Oo, N.W. and Htwe, T.M. (2010), “Myanmar Words Spelling Checking Using Levenshtein Distance Algorithm”, Fifth Local conference on Parallel and Soft Computing.

[7] Kumar, R., Bala, M. and Sourabh, K. (2018), “A study of spell checking techniques for Indian Languages”, JK Research Journal in Mathematics and Computer Sciences, Vol. 1, No. 1 , March.

[8] Kukich, K. (1992), “Techniques of Automatically Correction Words in Text”, ACM Computing surveys, Vol. 24, No. 4, December.

[9] Mon, A.M. and Thein, T. (2013), “Myanmar Spell Checker”, International Journal of Science and Research (IJSR).

[10] Indurkhya, N. and Damerau, F.J., "Handbook of Natural Language Processing", Second Edition.

[11] Pradhan, A. and Dalai, S.S. (2020), "Design of Odia Spell Checker with word Prediction", International Journal of Engineering Research & Technology (IJERT).

[12] Myanmar Syllable Dictionary, Date of Access: September 2022, https://github.com/trhura/pango-myanmar/blob/master/data/mydict-syllables.txt.

[13] Myanmar Word Dictionary, Date of Access: September 2022, https://github.com/mcfnlp/Head-Word.

[14] Indurkhya, N. and Damerau, F.J., "HANDBOOK OF NATURAL LANGUAGE PROCESSING", Second Edition.

[15] Syllable Segmentation using Regular Expression, Date of Access: September 2022, https://github.com/swanhtet1992/ReSegment.

[16] Naïve Bayes Classifier Algorithm, Date of Access: September 2022, https://www.javatpoint.com/machine-learning-naive-bayes-classifier.

[17] Burmese Language, Date of Access: September 2022, https://omniglot.com/writing/burmese.htm.

[18] မောင်ခင်မင်(ဓနုဖြူ), "ငယ်ပေါင်းကြီးဖော် မြန်မာစာ နှင့် အရေးအသား ပြဿနာများ".

[19] မြန်မာစာမြန်မာစကား, Department of Myanmar Language Commission, Ministry of Education, Union of Myanmar, August 1993.

[20] မြန်မာစာလုံးပေါင်းသတ်ပုံကျမ်းနှင့်ခွဲထား, Department of Myanmar Nationalities' Languages, Ministry of Education, Union of Myanmar, Second Edition, 2020.

[21] မြန်မာအဘိဓာန်, Department of Myanmar Language Commission, Ministry of Education, Union of Myanmar, August, 2008.

# Neural Machine Translation between Myanmar and Korean Languages

Hnin Nandar Zaw[1], Yi Mon Shwe Sin[2], Khin Mar Soe[2]
[1]*University of Computer Studies, Pyay,* [2]*University of Computer Studies, Yangon*
*hninnandarzaw@ucsy.edu.mm, yimonshwesin@ucsy.edu.mm, khinmarsoe@ucsy.edu.mm*

## Abstract

*No matter where they are in the world, individuals may now easily and affordably interact with one another because to the Internet. Additionally, Natural Language Processing (NLP) makes an effort to enable users to speak naturally to computers. Language, on the other hand, continues to be a significant barrier that makes it difficult to communicate with those from other countries. The process of translating text from one language into another using computer technology is known as machine translation (MT). Recently, neural machine translation (NMT) has been proposed and has improved in several language pairs. This paper's main goal is to design a system for neural machine translation between Myanmar and Korean. The proposed system contains two primary components. The first step is the development of new parallel corpus in Myanmar and Korean language. The attention-based neural machine translation system for the Myanmar–Korean language pair is introduced in the second. The proposed model's experiments are conducted using a word-based neural machine translation model. Recurrent neural network Encoder-Decoder architecture with an attention mechanism is used to evaluate the results of the translation between Myanmar and Korean using the BLEU score.*

## 1. Introduction

Language translation is becoming more widespread and diverse as globalization and information technology advance quickly. One of the key functions of a machine translation system in natural language processing (NLP) is to translate between languages. The field of NLP research has a particular focus in creating high-quality machine translation systems. Artificial intelligence is used in machine translation to translate text from one language to another without the need for a human translator. The application of machines to translate natural languages has gained popularity in recent years. Neural Machine Translation (NMT) makes an effort to create and train a single, sizable neural network that can read an input sentence and produce a correctly translated sentence. The translation quality of several language pairings has significantly improved using NMT and the attention-based encoder-decoder system. Both neural machine translation (NMT) and statistical machine translation (SMT) systems rely on sizable parallel data corpora for model training. Additionally, the size of the parallel data corpus has a significant impact on the effectiveness of neural machine translation systems.

On the other hand, Myanmar language is one of the low resource languages and Myanmar-Korean parallel corpus is rare. Therefore, in this system, a parallel corpus for Myanmar and Korean is first constructed, and then a neural machine translation system between the two languages is suggested. In order to create the Myanmar-Korean parallel corpus, Myanmar sentences from the UCSY-corpus [3], which comprises of the Myanmar-English language pair, are used. These Myanmar phrases are then manually translated into Korean. Additionally, this corpus includes parallel sentences from the spoken and written text books for school in both languages. More than 37K parallel sentences can be found in the Myanmar-Korean parallel corpus. The proposed system aims to implement the Neural Machine Translation System implementation between Myanmar and Korean.

In this paper, section 2 will describe related work of this paper. In the section 3, about Myanmar Language and Korean Language will be described. In section 4, neural machine translation and attention-NMT will be expressed. The proposed system design will be explained in

section 5. In section 6, the experimental setting of the system will be reported followed by the conclusion in section 7. Finally, reference papers of this system will be expressed.

## 2. Related Work

In this section, previous works on neural machine translation systems are introduced.

The authors first presented the Attention mechanism in [1], which creates an alignment model between the source and target characters or words. The parallel corpora used to evaluate the system's performance on the English-to-French translation tasks totaled 850M words. The RNN Encoder-Decoder (RNNencdec) and the suggested model are trained by the system using the equivalent settings (RNNsearch). The RNNencdec's encoder and decoder each have 1000 hidden units. One thousand hidden units each make up the forward and backward recurrent neural networks (RNN) that make up the encoder for the RNNsearch. Its decoder contains 1000 hidden units. Each model is trained by the authors using the minibatch stochastic gradient descent (SGD) technique with Adadelta (Zeiler, 2012). Each SGD update direction is computed using a minibatch of 80 sentences. The authors trained each model for approximately 5 days. On longer sentences, the presented models perform well, and the proposed RNNsearch greatly beats the traditional encoder-decoder model (RNNencdec).

An enhanced RNN neural network translation model is put out by the author in [2]. This article also compares the BLEU results of the Chinese-Korean corpus I, II, and III with the results of the conventional translation model. According to the results, the three corpora examined by this model have BLEU scores of about 45 points, compared to only 30 points for the traditional ones. The translation model in this paper's BLEU score has raised by roughly 15 points, which shows that the translation quality has greatly improved.

According to the author in [3], attention-based neural machine translation models are introduced based on word-to-word, character-to-word, and syllable-to-word levels and a parallel corpus for the Myanmar-English language pair is formed. The author used the default settings of the pytorch OpenNMT [9]. Moreover, to decrease the low resource problem, source side monolingual data are also used. The experimental results show that syllable (Myanmar) to word (English) level neural machine translation model obtains an improvement over the other systems.

Recurrent neural networks (RNN), transformer, and convolutional neural networks (CNN) were researched by the author in [4] and tested on a parallel text corpus in Myanmar and Rakhine. Additionally, word embeddings use the word byte pair encoding (Word-BPE) and syllable byte pair encoding (SyllableBPE) segmentation techniques. According to experimental findings, Syllable-BPE segmentation produces the best NMT and SMT performances for both types of translations.

## 3. Myanmar and Korean Languages

The primary language of the Republic of the Union of Myanmar is Myanmar language, which is spoken in that country. It's also referred to as Burmese language. Myanmar belongs to the Tibeto-Burman ethnic group. About 34.5 million people speak Myanmar as their first language, while another 10 million people speak it as a second language. Additionally, Myanmar is a language that is spoken in a few regions of the United States as well as in nearby nations like Bangladesh, Malaysia, and Thailand. Despite speaking their own native tongues, ethnic groups also speak Myanmar as a second language.

Myanmar script consists of (33) consonants: (က…to…အ), independent vowels (အ၊ ဣ၊ ဤ၊ ဥ၊ ဦ၊ ဧ၊ ဩ၊ ဪ), dependent consonant signs (also known as Medials) (ျ၊ ြ၊ ွ၊ ှ), dependent vowels signs (ါ၊ ာ၊ ိ၊ ီ၊ ု၊ ူ၊ ေ၊ ဲ၊ ံ၊ ့၊ း၊ ်), dependent various signs (also known as Pali Word) (ဉ၊ ၍၊ ၎၊ ၏၊ ္), punctuation (၊ and ။) and digits. (၀၊ ၁၊ ၂၊ ၃၊ ၄၊ ၅၊ ၆၊ ၇၊ ၈၊ ၉). The writing system used in Myanmar is left to right. The spoken style and the written style are the two types of language [3]. Additionally, the sentences' construction is subject-object-verb (SOV).

The spacing between words in the Myanmar language is not specified. Usually, there is no space between sentences. It is occasionally written with spaces between phrases. Sentences can be easily determined with sentence boundary maker "။" which is called ပုဒ်မ and pronounced as "Pou

ma". However, there is no set guideline on how to write in Myanmar. The sentence is made up of one or more words or phrases, following Myanmar sentence structure. There are one or more syllables in each word. And one or more characters make up a syllable. However, a word might sometimes just have consonants and no vowels.

Both North and South Koreans speak Korean as their official and native tongue. However, the two Koreas have acquired certain distinct vocabularies throughout the previous 74 years of political division. There are about 80 million South Korean speakers in the world [8]. The term "Korean" can refer to a language, a group of people, or a feature of a culture. The Democratic People's Republic of Korea (DPKR) is the name of North Korea, while ROK is the name of South Korea (Republic of Korea). The majority of Korean language students are learning the South Korean dialect, known as 한국어 (hangugeo). The other language spoken on the Korean Peninsula is referred to as North Korean 문화어 (munhwaeo). Subject-Object-Verb (SOV) is the standard Korean sentence structure [10]. The structure of Korean sentences is the same to the sentence structure of Myanmar language.

## 4. Neural Machine Translation (NMT)

Large-scale neural networks are trained end-to-end for language translation using neural machine translation (NMT), which is known as end-to-end training. A high quality NMT can determine the context of the translation and apply models to provide a more accurate translation. NMT is an algorithm that is used to translate words from one language to another. NMT is a recently advanced method of machine translation. The three most popular neural machine translation models are the Transformer, Attention, and Sequence-to-Sequence models.

### 4.1. Attention-based Neural Machine Translation

An attention-based NMT (Bahdanau et al., 2014) is an encoder-decoder network. Traditional encoder-decoder neural network models consist of two parts: encoder and decoder. Additionally, there are sequence-to-sequence encoder-decoder models based on recurrent neural networks (RNNs). An RNN decoder produces data for another sequence whereas an RNN encoder accepts input for one sequence.

The encoder in attention-based encoder-decoder architecture uses a bi-directional recurrent neural network (BiRNN) in particular because it performs better with longer sentences. Each source word's annotation is encoded by the encoder in order to obtain the word that comes before it and the word that comes after it. A BiRNN consists of two types of recurrent neural network: forward RNN and backward RNN. A forward RNN ( $\overrightarrow{f}$ ) takes the input sequence of words as in direction from left to right. And then it calculates a sequence as forward hidden states ( $\overrightarrow{h_1}, \ldots\ldots, \overrightarrow{h_{T_x}}$ ). The sequence is also taken by a backward recurrent neural network ( $\overleftarrow{f}$ ), which goes from right to left in the opposite way. It actually means from the beginning of the sentence to the end. And it results in a sequence of backward hidden states ($\overleftarrow{h_1}, \ldots\ldots\ldots, \overleftarrow{h_{T_x}}$). An annotation for each word $x_j$ (an input sequence x like [ $x_1$ ,....., $x_{T_x}$ ]) is obtained by joining the forward hidden state $\overrightarrow{h_j^T}$ and the backward hidden state $\overleftarrow{h_j^T}$ , i.e., $h_j$ = $\left[\overrightarrow{h_j^T}, \overleftarrow{h_j^T}\right]^T$.

It is suggested that the use of attention can align and translate. The only issue with machine translation is alignment. Alignment is the process of determining the connections between the words in the input and output, whereas translation is the act of applying this knowledge to select the appropriate output. This alignment is known as "attention" in the field of neural machine translation, and encoder-decoder models with attention are now frequently utilized. As a result, in addition to encoders and decoders, attention mechanisms are included in attention-based models. The decoder and alignment model compute the context vector using the sequence of annotations. The decoder accepts inputs that include word predictions, the prior concealed state, and a representation of the input context. The following output word prediction and new hidden decoder state are then generated:

$$p(y_i \mid y_1, \ldots\ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i),$$

A sequence of hidden states $si$ which are computed from the previous hidden state $si-1$, the embedding of the previous output word $y_{i-1}$, and

the input context $c_i$ as follows:

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

To compute the context state $ci$, the decoder gave the output as a sequence of word representations $hj = (\overrightarrow{h_j}, \overleftarrow{h_j})$.

The context state $c_i$ that the attention mechanism creates is informed by the prior hidden state of the decoder $s_{i-1}$ and all input word representations $h_j = (\overrightarrow{h_j}, \overleftarrow{h_j})$ and is calculated as follows:

$$\alpha_{ij} = \frac{\exp(a(s_{i-1}, h_j))}{\sum_k exp(a(s_{i-1}, h_k))}$$

Finally, the contribution of the input word representation $h_j$ to the context vector $c_i$ is valued using the normalized attention and finished.

$$c_i = \sum_j \alpha_{ij} h_j$$

## 5. Proposed System

There are two primary modules in the proposed system. Training Module is the first, while Testing Module is the second. First, a parallel corpus for Myanmar and Korean needs to be built for the training module. We must tokenize and clean the corpus data in preparation for data pre-processing. The NMT models are then trained using the PyTorch OpenNMT toolkit [9]. We must pre-process the input source sentences in the testing module. After that, a trained NMT model is used to translate sentences and produce the translated sentences.

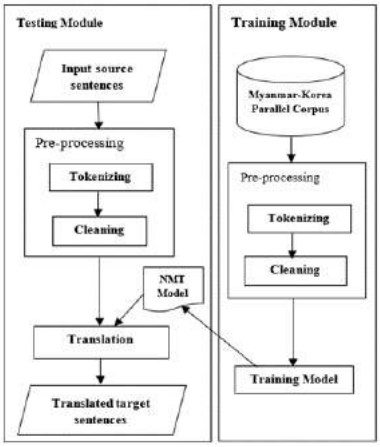



**Figure 1. Overall architecture of the proposed system**

## 6. The Experimental Setting

### 6.1. Dataset and Pre-processing

One of the low resource languages is Myanmar. There are not many parallel corpora between Myanmar and Korea at the present. For this system, we must therefore construct a parallel Myanmar-Korean corpus. In order to create a new Myanmar-Korean parallel corpus, Myanmar texts from the UCSY Myanmar-English Corpus [3] are collected and manually translated into Korean. About parallel sentences from local news, travel-related articles, school textbooks, and spoken textbooks in both languages are included in the corpus. More than 37K parallel sentences may be found in this parallel corpus. This corpus' parallel sentences are divided up at the word level. UCSY NLP Word Segmenter tool [13] was used for the word segmentation task, and Korean sentences were segmented manually. For the cleaning the corpus, Moses's clean scripts [12] was used.

We must follow Korean sentence grammar when performing manual segmentation for the Korean language. In Table 1, that is explained.

**Table 1. Segmentation Scheme for Korean language**

| Parts of speech | Korean meaning | Myanmar meaning |
|---|---|---|
| Nouns | 저, 나, 너, 당신, 그, 그녀, 우리, 저희 | ကျွန်တော်၊ ကျွန်မ၊ ငါ၊ မင်း၊ ခင်ဗျား၊ သူ၊ သူမ၊ ငါတို့၊ ကျွန်တော်တို့ |
| Verbs | 먹다 | စားသည် |
| Adjectives | 좋다 | ကောင်းသော |
| Adverbs | 빠르게 | မြန်မြန် |
| Feeling Verbs | 좋아하다 | ကြိုက်သည် |

For example, the current word-level segmented Myanmar sentence "အဲဒါက အဓိက ပြဿနာ ပါ ။" is equal to the translated Korean sentence "그것이 주요 문제입니다."

The Myanmar-Korean parallel corpus is randomly divided into three division files as shown in the following Table 2 in order to train the Myanmar-Korean NMT models:

**Table 2. Statistics of Korean-Myanmar parallel corpus**

| Files | No. of sentences |
|---|---|
| Training File | 33925 |
| Validation File | 3017 |
| Testing File | 700 |
| Total Sentences | 37642 |

### 6.2. Neural Machine Translation Model

Nowadays, neural machine translation systems have succeeded in almost all language pairs, and the field is developing quickly. Additionally, there are numerous toolkits accessible for the research, development, and application of neural machine translation systems. There are various NMT implementations in use right now. The Myanmar-Korean neural machine translation models were developed using PyTorch OpenNMT, which is available on GitHub [9]. 500 hidden units each on the encoder and decoder of a 2-layer long short-term memory are employed for the translation system. Drop-out was set to 0.3, and each direction's computations were done with 64 batches of data and a learning rate of 1.0. The Korean language has 25,282 words in its vocabulary, while the Myanmar language has 15,188 words.

### 6.3. Evaluation Details

Bilingual Evaluation Understudy (BLEU), one of the de facto recognized automatic evaluation metrics, is employed to examine the experiments of Myanmar-Korean Neural Machine Translation models. These experiments' BLEU scores are displayed in Table 3.

The experimental results show the performance of Korean-to-Myanmar model is better than the Myanmar-to-Korean model. In translation results, we found that some Foreign Names are unable to translate, and some Korean sentences are not using the subject word to translate. That is why the BLEU score of Korean-to-Myanmar NMT model is better than that of Myanmar-to-Korean NMT model.

**Table 3. Evaluation result of Korean-Myanmar NMT models**

| NMT Model | BLEU |
|---|---|
| **Myanmar-Korean** | **12.58** |
| **Korean-Myanmar** | **20.32** |

## 7. Conclusion

In this paper, we provided attention-based neural machine translation models for translations in both directions between Myanmar and Korean. Firstly, a corpus of over 37K parallel sentences between Myanmar and Korea has been constructed. Then word level segmentation is used to train models for Myanmar-Korean neural machine translation. Utilizing BLEU to evaluate the tests, the results are 20.32 for the Korean-Myanmar model and 12.58 for the Myanmar-Korean model. Since the size of the parallel corpus has a significant impact on the accuracy of Neural Machine Translation (NMT) systems, the current Myanmar-Korean parallel corpus will not be adequate for training translation models in the upcoming experiments. In order to improve translation performance, more data must be collected and other neural models must work harder.

## References

[1] Dzmitry Bahdanau, Jacobs University Bremen, Germany and KyungHyun Cho, Yoshua Bengio, Universite de Montreal, published as a conference paper at ICLR 2015 [Neural Machine Translation By Jointly Learning To Align And Translate]

[2] Yang Dong, Nanyang Institute of Technology, Nanyang, Henan 473000, China [RNN Neural Network Model for Chinese-Korean Translation Learning]

[3] Yi Mon Shwe Sin, Khin Mar Soe, International Journal on Natural Language Computing (IJNLC) Vol.8, No.2, April 2019 [Attention-based syllable level neural machine translation system for Myanmar to English Language Pair]

[4] Thazin Myint Oo, Ye Kyaw Thu and Khin Mar Soe, UCSY NLP Lab, Myanmar Language and Semantic Technology Research Team (LST), NECTEC, Thailand Language and Speech Science Research Lab, Waseda University, Japan [Neural Machine Translation Between Myanmar (Myanmar) and Rakhine (Arakanese)]

[5] Yongkeun Hwang, Yanghoon Kim (Department of Electrical and Computer Engineering, Seoul National University) and Kyomin Jung (Automation and Systems Research Institute, Seoul National University) [Context-Aware Neural Machine Translation for Korean Honorific Expressions]

[6] Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita, National Institute of Information and Communications Technology, ACM Trans. Asian Low-Resource Lang. Inf. Process., Vol. 15, No. 4, Article 22 [Word Segmentation for Myanmar (Myanmar)]

[7]https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/

[8]https://en.wikipedia.org/wiki/Korean_language

[9]Pytorch-OpenNMT:http://github.com/OpenNMT/OpenNMT-py

[10]https://www.90daykorean.com/how-to-lear n-the-korean-alphabet/

[11] Attention Model Intuition. https://www.youtube.com/watch?v=SysgYptB198

[12] Moses Toolkit: http://www2.statmt.org/moses/

[13] Win Pa Pa, Nilar. Thein, "Myanmar Word Segmentation using Hybrid Approach", Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, pp-166-170.

# Offensive Speech Detection Using Machine Learning Model

Ei Phyoe Hein[#1], Ei Ei Thu[#2]
*University of Computer Studies (Taunggyi), University of Computer Studies (Pinlon)*
*eiphyoehein@ucstgi.edu.mm, eieithu@ucspinlon.edu.mm*

## Abstract

*Nowadays, Social Media platforms becomes the dissemination of offensive speech. As the volume of online content continues to grow, the offensive speech spread rapidly. Therefore, many countries have developed law to prevent online offensive speech. Offensive speech identification is to detect the inflict tweets, abusive comments; execrate words on social media platforms. In order to detect hurtful tweets, the efficient offensive speech detection technique is used. This system provides a simple machine learning model using logistic regression and support vector machine method. The system intends to solve the problem of offensive speech detection. In order to detect offensive speech, this system used Offensive Speech Dataset. Tweets data are collected from Twitter social media web site. The experiments are conducted with 80% for training data and 20% for the testing. Results show that the system logistic regression and support vector machine approach gains high accuracy in offensive speech detection process.*

## 1. Introduction

Through various social media platforms like Facebook, Twitter, Viber, Telegram etc., communication through the internet has grown more faster than any other medium on the world. This paper focused on classification from Twitter social media website. Speech that is offensive attacks or shows hatred toward an individual or group on the basis of their race, religion, sex, or sexual orientation. Typically, disparaging remarks about a person or a group based on their race, color, national origin, sex, gender, handicap, religion, or sexual orientation are considered to be offensive speech. Correctly detecting offensive word is still difficult task. It is challenging problem to identifying the word whether it is abusive or not. Machine learning plays vital role in identification and detection domain. In machine learning, there are many methods to detect offensive speech. Among them, logistic regression and support vector machine obtains the higher accuracy in detection than other methods. The detection method described in this paper is based on the logistic regression and support vector machine model. Efficient Logistic Regression and Support Vector Machine model is proposed to predict the offensive speech. This system intends to predict tweets from a Twitter dataset [7] whether it is offensive or not. The main contribution of this system is the dataset that was re-created from the Twitter dataset that is collected from the Kaggle website. The system intended for the offensive speech therefore remove hate speech column and hate speech 1430 tweets. The system uses 23353 tweets and 6 columns.

## 2. Related Work

In recent works, many of the research have been conducted to detect offensive speech. Some of the related works are presented in this section.

The researchers [1] suggested a method for detecting hate speech in Amharic language. They collect of 6120 instances of Amharic posts from Facebook, and they labeled the speech as “hate” and “not hate” using word2vec and term frequency-inverse document frequency (TF-IDF) feature extraction. In their system, they used Naïve Bayes (NB) and random forest (RF) classifier to detect the features of “hate” and “not hate” speech. According to their result, NB classifier is outperformed than RF classifier.

The researchers [2] detected hate speech in social media using machine learning. The authors used character-based 4grams feature engineering approaches to produce numerical features for their study. The Support Vector Machine classifier was fed the obtained numerical characteristics by the authors.

The researchers [3] created a new dataset for the classification of tweets as hate speech using the Twitter API, offensive language, or neither. They

included a set of 85.4 million Twitter samples from roughly 33 000 Twitter users in their dataset. They then created a collection of 24k labeled Twitter samples. Bigram, Unigram, and Trigram features were employed for the classification job and were weighted according to their TF-IDF. Additionally, there were binary and count indications for URLs, hashtags, mentions, and retweets. Numerous classifiers, including logistic regression, Naive Bayes, decision trees, random forests, and linear SVMs, were put to the test (Support Vector Machine). Through their tests, it was discovered that Linear SVM and Logistic Regression tended to produce better outcomes.

The researchers [4] classified messages as offensive or not offensive using the Cable News Network (CNN), Support Vector Machine (SVM) and bidirectional Long Short-term Memory (BiLSTM) machine learning model. According to their experiments, the BiLSTM better than the other model.

The researchers [5] detected hate speech and cyber bullying using three machine learning models. These models are XGBoost, Logistc Regression and Decision Tree. According to their result, the Logistic Regression and Decision Tree models outperformed than other machine learning model.

The researchers [6] compared Logistic Regression, Naïve Bayes, Random Forest, Support Vector Machine, K-nearest Neighbors (KNN), AdaBoost and Multilayers perceptron (MLP) machine learning models. According to theirs experiments, Support Vector Machine outperformed other machine learning with accuracy 79%.

The researchers [7] classified hate speech on twitter to create the numerical vectors for their research, they used character N-grams feature engineering approaches. The authors gave the LR classifier the generated numeric vector and received an overall F-score of 73%.

## 2. Machine Learning

Machine learning's main goals are to learn useful models of the input data and to convert the input data into meaningful outputs. Through supervised and unsupervised learning, the machine "learns" and applies its algorithms. The process of teaching a computer to convert input data into a predefined actual output is known as supervised learning. Unsupervised learning is the process of identifying new patterns in data without any prior knowledge and training [9]. Many actual industrial problems have been successfully solved using machine learning techniques. One of the commonly used techniques is support vector machine and logistic regression. Support Vector Machine and Logistic Regression techniques belongs to the family of supervised learning algorithms.

### 2.1. Supervised Learning

Supervised machine learning is another name for supervised learning. A subset of artificial intelligence (AI) and machine learning is supervised learning. Its use of label datasets to train algorithms that can accurately classify data or predict outcomes defines it. When using data mining, supervised learning may be divided into two categories of issues: classification and regression. In supervised machine learning processes, several algorithms and computation techniques are used. There are neural network, naïve bayes, logistic regression and support vector machine etc.

### 2.2. Logistic Regression

The method of modeling the likelihood of a discrete result given an input variable is known as logistic regression [9]. A binary result, such as true or false, yes or no, and so on is modeled by the most popular logistic regression models.

The logistic function, also referred to as the sigmoid function, is where logistic regression first began. This function is used by logistic regression to forecast the likelihood that a given point will belong to a class. Any real value can be converted by the function into a value between 0 and 1.

In Logistic Regression, the input values ($x$) are combined linearly using coefficient values (β), called weights to predict an output value ($y$). This system uses probability Equation (1).

Formula of Logistic Regression:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_i \qquad (1)$$

where, for i=n observations:

$y_i$ = dependent variable

$x_i$ = independent variables

$\beta_0$ = y-intercept

$\beta_n$ = coefficients for each independent variable

### 2.3. Support Vector Machine

One of the supervised machine learning techniques used for various classification issues is the Support Vector Machine (SVM) algorithm. The system searches for a hyperplane that best divides the two classes in this supervised machine learning issue. The main of this algorithm is to come up with the hyper-plane such that the marginal width is maximum. Marginal width is the distance between the 2 marginal lines where each marginal line is at closest data point of each of the classes present in our dataset. These closest points belonging to each class from the hyper plane are called support vectors. SVM works very well in the cases of high dimensional data and also in the cases where number of dimensions are greater than the number of instances or data points. SVM can work with unstructured and semi-structure data like text and image.

### 2.4. TF-IDF

In order to measure the frequency of word in document, TF-IDF method is used. TF means term frequency and it can obtain by dividing count of one word by told number of words. IDF means inverse document frequency and it can obtain by taking log value in dividing number of specific terms including documents and total number of documents. In Equation (2) describe the calculation of tfi-df.

$$tfidf(d,t) = tf(t) * idf(d,t) \quad (2)$$

In this paper, TF-IDF is used to transform tweeted words into numeric word vector.

## 3. Methodology

### 3.1. System Design and Architecture

The aim of this system is to detect offensive speeches. In order to detect offensive speech, the proposed system build efficient machine learning model using logistic regression method. There are three main parts in the proposed system architecture. Figure 1 represents the overview architecture of the system.

1. Data Pre-processing
2. Feature Selection
3. Model Building

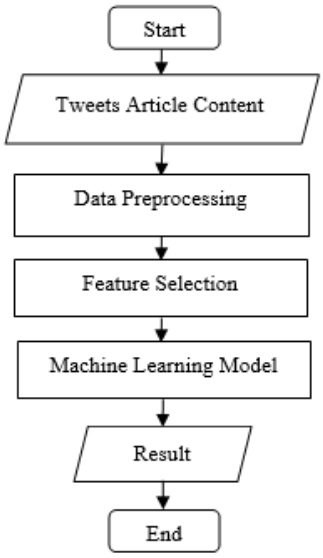



**Figure 1. System Architecture**

### 3.2. Data Pre-Processing and Dataset

In order to pre-process the tweeted data, the system tokenize, stem and remove the username, punctuations, special characters, emoji, hash symbol and hashtag using NLTK, PANDAS, NUMPY, Scikit-learn, Porter Stemmer, Word-net Lemmatizer, Stopword remover under the NLTK library. After removing unnecessary characters, the system changes all texts into lower case.

**Natural Language Library (NLTK) is mainly used for interaction between computer and human language.** In this system, NLTK use for tokenization, porter stemmer, lemmatizations, lower case conversion, stop words removal, remove twitter handle (e.g.; @user), remove url from tweet (e.g.; http://www.twitter.com), remove punctuations, numbers (e.g.; 1, 2, 3…) and remove special characters. In this system, PANDAS is used for handling data that can load dataset files. NUMPY is a library that offers effective multi-dimensional array objects and a number of operations to use with it. There are six features in offensive dataset [10]. These are id, count, offensive, not offensive, label and tweet. Column ‘id’ contains 0 to 23352, column ‘count’ contains the number of CrowdFlower (CF) users who coded each tweet, column ‘offensive’ contains the number of CF users who judged the tweet to be offensive, column ‘not offensive’ contains the

number of CF users who judged the tweet to be not offensive, column and column 'label' contains class label for majority of CF users. 1 – offensive speech and 0 – not offensive. The number of offensive speeches from the data is 19190 and4163 as not offensive. Column 'tweet' contains the user tweet content. The experiments are conducted with 80% for training data and 20% for testing data. In Figure 2 depicts the dataset example.

### 3.3. Feature Selection

When using the Unigram technique, each word is parsed manually before being entered into the term frequency computation. The proposed use word n-gram of minimum order (1,2).

The two significant measurements that show the specificity and relevance of terms with the information carried by the documents are the term frequency (TF) and inverse document term frequency (IDF). The system has used TF-IDF weights for the n-gram features extracted from the tweets in the datasets.

| id | count | offensive | Not offensive | label | tweet |
|---|---|---|---|---|---|
| 0 | 3 | 0 | 3 | 0 | !!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your h|use. & as a man you should always take the trash out... |
| 1 | 3 | 3 | 0 | 1 | !!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!! |
| 2 | 3 | 3 | 0 | 1 | !!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit |
| 3 | 3 | 2 | 1 | 1 | !!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny |

**Figure 2. Dataset Example**

### 3.4. Model Building

This method, which is the output of the feature selection process, uses data train in term weighting form. In this approach, text was classified using machine learning, and a data mining classification technique was utilized as the method to create a model that predicted whether or not the text would be considered offensive. The process output is a classification model based on the Support Vector Machine and Logistic Regression algorithms.

This classifier's goal is to locate a hyper-plane in a space with n dimensions, where n is the total number of attributes used to classify the data points. Finding boundary points that can divide objects into different groups is the main objective of the training process. Support vectors are used to improve the classifier's margin. These important points affect the hyperplane's direction because they are nearer to it. Each message must be converted into a numerical feature vector as the first step in setting up the data for the Support Vector Machine and Logistic Regression to execute. One method that is frequently used in the literature is TF-IDF (Term Frequency - Inverse Document Frequency).

### 3.5. Evaluation Matrix

The number of instances that are correctly classified is known as a true positive (TP) and false positive (FP) in classification. The number of classifications that the classifier successfully identified as the correct response is a true positive. The negative, or a true negative, is the number of classifications that the classifier properly predicted wouldn't be the solution. The improperly classified occurrences are true negatives (TN) and false negatives (FN). When an incorrect instance is labeled as correct, a false positive result. When a correct occurrence is labeled as incorrect, a false negative result.

## 4. Result and Discussion

In order to measure the accuracy of the system, an accuracy score metric is used. This metric compares the predicted result with actual results. After testing the model, this system achieves 96% accuracy. The confusion matrix for the system Logistic Regression and SVM models is show in Figure 3 and 4.

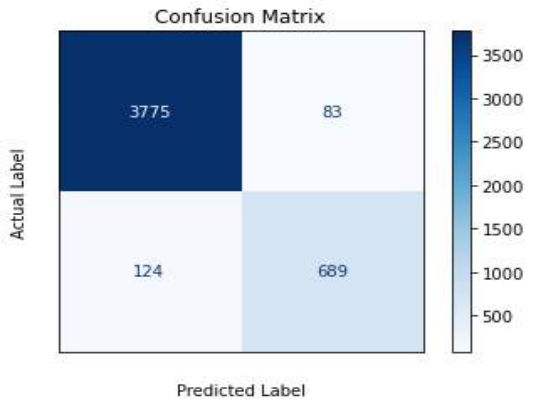



**Figure 3. Logistic regression confusion matrix on test dataset**

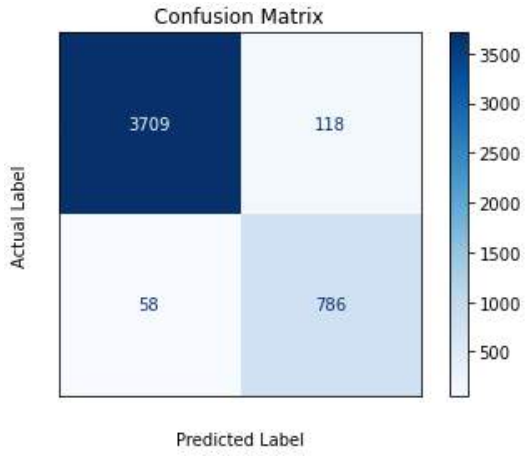



**Figure 4. Support Vector Machine confusion matrix on test dataset**

### 4.1. Confusion Matrix

Confusion Matrix: It is the representation of the information about the actual and predicted classifications by the classification algorithm. It gives the insight into the accuracy of each of the classes that are present in the target feature meaning that the system get the number of test instances for each class that are correctly as well incorrectly predicted by the system model in Table 1.

**Table 1. Confusion Matrix**

| | | Predicted Label | |
|---|---|---|---|
| | | Offensive | Not Offensive |
| Actual Label | Offensive | TP | FN |
| | Not Offensive | FP | TN |

### 4.2. Comparative Study

The comparative study results obtained using the Logistic Regression and Support Vector Machine classifier are displayed below Table 2. The confusion matrix that generated was offered by this system. The model's F1-score and accuracy score are as follows:

**Table 2. Result of LR and SVM Model**

| | Accuracy | Precision | Recall | F1- Score |
|---|---|---|---|---|
| LR | 0.95 | 0.98 | 0.97 | 0.97 |
| SVM | 0.96 | 0.97 | 0.98 | 0.97 |

Logistic Regression can only be handled in linear solution while Support Vector Machine can be handled in both linear and nonlinear solution. Therefore, this system achieved the Support Vector Machine is better than the Logistic Regression with accuracy 96% and 95% respectively. In [8] paper, the authors achieved Support Vector Machine and Logistic Regression with accuracy 71% and 70% respectively. According to this system result, Support Vector Machine model outperform than Logistic Regression model.

## 5. Conclusion

A simplified machine learning model based on logistic regression and support vector machines is presented. The problem of offensive speech detection is solved by using the system model. The results of the experiments demonstrate that the system approach is effective in identifying offensive speech on Twitter. The offensive speech detection for other social media platform (Facebook, Instagram, Twitter, etc.) will conduct in future. Finally, the system led to the fact that the logistic regression and support vector machine model by TF-IDF the previous performance achieving the accuracy of 0.95 and 0.96 respectively on this dataset. According to the result, SVM model is performed consistently better than LR model.

## Acknowledgements

I would like to express my graditude to my Pro Rector, Dr. Khin Sandar Aung for giving me a golden opportunity to do this thesis.

I would like to express very special thanks to Dr. Darli Myint Aung, Professor and Dean of Master Course, University of Computer Studies,

Taunggyi for her valuable advice and administrative support for completion of the thesis.

And I'd like to express my deep gratitude to Lecturer Dr. Ei Ei Thu, my research supervisor, for her patient guidance, enthusiastic encouragement, and useful critiques of this thesis work.

Finally, may I extend my grateful thanks to my parents and my siblings for their encouragement, understanding and support throughout the period of doing this thesis.

## References

[1] Z. Mossie, J.H. Wang, "Social network hate speech detection for Amharic language*",* In Proceedings of the 6th International Conference on Computer Science and Information Technology, Copenhagen, Denmark, 28–29 April 2018; pp. 41–55.

[2] Malmasi S. and M Zampieri, "Detecting hate speech in social media", 2017.

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the 11th International AAAI Conference on Weblogs and Social Media, ser. ICWSM '17, 2017

[4] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media", In Proceedings Conference of the North American Chapter of the Association of Computational Linguistics: *Human Language Technologies, vol. 1*, June, 2019, pp. 1415-1420.

[5] O. ORIOLA and E. KOTZÉ. "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets", 2020.

[6] S. Abro, S. Shaikh2, Z. Ali4, and Z. H. Khand, "Automatic Hate Speech Detection using Machine Learning: A Comparative Study", *International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8*, 2020

[7] Waseem, Z. and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. in Proceedings of the NAACL student research workshop. 2016.

[8] V. Pathak, M. Joshi, P. Joshi, M. Mundada, T. Joshi, "Using machine learning for detection of hate speech and offensive code-mixed social media text", CEUR Workshop Proceedings 2826 351–361, 2020.

[9] https:// chatbotsmagazine.com/machine-learning-neural-networks-and -algorithms-5c0711eb8f9a

[10] https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

# Intent Classification of Users' Comments in Myanmar Language on Social Media Shopping Pages

Ei Myat Myat Noe, Hsu Myat Mo
*University of Computer Studies, Yangon*
*eimyatmyatnoe@ucsy.edu.mm, hsumyatmo@ucsy.edu.mm*

## Abstract

*According to the social media usage statistics, there are about four billion total social media users across all platforms. Moreover, people spend more time on social media than before and even perform daily activities on them. Because of the convenience of social media services, activities such as online shopping can be easily done on social media. Social media commerce has been one of the popular ecommerce trends in recent years. AI applications in customer services such as chatbots and personalization help business to understand their customers better and improve customer experience. Intent classification is one of the techniques used in these applications. This paper focuses on the classification of users' intentions based on the user comments posted in Myanmar Language on social media shopping pages. Convolutional Neural Network (CNN) is applied to classify the users' comments to one of the predefined intent categories. According to the experimental result, intent classification model with name normalization plus word segmentation preprocessing can give the F-score value of 0.86.*

## 1. Introduction

Social media where people communicate, share thought and opinions about a particular topic, exchange information and express feeling, has made people's lives easier and has become an integral part of one's life. Social media enables communication for not only one's personal life but also for business life. People started adopting digital marketing instead of marketing offline and social media plays a vital role in promoting online business.

Most of the social media users usually post comments as their opinion, feedbacks of the services or products, and also as their inquiry about particular information. Analyzing the users' generated content such as comments on social media post has become an essential task to know the intentions behind those comments which can be very beneficial for the business.

Intent classification (sometimes called as intent recognition) is a technique used in Natural Language Processing (NLP) and the fundamental concept of Natural Language Understanding (NLU). Intent classification is a kind of text classification [4]. Intent classification takes the written or spoken text as input for processing and uses machine learning (ML) and NLP techniques to analyze the intentions behind the input written or spoken texts and assigns or categorizes them into their predefined intent automatically.

AI applications in customer services such as chatbots and personalization help businesses to understand their customers better and improve customer experience and can also provide better services to their customers. Intent classification is one of the techniques used in these applications. It is useful to understand the intensions behind customer queries, automate process, and gain valuable insights. By understanding the users' intention, business can give a more accurate response to their users and customers.

Intent classifier can be modeled by applying traditional machine learning classification models such as k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes and so on. However, the diverse nature of contents on social media can make difficult for those traditional machine learning methods to classify intent classes. In recent years, deep learning models have proved that these models give impressive results on image processing and computer vision, speech recognition, NLP tasks such as machine translation, text classification, and many other tasks.

The filters/ kernel in CNNs can help identify relevant patterns in text data depending on kernel size. Since CNNs are translation invariant, they can detect the relevant patterns irrespective of their position in the sentence. Each filter/ kernel detects a specific feature, such as if the sentence

contains some key phrases terms anywhere in the sentence. Most test classification tasks such as intent classification are determined by the presence or absence of some key phrases present anywhere in the sentence. Therefore, text classification tasks like intent classification task can be effectively modeled by CNNs which are good at extracting local and position-invariant features from data. Most comments on social media are written in free ordered and no regular standard grammar order. For this reason, CNN has been chosen to be applied for this intent classification task of users' comments in Myanmar text on social media shopping pages.

Myanmar language is free order language and the scripts are written from left to right without placing a regular space between words or phrases. Word segmentation step is one of the most important preprocessing tasks for Myanmar NLP research. Moreover, name words are also needed to be tokenized correctly because it can also affect the model accuracy. In this paper, before intent classification step, word segmentation and name normalization steps are performed as preprocessing tasks. Without name normalization step but only with word segmentation processing step, the model accuracy is not as good as the model with word segmentation and name normalization preprocessing steps. Word level tokens are considered as input tokens to CNN model training for this intent classification task. The results show that CNN can give the promising result even though performed on the unbalanced and low amount of data.

In the following section, some of the research that are related to this intent classification, are described. In Section 3, the explanation of CNN is presented followed by the detail explanation of intent classification model training on users' comments in Myanmar scripts in Section 4. Experiment results are shown in Section 5 and then finally the paper is concluded by giving the performance analysis of the model.

## 2. Related Work

Intent classification uses machine learning and NLP to automatically associate text or expressions with a particular intent. There are a lot of machine learning algorithms available for intent classification and deep-learning based neural network architectures are also used to perform intent classification of text.

The authors of [1] considered the task of annotating travel-related reviews with travel intents that best represent the reviewer's reasons for visiting the place of interest. They classified the reviews into eight travel intent, business, eating out, education, health, holidays, religion, shopping and socializing. They applied Naïve Bayes as a baseline method and compared with other classification models such as DNN, SVM and RF.

The authors of [2] addressed the problem of multiclass classification of intent with a use-case of social data generated during crisis events. The crisis data was classified into three different intents, seeking, offering, and none (neither seeking nor offering). The hybrid approach that combines knowledge-guided patterns with syntactic features based on bag of tokens.

The authors of [3] focused on the intent classification of users' generated comments on social media posted in Myanmar text. Comments posted on telecommunication service were classified into five intent categories, application, auto-subscription, bill, customer service and internet. The combination of pretrained model and deep-learning model was applied in their experiments. Continuous Bag of words (CBOW) model was used to train the task-specific Word2Vec model which is input for the CNN to classify the users' comments into their corresponding intents. Their proposed model gains F-score value of 0.94.

## 3. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a type of artificial neural network [5]. Although CNNs are mainly used in image processing and computer vision problems, lately they have been applied in NLP tasks with interesting outcomes.

In this proposed CNN model, there are four layers: embedding layer, convolutional layer, pooling layer, and fully connected output layer. The incoming input sentences are needed to convert into dense vectors before further processing. From the convolutional layer, various features form the input can be extracted. By sliding the filter and the parts of the input with respect to the size of the filter, the output is

termed as the feature map which gives the information about the sentence. The pooling layer reduces the size of the feature map which means it only retains the important features of the input. Pooling can be performed either average pooling or max pooling. In the fully connected layer, an activation function is used to calculate the probability of each input to every output vector. Figure 1 shows the CNN model architecture for intent classification of is presented.

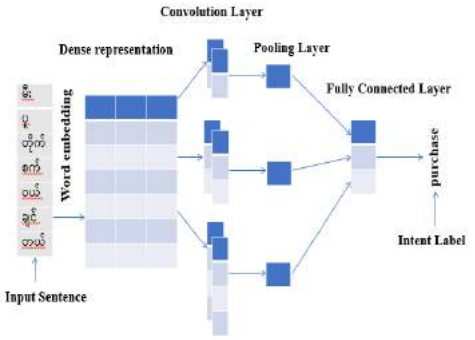



**Figure 1. CNN Architecture for this Intent Classification Model**

## 4. Intent Classification Model

In this section, the overall work flow of the proposed CNN intent classification modeling will be described.

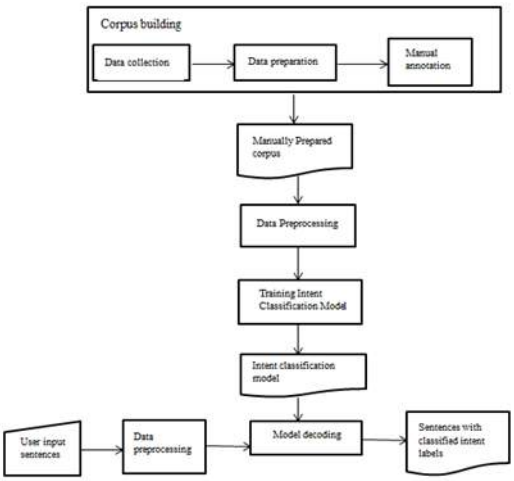



**Figure 2. Work Flow of the system**

Figure 2 show the work flow of the proposed intent classification modeling. Firstly, Myanmar language is a low resource language which means there no efficient linguistic resources such as data that are needed to process with machine learning algorithms.

As a very step for this intent classification model, a corpus was manually constructed. Data preprocessing steps (word segmentation and name normalization) are performed on this manually prepared corpus before training process. After that intent classification model training is conducted with CNN neural architecture.

In decoding stage, the obtained CNN based intent classification model is used to classify the user input comments into their corresponding intent categories.

### 4.1. Corpus Building

There are many vendors and shopping malls in Myanmar and most of them have their own official social media pages (Facebook pages). Users' comments written in Myanmar language from these pages are collected as raw data to build the corpus.

It is important to prepare the data before any other tasks. Comments on social media are written in different fonts. Firstly, all the collected data are converted from non-Unicode fonts such as Zawgyi font to Unicode font in order to get the unique Unicode encoding, and are corrected spelling errors and mistyped errors in text. Moreover, noise data are cleaned; such as by eliminating emojis from the sentences, and all the foreign words are transliterated.

After being performed these data preparation tasks, all the prepared data are manually annotated with the predefined seven types of intent categories. In this work, seven different types of intent that are related to the shopping activities are defined to classify the users' comments.

The intent "purchase" is used to annotate the comments that are written with the intention of buying a product or something from the vendors. The intent "payment" is for comments that are related to the action or process of paying for the products. The intent "delivery" is used to annotate the users' comments that are related to the delivery services. The intent "job" is used when the comments appear like want to apply for job or inquiry for job position. The intent "opinion" is applied when the users' comments seek like expressing their opinion about the products or services. When users' comments are

questions asking, they want to know something about the services or products, these texts are annotated with the intent "general_info". The next intent "contact_info" is considered for the annotation of users' comments that are posted with the intention of inquiring the contact information such as address or phone no of the vendors or services and where they can buy the products. Table 1 shows the examples of defined intent categories and their usage in the annotation process.

There are over 1K sentences in this manually annotated intent corpus. Table 2 shows the data statistic and distribution of each intent category in this manually annotated corpus. It contains the intent category "general_info" the most; over 27% over the whole dataset; and the intent category "job" is the least; this type of intent is about 5% in the dataset.

**Table 1. Description of defined intent classes and sample data**

| No | Defined Intents | Example Sentences |
|---|---|---|
| 1 | purchase | မီးပူတိုက်စက်ရှိလားရှင့်။ |
| 2 | payment | ဆိုင်မှာလာဝယ်ရင် ကေပေး နဲ့ရှင်းလို့ရလားရှင့်။ |
| 3 | job | ဝန်ထမ်းခေါ်လားဗျ။ |
| 4 | contact_info | ဘိုကလေးဈေး ဂမုန်းပွင့် ဖုန်းနံပါတ်သိချင်ပါတယ် |
| 5 | delivery | နယ်ကမို့ပါ။ လူကိုယ်တိုင် မလာဝယ်နိုင်တဲ့အတွက် စာတိုက်ကို ပို့ပေးနိုင်ပါသလားရှင့်။ |
| 6 | general_info | ကားပါကင်က တစ်နာရီကို တစ်ထောင်လားရှင့်။ |
| 7 | opinion | ကောင်းလိုက်တဲ့ ဝန်ဆောင်မှုပါ။ |

**Table 2. Data statistic and data distribution of dataset**

| Data | Size | Data distribution % |
|---|---|---|
| Total Data | 10335 | |
| purchase | 2165 | 20.948% |
| payment | 586 | 5.670% |
| job | 158 | 1.529% |
| contact_info | 1196 | 11.572% |
| delivery | 1837 | 17.775% |
| general_info | 2858 | 27.654% |
| opinion | 1535 | 14.852% |

## 4.2. Preprocessing

Myanmar scripts are written in free order and are also written continuously in sequence with or without breaking between words which means that there is no standard placement of white space between words or phrases in sentences. Therefore, useful information about words that are very important to text processing with machine cannot be got. Word segmentation is a very important preprocess for Myanmar NLP tasks. In this work, word level is considered as basic input tokens to the CNN input layer. For this word segmentation process, users' comments are segmented into words by using word-segmentor[1] for Myanmar language from UCSY-NLP research lab.

Moreover, many name words appear in data and it is necessary to correctly recognize the name word in text. It can also affect the classification accuracy. For this reason, named normalization process is carried out to correctly segment name words and recognize names in text. For this name normalization process, named entity recognizer for Myanmar language from UCSY-NLP research lab [6] was firstly applied to recognize names in comments. After that, named entities are normalized with respective name words.

During intent classification modeling, two different types of experiments are carried out: the first one is carried out with data that are only preprocessed with word segmentation process; and another is carried out with data that are obtained from both word segmentation and named normalization processes. The experimental results show that as data preprocessing step, named normalization is as important as word segmentation for the text processing.

## 4.3. Training Set Up

For the CNN neural training for this intent classifier modeling, TensorFlow[2], Keras[3] deep learning python framework is used. All the experiments are conducted on Google Colab service provided by Google. Data is partitioned into 80% for training and 20% for testing. (See in Table 3).

Training data are passed to the embedding layer of CNN for dense vector representation of each input word. Embedding dimension size is set as 60. Input maximum length is 10.

[1] http://www.nlpresearch-ucsy.edu.mm/wordsegmentation.html

[2] https://www.tensorflow.org/

[3] https://keras.io/

One-dimensional filters of size 3 convolve over one-dimensional input. 64 filters are applied and for each filter, values of input word vectors are pairwise multiplied with associate weights in the filter and then results are being summed up and passed into ReLU activation function to produce output feature maps.

As for the pooling layer, pooling is performed with one-dimensional max-pooling to capture the important features. Outputs from pooling layer are flattened and passed into fully connected output layer to predict the probability of each intent class. In this layer, softmax is applied as activation function.

**Table 3. Data partition for training and testing**

| Data | Size | % |
|---|---|---|
| Total Data | 10335 | - |
| Training data | 8267 | 80% |
| Testing data | 2067 | 20% |

## 5. Experimental Result

Performance of the trained models is measured with the F1-score. Firstly, intent classification model is trained and tested with data that are preprocessed by only word segmentation. This model gives the F1-score of 0.70%. The proposed model, which is trained with the named normalized and word segmented data, outperforms; and F1-score value of 0.86% is obtained. (See in Table. 4).

Although CNN is effective when the data is well prepared and trained on very large amount of data, from this experiment, it can be seen that it still works properly while training on the low amount of data if the data is well prepared. Therefore, more data is needed to get the full effectiveness of CNN. However, the experimental result is promising.

**Table 4. Comparison of F1-score for each model**

| Model | F1-score |
|---|---|
| With word segmentation | 0.70 |
| With word segmentation and name normalization | 0.86 |

## 6. Conclusions

In this paper, a deep-learning based Convolution Neural Network (CNN) is applied to perform intent classification of users' comments in Myanmar language on social media shopping pages. In data many name words appear, thus name normalization is also applied as one of the preprocessing steps before training and it is more efficient than the model trained with only word segmentation process. Therefore, this paper describes the comparison of two intent classification models.

In the future, more data will be collected for existing intent categories and for new intent categories. There is also a plan to experiment with the combination of other pretrained model and deep learning model to improve the performance of the deep learning model.

## References

[1] Z. M. Kim, Y. S. Jeong, J. Hyeon, H. Oh, H. J. Choi, "Classifying Travel-related Intents in Textual Data," Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 3, Issue 1 (2016) ISSN 2349-1469 EISSN 2349-1477.

[2] H. Purohit, G. Dong, V. Shalin, K. Thirunarayan, A. Sheth,"Intent Classification of Short-Text on Social Media," 2015 IEEE International Conference on Smart City/SocialCom/ SustainCom (SmartCity).

[3] T. N. Tun, K.M. Soe,"Intent Classification on Myanmar Social Media Data in Telecommunication Domain Using Convolutional Neural Network and Word2Vec," 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA).

[4] Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

[5] K. D.a Chaudhuri, "Intent Classification with Convolutional Neural Networks".

[6] H. M. Mo, and K.M. Soe, "Syllable-Based Neural Named Entity Recognition for Myanmar Language", International Journal on Natural Language Computing (IJNLC), Vol.8, No.1, February 2019.

[7] H. M. Mo, and K.M. Soe, "Myanmar named entity corpus and its use in syllable-based neural named entity recognition ", International Journal of Electrical and Computer Engineering (IJECE), Vol.10, N0.2 pp. 1544~1551, April, 2020.

# Statistical Machine Translation between Myanmar and Lisu Languages

Zaw Mee[1], Win Pa Pa[2]
[1]*University of Computer Studies, Myitkyina,* [2]*University of Computer Studies, Yangon*
*zawmee@ucsy.edu.mm, winpapa@ucsy.edu.mm*

## Abstract

*Natural language processing, as an essential component of artificial intelligence technology, is rooted in a variety of disciplines, including linguistics, computer science, and mathematics. Natural language processing's rapid advancements provide strong support for machine translation research. The process of translating text from one language into another using computer technology is known as machine translation (MT). Recently, statistical machine translation (SMT) has been proposed and has improved in several language pairs. The primary objective of this study is to develop a system for statistical machine translation between Lisu and Myanmar. There are two key parts to the system overview. Making a new parallel corpus in Lisu and Myanmar is the first stage. The second section introduces the phrase-based statistical machine translation system for the Myanmar-Lisu language pair. Experiments with the proposed model are carried out using a word-based statistical machine translation model. Using the BLEU score, this system is used to evaluate the results of the translation between Myanmar to Lisu and Lisu to Myanmar Languages.*

## 1. Introduction

As globalization and information technology advance, language translation becomes more widespread and diverse. Natural language is a fundamental aspect of human behavior and an essential component in our daily lives. It is a tool for communicating with people all over the world. Natural language processing (NLP) refers to computers' ability to generate and interpret natural language. The use of computers to translate text and speech from one natural language to another is known as machine translation. Machine Translation is the use of computers to automate some or all of the tasks associated with translating between human languages. Rule-based Machine Translation, Example-based Machine Translation, and Neural Machine Translation are the three major machine translation techniques. This work presents the statistical machine translation system (SMT).

The proposed system is intended to implement the Statistical Machine Translation System between Myanmar and Lisu. Section 2 of this paper explains related work for this paper. Section 3 discusses the nature of Myanmar and Lisu languages. Section 4 discusses statistical machine translation and phrase-based translation (SMT). The suggested system overview in Section 5 is described. Section 6 presents the system's implementation, followed by Section 7's conclusion. Finally, at the end of this system, there are reference papers.

## 2. Related Work

This section describes research on statistical machine translation.

Bidirectional statistical machine translation, which creates an alignment model between the source and target characters or words, was introduced by the authors in [1]. In this paper, phrase alignment pairs were tested for statistical machine translation. English-Afaan Oromo to English-Afaan Oromo. This study created English and Afaan Oromo language models using monolingual corpora of 19300 and 12200 sentences, respectively. The experimental results show that BLEU scores of 18% and 35% were obtained from English-Afaan Oromo and Afaan Oromo-English, respectively.

In [2], the authors presented statistical machine translation for the second time. This paper contributes to the quality assessment of Statistical Machine Translation between Myanmar and Kayah languages. This system also created a Myanmar-Kayah parallel corpus of 6590 sentences based on the ASEAN MT corpus

for Myanmar. According to the results, the PBSMT, HPBSMT, and OSM approaches achieve the BLEU score for Myanmar to Kayah and Kayah to Myanmar translation.

In [3], the author suggests a phrase-based statistical machine translation model and decoding algorithm that enables us to assess and contrast numerous other phrase-based translation models that have also been proposed. In order to better understand and explain why phrase-based models perform better than word-based models, this framework conducts a number of experiments. The empirical results, which are consistent across language pairs investigated, suggest that the best levels of performance can be attained using relatively straightforward techniques, such as lexical weighting of phrase translations and heuristic learning of phrase translations from word-based alignments. Interestingly, learning phrases from word alignment models with high accuracy and learning phrases with more than three words have little effect on performance.

## 3. Myanmar and Lisu Languages

Myanmar language, which is spoken in that country, is the primary language of the Republic of the Union of Myanmar. It is also known as Burmese language. Myanmar is a member of the Tibeto-Burman ethnic group. Myanmar is spoken by approximately 34.5 million people as their first language, with another 10 million speaking it as a second language. Furthermore, Myanmar is a language spoken in a few areas of the United States as well as neighboring countries such as Bangladesh, Malaysia, and Thailand. Ethnic groups speak Myanmar as a second language in addition to their native tongues.

Myanmar script is made up of (33) consonants: (က...to…အ), Medias (ျ၊ ြ၊ ွ၊ ှ), vowels (ါ၊ ာ၊ ိ၊ ီ၊ ု၊ ူ၊ ေ၊ ဲ၊ ံ), special characters (ဥ် ၊ ၍ ၊ ၏ ၊ ဿ ၊ ဤ ), punctuation (၊ and ။) and digits (၀၊ ၁၊ ၂၊ ၃၊ ၄၊ ၅၊ ၆၊ ၇၊ ၈၊ ၉). Myanmar's writing system is left to right. Furthermore, the sentence structure is subject-object-verb (SOV).

Speaking countries for Lisu include China, Myanmar, Thailand, India, and Laos. The bulk of Lisu speakers still reside in NW Yunnan, China, which is considered to be their current homeland. 575,000 people in West Yunnan, Sichuan, and the higher portions of the Salween and Mekong River regions in China currently speak it (1990 census). Around Lashio, in Wa State, Myitkyina and Bhamo, around Putar, and around Loilem in Shan States, 126,000 Lisu speakers reside (1987 estimate).

The provinces of Chiangmai, Chiangrai, Maehongson, Kamphaeng, and Phet in Thailand are home to more than 25,000 Lisu speakers. To the northwest of Thailand, some Burmese have moved. There are more than 1000 Lisu speakers in India (Bradley 1994). The overall population is 657,000 according to Ethnologue, 635,000 according to Wurm (1981). Four primers, a sizable number of graded readers, health books, folktales, a Bible, hymnals, and Bible commentaries are all available in Lisu. Table 1 describes Lisu Fraser Scripts Consonants, Semi Vowels and Vowels.

**Table 1. Lisu Fraser Scripts Consonants**

| | | | | | | |
|---|---|---|---|---|---|---|
| **Consonants** | B | P | ɗ | D | T | Ꞇ |
| | G | K | ƙ | J | C | Ɔ |
| | Z | F | Ƒ | M | N | L |
| | S | R | Ʀ | H | Ɠ | ſ |
| | X | | | | | ƕ |
| **Semi Vowels** | W | Y | | | | |
| **Vowels** | A | Ɐ | E | Ǝ | I | O |
| | U | ∩ | | | | |

In Table 2, Tone indication (Single Tones, Double Tones), Punctuations and Digits are showed.

**Table 2. Tone Indication, Punctuations and Digits**

| | | | | | | |
|---|---|---|---|---|---|---|
| **Single Tones** | ˙ | ˌ | ˷ | ; | : | ‥ |
| **Double Tones** | ˙; | ˌ; | ˙: | ˌ: | ...: | ...; |
| | ˷; | ˷: | | | | |
| **Punctuation** | - | -. | ? | ! | “ ” | 0 |
| **Digits** | 0 | 1 | 2 | 3 | 4 | 5 |
| | 6 | 7 | 8 | 9 | | |

Subject-Object-Verb (SOV) language is used in Lisu sentences. The direct Object is followed by the Verb, which is followed by the Subject. A clause is made up of one verbal proposition and the words that go with it. It is normal for the verb to be the last word in an unmarked sentence. The majority of the grammatical marking is carried by particles that follow nouns and verbs. For example, noun phrases contain required words to indicate topic nya, possession, location, instrument, and so on. Following subjects and objects, these particles are optional. The use of numbers with or without nouns requires use of classifiers.

## 4. Statistical Machine Translation (SMT)

An offspring of empirical machine translation (EMT) systems is the SMT system. Despite the fact that they rely on a large number of parallel aligned corpora, these systems learn to translate a large number of previously translated texts to another language. A structure for translating text from one natural language to another is a statistical machine translation system. To be effective, sentences in both the source and the destination languages or bilingual or multilingual languages are required.

A statistical machine learning algorithm is used to build the statistics tables. The procedure is referred to, and the statistical tables contain the statistical data. This statistical data is used to determine the best decoding phrase outcome.

SMT (Statistical Machine Translation) creates a correctly translated sentence using decoders and input sentences for training. Using SMT and the phrase-based decoder system, the translation quality of several language pairings has significantly improved. Large parallel data corpora are used by statistical machine translation (SMT) systems for model training. Furthermore, the size of the parallel data corpus influences the effectiveness of statistical machine translation systems.

On the other hand, Myanmar is a low-resource language, and parallel corpora in Myanmar and Lisu are uncommon. As a result, in this system, a parallel corpus for Myanmar and Lisu is created first, followed by a statistical machine translation system between the two languages. Myanmar sentences from the UCSY-corpus [4], which includes the Myanmar-English language pair, are used to construct the Myanmar-Lisu parallel corpus. These Myanmar phrases are then handwritten into Lisu. This corpus also includes parallel sentences from both spoken and written school text books in both languages. The Myanmar-Lisu parallel corpus contains over 15K parallel sentences.

### 4.1. Phrase-based Statistical Machine Translation

Phrasal units serve as the foundation for PB-SMT translation models [3]. A phrase in this case is only a group of words that are together it does not have linguistic significance. The phrase translation model is based on the noisy channel model. Find the translation e that, given the source phrases, maximizes the translation probability $P(e|f)$. In this case, the source language is Myanmar, and the target language is Lisu. The translation of a source sentence f into a target sentence e is modeled as equation 1.

$$e = argmax\ e\ P(e|f) \tag{1}$$

The mathematical formulation of phrase-based model is as equation2.

$$P(e|f) = argmax\ P(f|e)P(e) \tag{2}$$

This allows for a language model $P(e)$ and a separate translation model $P(f|e)$.

Using decoding, the input source sentence f is segmented into a sequence of I phrases $f_1^{-I}$.

Each source sentence $f_i^- in\ f_1^{-I}$ is translated into a target language$e_i^-$. The arrangement of the target phrase is flexible. A probability distribution is used to model phrase translation $\emptyset(f_i^-|e_i^-)$ . Keep in mind that the Bayes rule, from the perspective of modeling, inverts the translation direction.

In order to mimic the reordering of target output, a relative distortion probability distribution is used d $(a_i - b_{i-1}), a_i$ d the begin position of the source phrase that was translated into the I target and $b_{i-1}$ indicates where the source language ends when it is translated into the target language ( i -1).

All of the studies employ a joint probability model to train the distortion probability distribution d(.). Alternately, a more straightforward distortion model is applied$d(a_i - b_{i-1}) = \propto^{|a_i - b_{i-1}-1|}$ with an appropriate value for the parameter $\propto$.

In order to calibrate the output length, it introduce a factor for each generated target word in addition to the trigram language model $p_e$. This is a straightforward method for improving performance. This factor is usually greater than one, biasing longer output.

In conclusion, according to the model, the best target output sentence $e_{best}$ given a source input sentence f is

$$P(f_1^{-I}|e_1^{-I}) = \prod_{i=1}^{I} \emptyset(f_i^-|e_i^-)\ d(a_i - b_{i-1}) \quad (3)$$

All phrase pair are gathered to align the word alignment: Only the words within a pair of legal phrases are aligned words from the other phrase are not. Using the phrase pairs that have gathered, it estimates the phrase translation probability distribution by relative frequency:

$$\emptyset(f^-|e^-) = \frac{count(f^-, e^-)}{\sum_{f-} count(f^-, e^-)} \quad (4)$$

## 5. System Overview

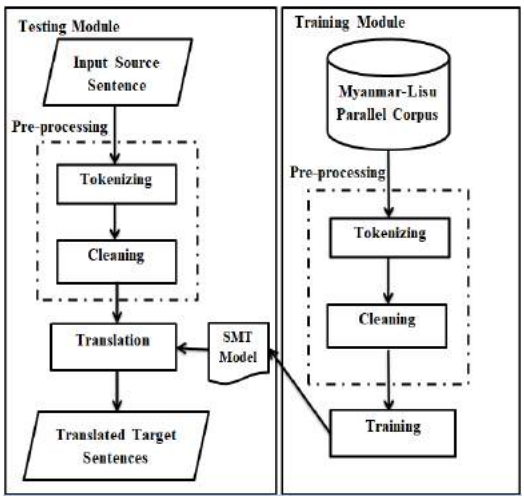



**Figure 1. The system flow of the Proposed System**

The proposed system has two primary modules. The first is the Training Module, and the second is the Testing Module. To begin, a parallel corpus for Myanmar and Lisu is required for the training module. To prepare for data pre-processing, tokenization and cleaning are used for the corpus data. The Moses toolkit is then used to train the SMT models [9]. In the testing module, data must be pre-processed the input source sentences. Following that, a trained SMT model is used to translate and produce translated sentences.

## 6. Implementation of the System

In this section, the implementation of the proposed system is described as detail.

### 6.1. Dataset and Pre-processing

One of the low-resource languages is Myanmar. There aren't many parallel corpora between Lisu and Myanmar at the moment. As a result, a parallel Myanmar-Lisu corpus is created for this system. Myanmar texts from the UCSY Myanmar-English Corpus [4] are collected and manually translated into Lisu to create a new Myanmar-Lisu parallel corpus. For the cleaning the corpus, Moses's clean scripts [9] was used. The corpus contains approximately parallel sentences from school textbooks, and spoken textbooks in both languages. This parallel corpus contains over 15K parallel sentences. The parallel sentences in this corpus are divided at the syllable segmenatition. For the syllable segmentation tool [5] was used, and Lisu sentences were manually segmented. In order to train the Myanmar-Lisu SMT models, the parallel corpus is randomly divided into three division files, as shown in Table 3.

**Table 3. Statistics of Myanmar-Lisu parallel corpus**

| Files | No. of sentences |
|---|---|
| Training File | 13590 |
| Tuning File | 758 |
| Testing File | 749 |
| Total Sentences | 15096 |

### 6.2. Moses SMT System

This system utilized the PBSMT technology offered by the Moses toolkit in [9] to train the PBSMT statistical machine translation systems. The word segmented source and target languages

were then aligned using GIZA++. The alignment was symmetric zed using a heuristic. The msd-bidiretional-fe option was also used to train the lexicalized reordering model.

The 5-gram language model is then trained with Kneser-Ney discounting using KenLM by the system. The Moses decoder was used to decode, and the system was trained using the standard methods of MERT and decode. Moses' default settings are also used for all experiments.

### 6.3. Experimental Results

Several automatic MT evaluation techniques have been put out in recent years. The BiLingual Evaluation Understudy is one of them (BLEU). Many MT researchers have utilized BLEU to demonstrate the potency of their inventive approaches to building MT systems. A rating system called BLEU compares the accuracy of N-grams to reference translations made by human translators. BLEU is then used the experiments of Myanmar-Lisu Statistical Machine Translation models. The results of these trials' BLEU tests are shown in Table 4.

**Table 4. Evaluation result of Myanmar-Lisu SMT models**

| SMT Model | BLEU |
|---|---|
| Myanmar-Lisu | 47.06 |
| Lisu-Myanmar | 43.59 |

The Myanmar-Lisu model works better than the Lisu-Myanmar model, according to the experimental findings. The finding is that some names cannot be translated and that some Lisu sentences do not use a particular verb when being translated. This is the reason why the Myanmar-Lisu SMT model has a greater BLEU score than the Lisu-Myanmar SMT model.

## 7. Conclusion

The system investigated the phrase-based statistical machine translation models for translations between Myanmar and Lisu in both directions. A corpus of over 15K parallel sentences between Myanmar and Lisu Languages are constructed by using Perl and Python programming languages. The parallel corpus is then used to create a translation model, and a monolingual corpus is prepared to build language models for the two languages. The experimental results for the Myanmar-Lisu model are 47.06 and 43.59 for the Lisu-Myanmar model. This system results show that the performance of Myanmar-Lisu model is better than the Lisu-Myanmar model. More data must be collected in order to improve translation performance.

## References

[1] Wendesen Endale, “Bi-directinal Afaan Oromo-English Statistical Machine Traslation”, Department computer science Mettu University, Ethiopia, The International Journal Research Published paper at ISSN:2251 2015]

[2] Zar Zar Linn, Ye Kyaw Thu and Pushpa B. Patil, “Statistical Machine Translation between Myanmar (Burmese) and Kayah”, Journal Of Intelligent Informatics and Smart Technology, Vol.4, April 2020].

[3] P.Koehn,F.J.Och and D. Marcu, “Statistical phrase-based translation.” in Proc. of HTL-NAACL, 2003, pp. 48– 54.

[4] Yi Mon Shwe Sin and Khin Mar Soe and Khin Yandanar Htwe “Large Scale Myanmar toEnglish Neural Machine TranslationSystem”. Proceeding of the IEEE 7th Global Conference on Consumer Electronic (GCCE 2018).

[5] D. Chenchen, H. T. Z. Aye, W. P. Pa, K. T. Nwet, K. M. Soe, U. Masao, S. Euchiro, “Towards Burmese (Myanmar) Morphlogical Analysis: Syllable-based Tokenization and Part-of-speech Tagging”, ACM Transactions on Asian and Low-Resource Language Information Proceessing (TALLIP), Vol. 19, No. 1, pp. 5, 2019.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C.Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation”, in Proc. of ACL, 2007, pp. 177–180.

[7] Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, “A Large-scale Study of Statistical Machine Translation Methods for Myanmar Language”, in Proc. of SNLP2016, February 10-12, 2016.

[8] Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages", 5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU Workshop), 09-12 May, 2016, Yogyakarta, Indonesia, Procedia Computer Science, Volume 81, 2016, pp. 250–257.

[9] Philipp Koehn, "Moses Statistical Machine Translation System", User Manual and Code Guide,University of Edinburgh, http://www.statmt.org/moses/manual/manual.pdf

[10] Marcu, D. and Wong, W. (2002). "A phrase-based, joint probability model for statistical machine translation". In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNL.

[11] Statistical machine translation - Wikipedia

[12] Lisu language - Wikipedia

[13] Myanmar language - Wikipedia

[14](PDF) How Do You Write Lisu? (researchgate.net)

# A Comparative Study of Lexicons on Aspect-Based Opinion Mining Using Support Vector Machine

Nway Nway Aung, Moe Moe San
Hnin Cherry, Soe Kalayar Naing
*University of Computer Studies, Pathein*
*nwaynwayaung@ucspathein.edu.mm, moemoesan@ucspathein.edu.mm*
*hnincherry@ucsy.edu.mm, soekalayarnaing@ucspathein.edu.mm*

## Abstract

*Text is the primary means of information transmission in the digital age. The internet is full of messages, news items, reviews, and opinionated pieces of information. Online shopping and product reviews are commonplace among users. There is a need for a mechanism to extract and summarize pertinent information to improve the decision-making process because this feedback is made public to help others with their purchasing decisions. Data mining and natural language processing techniques are combined to create the technology known as aspect-based opinion mining. This system's main goal is to extract user feedback and product attributes from laptop reviews using common words and natural language. Depending on the laptop models, five distinct datasets will be used for analysis. It will be suggested to categorize laptop model reviews as favorable, unfavorable, or neutral. It will be taken out three aspects (RAM, Processor and Battery) by using Support Vector Machine classifier. It will be decided better accuracy method by comparing SentiWordNet and SentiStrength lexicons in this proposed system. It can be evaluated performances by calculating of accuracy, precision, recall and f-score measurement. According to the experimental findings of this work, Sentiwordnet performs better than the sentistrength lexicon. The aspect classified result showed an increase in SVM accuracy from 44% to 95%.*

**Keywords:** SentiWordNet, SentiStrength, SVM

## 1. Introduction

Aspect based opinion mining is an established scientific field. The analysis of subjective data toward various entities is known as opinion mining. Typically, the term "entity" refers to a product, an organization, a service, or any of these characteristics.

Natural Language Processing is main objective of opinion mining to locate and extract opinions from written material, such product reviews for laptops. A wonderful source of information to obtain viewpoints on a particular issue is the Internet. Organizations are leveraging online content to inform their manufacturing and purchasing decisions as the Web has grown. As a vast amount of data in website, it is difficult quickly to get important information.

A system that can automatically summarize paper is becoming more and more useful. To enable users to make educated decisions, a system likes as gathers pertinent data and presents it in a way that is simple to read and understand. The suggested technique utilizes a linear Support Vector Machine (SVM) classifier to extract the aspect from online laptop product customer reviews. The laptop reviews are to be categorized as positive, negative, or neutral for each feature. By contrasting two lexicons, this opinion mining process will move forward. (SentiWordNet and SentiStrength).

## 2. Related Work

[1] The Authors presented the implementation of a sentiment categorization model using a convolutional neural network and a bidirectional gated recurrent device. That system suggested using the laptop dataset and the restaurant dataset in parallel using bi-GRU. Use bi-GRU and CNN simultaneously because the bi-GRU components used as the CNN input may not yield as much useful feature space from the original dataset as the CNN. Accuracy measured using CNN, LSTM, SVM, and BiLSTM. On the SemEval

Restaurant dataset, BiGRU-CNN was shown to have the highest accuracy of all models, with a score of 79.94. On the laptop dataset, it has a precision of 73.54, whereas CNN has the lowest accuracy (57.14). It has been demonstrated that the combination of BiGRU with CNN helps maintain the linear performance of the BiGRU improve outcomes.

[2] Highlighted aspect-based opinion mining and concentrated on it on restaurant review reviews. This system used to extract data from the document, including SentiWordNet, dependency parser, and POS tagging. Categorized reviews as favorable, unfavorable, indifferent, and major components of reviews include location, fare, service, timing, and personnel. Says that 70% of the research objective is satisfied by the results. It is not a domain-specific dictionary, SentiWordNet. Furthermore, the proposed system will be able to recognize more corrected beliefs if domain-specific SentiWordNet files can be created.

[3] Presented a comparison of sentiment analysis for the Myanmar language using three different machine learning algorithms, including logistic regression, support vector machines, and random forests. To analyze the effectiveness of the word vector approaches, Word2Vec, TFIDF test feature vector representation, and pre-trained Word2Vec were used. By attaining 80% of the F1-score, it was demonstrated that the Logistic Regression classifier with Word2Vec performed better than the other two Machine Learning algorithms.

[4] Introduced an unsupervised method for discovering aspect categories using association rule mining on co-occurrence frequency data from a corpus. This system used a supervised variation with an F1-score of 84% that outperforms current techniques. The author is interested in learning how incorporating outside data might enhance the outcomes. They are more interested in adopting more semantic alternatives, such as ontologies or other semantic networks, but lexicons are a wonderful way to accomplish this, as shown by.

[5] Presented for identifying and interpreting underlying themes and viewpoints in English airline reviews. Eight distinct airline industry-specific features that can be used for opinion mining jobs have been identified. Small-scale elements included the cabin, amusement, food, on-board services, off-board services, seat, personnel and valuables. That system utilized conditional random fields (CRF) with stochastic gradient descent optimization and pre-trained word embedding for sequential analysis before applying machine and ensemble learning algorithms for the second stage. Optimized CRF outperforms some baseline systems with ROC-AUC score of 96% and F1 score of 94%. Although limited to a few reviews, future research could use a larger dataset and the approaches suggested in this publication.

[10] Presented lexicon-based methods for determining the polarity score and categorizing the reviews as favorable or unfavorable. The movie reviews in this article have been categorized using the NLTK, Text Blob, and VADER Sentiment analysis tools. According to the results of the testing, VADER performs better than the Text blob.

## 3. Background Theory

A kind of natural language processing is that for monitoring public opinion of a certain product. It is also known as sentiment analysis, and it entails developing a system to gather and examine customer reviews of a product. There are numerous lexicons for opinion mining such as Senti-net, Sentiwordnet, Sentistrength, HowNet, etc. Among them, SentiWordNet and SentiStrength will be utilized in this suggested system.

### 3.1. SentiWordNet

SentiWordNet is a lexical database used for sentiment analysis. SentiWordNet assigns three sentiment scores—positive, negative, and neutral—to each synset of wordnet. By including data on the sentiment-related properties of text phrases, SentiWordNet is frequently used to enhance text representation in opinion mining (OM) applications. OM has a wide range of uses, from customer contact management to tracking people' opinions about products or political candidates as stated in internet forums. Figure 1 displays a portion of SentiWordNet.

| POS | ID | PosScore | NegScore | SynsetTerms | Gloss |
|---|---|---|---|---|---|
| a | 00001740 | 0.125 | 0 | able | (usually followed ... |
| a | 00002098 | 0 | 0.75 | unable | (usually followed ... |
| a | 00002312 | 0 | 0 | dorsal | facing away from ... |
| a | 00002312 | 0 | 0 | abaxial | facing away from ... |
| a | 00002527 | 0 | 0 | ventral | nearest to or faci... |
| a | 00002527 | 0 | 0 | adaxial | nearest to or faci... |
| a | 00002730 | 0 | 0 | acroscopic | facing or on the s... |
| a | 00002843 | 0 | 0 | basiscopic | facing or on the s... |
| a | 00002956 | 0 | 0 | abducting | especially of mus... |
| a | 00002956 | 0 | 0 | abducent | especially of mus... |
| a | 00003131 | 0 | 0 | adductive | especially of mus... |
| a | 00003131 | 0 | 0 | adducting | especially of mus... |
| a | 00003131 | 0 | 0 | adducent | especially of mus... |
| a | 00003356 | 0 | 0 | nascent | being born or beg... |
| a | 00003553 | 0 | 0 | emerging | coming into exist... |
| a | 00003553 | 0 | 0 | emergent | coming into exist... |

**Figure 1. SentiWordNet Data Sample**

## 3.2. SentiStrength

Sentistrength analyzes shorter sentences favorable and unfavorable emotional strengths even colloquial expressions. With the exception of radical sentences, it is as accurate as a human being for brief shared network sentences for English. Reports on sentistrength two sentimental strengths:

- -1 (not negative) to -5 (extremely negative)
- 1 (not positive) to 5 (extremely positive) SentiStrength can also report
- binary (positive/negative)
- trinary (positive/negative/neutral)

Part of SentiStrength library is shown in figure 2.

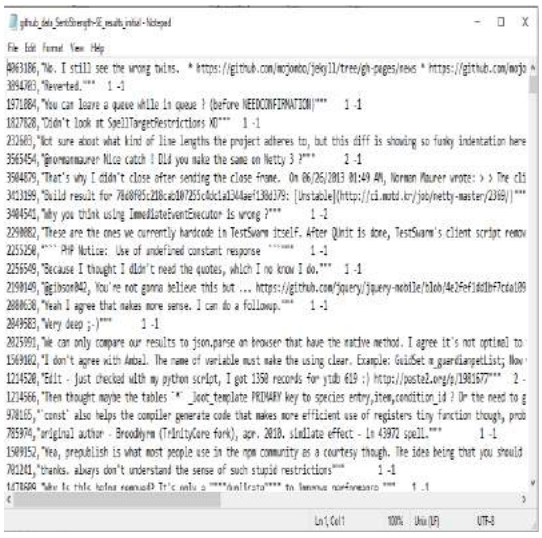


**Figure 2. SentiStrength Data Sample**

## 3.3. Term Weighting Schemes (TF-IDF method)

The term weighting method known as "Term Frequency" (TF) is based on how frequently terms appear in a document. The impact of a word on a document increases as a word's TF value increases. The weighting approach known as Inverse Document Frequency (IDF) bases on the calculations of the number of words that appear in each document. One of the most straightforward and effective weighting methods for the data is TF-IDF. Due to its straightforward formulation and effective operation on a variety of different data sets, TF-IDF and its algorithm version are the default choice [9]. This method's formulation is as follows.

$$W(d, t) = tf(t, d) * \log(N / n_t)$$

Where:

W (t, d) = term weight in document d
tf (t, d) = term frequency in document
N = the total number of document
$n_t$ = number of documents that have term t

## 3.4. Support Vector Machine (SVM)

SVM is a classifier that can distinguish between examples that belong to several categories by creating hyperplanes. In order for the SVM classifier to learn how to categorize things correctly, it must first be trained using a set of user reviews. A hyperplane that effectively distinguishes classes for the application is defined through iterative training. Let X Y represents a mapping. Y can take real numbers and is an aspect score; X is the extracted judgment about the aspect of a product.

The set of tuples that contain the aspect and the opinion itself is the work's intermediate outcome.

The following tuple represents the classifier output in storage,

$$P\,orient = (ri, aij, polij)$$

Where, *ri* is the ith review, *aij* is the feature j in review i, *polij* is the polarity of $j^{th}$ feature in $i^{th}$ review.

After: the following formula can be used to aggregate the polarities of the opinions in all the reviews and identify the aspect-opinion tuples,

For each aspect j of the product,

Orient aggregate[j] = $\Sigma i\, polij$

The total polarity value for each feature is normalized using the total number of reviews considered [7].

Orient norm[j] = $Orient\ aggregate(j) / \Sigma_i$

The opinionated text of these users is extremely unstructured, necessitating the use of a variety of natural language processing methods. During the training phase of aspect based sentiment analysis, the various characteristics of a product are identified. The qualities of a laptop's processor, RAM, battery life are examples. The support vector machine classifier is used to conduct the quantitative analysis of each aspect.

## 4. Methodology

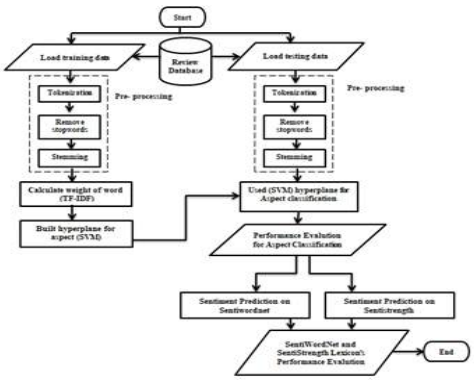



**Figure 3. Proposed System Design**

There are steps in this proposed system such as Data Collection, Data Pre-processing, Calculation weight of each word (TF-IDF), aspect classification, sentiment prediction and Accuracy Results. This system trained datasets until computation hyperplane values for aspects using SVM.

After the testing data is loaded, aspects (Ram, Processor, battery) are classified using the trained hyperplane values. Then aspect classification accuracy is measured. For the sentiment prediction, SentiWordNet and SentiStrength lexicons are loaded to classify the sentiment natures. Finally, performance evaluation is computed to compare two lexicons.

### 4.1. Data Collection

Numerous reviews are gathered from the web in order to determine the polarity of the sentences depending on their elements. There are numerous websites online where you may get a lot of consumer reviews. The reviews are gathered via the Amazon website (www.Amazon.Com).

The dataset was created by crawling public comments on the "Laptop Model" page on Amazon. There are five distinct datasets (MSI, Acer, Dell, Lenovo, ASUS) will be used for analysis. Each dataset for the reviews includes 10,000 sentences


1. The speed is fine.
2. Needs more RAM and it can't be expanded.
3. I upgraded the RAM to 16GB as I run a virtual lab for testing/school.
4. The battery life is lacking.
5. My laptop now has no battery.
6. long battery life.
7. Fast, good processing.
8. Low end processor.
9. I like the speed of the processor.
10. CPU can easily handle that type of workload for 1080p video.
11. I gave 5 stars because it's being compared to other gaming laptops and the battery life in this category is great imo.
12. To be fair I can't rate battery life just yet as I just got it but by every other measure I don't expect to be disappointed.
13. The only cons I have: the battery life is not great, it tends to run hot so I keep the second fan running which is obviously noticeably louder.
14. Great gaming laptop. Runs latest titles in have on epic settings. It's a tad loud but it's a gaming laptop and expected, same with battery life. Screen looks great, runs very smoothly.
15. The battery life is awful on this otherwise great setup. MSI Synapse software is finicky and requires regular tinkering.
16. I wanted good performance and still have some resemblance of battery life.
17. I don't like it the battery life that much but it's something you can deal with
18. The battery doesn't last long but I'm sure an upgrade battery would solve that problem.
19. The only clown fall is the battery only last 1.5-2 Hrs when not plugged in.
20. Aside from the trial software and the short battery life, lack of a webcam, its great.
21. the battery is completely shit.
22. I will be rating the laptop seven out of ten in this section of the review.

**Figure 4. Example of MSI Laptop Model Dataset**

### 4.2. Data Pre – processing

Among the five collected datasets, this system will select desired dataset and do the preprocessing step. The data cleaning processes includes tokenization, stop-word removal, and stemming during the pre-processing phase. The following step of feature selection and extraction will employ the clean dataset.

- Tokenization is the process of separating the text corpus into its component parts.

| I liked the speed of the processor. | | | | | | |
|---|---|---|---|---|---|---|
| I | liked | the | speed | of | the | processor |

Removing Stop Words and Stemming

- Stop words are extraneous words that frequently appear in writing.

- For instance, phrases like “so,” “and,” “or,” “the,” etc. First, all stop words are eliminated. The stop words in the illustration below, you, are, that, have, and the, are eliminated using this method.
- The system removes the stopwords and stemming which entails decreasing a word down into its stem, or root form.

| I | liked | the | speed | of | the | processor |
|---|---|---|---|---|---|---|
| like | | | speed | | | processor |

### 4.3. Calculation weight of each word

The cleaning words are pre-trained by using TF.IDF for weight calculation of each words.

**Example**: W (d, t) = tf ( t, d ) * log ( N / $n_t$ )

TFIDF for training record id1:
Let; Total Record=10
Total word count for record ID 1=7
To calculate= like, speed, process
TF.IDF (like) = (1/7)* log (10/1) =0.14
TF.IDF (speed) = (1/7)* log (10/2) =0.09
TF.IDF (processor) = (1/7)* log (10/3)=0.07

### 4.4. Aspect Classification by using SVM

After the calculation weight of each word, Laptop review data aspect classification is ready to classify. Three aspects (RAM, Processor, and Battery) are classified by using the Support Vector Machine Classifier. Section 3.4 provides an in-depth explanation of the SVM’s step-by-step process.

TF.IDF →(0.07, 0), (0.09, 0), (0.14, 2)
S1= (0.07, 0), S2= (0.09, 0), S3= (0.14, 2)
Bias=> S1= (0.07, 0, 1), S2= (0.09, 0, 1), S3= (0.14, 2, 1)
*α1*(0.07,0,1) (0.07,0,1)+ *α2*(0.09,0,1) (0.07,0,1)+ *α3*(0.14,2,1) (0.07,0,1)=0
*α1*(0.07,0,1) (0.09,0,1)+ *α2*(0.09,0,1) (0.09,0,1)+ *α3*(0.14,2,1) (0.09,0,1)=0
*α1*(0.07,0,1) (0.14,2,1)+*α2*(0.09,0,1) (0.14,2,1)+ *α3*(0.14,2,1) (0.14,2,1)=2
1.005 *α1+1.006 α2+1.1 α3=0*
1.006 *α1+1.008 α2+1.01 α3=0*
*1.009 α1+ 1.01 α2+5.02α3=2*
*α1=-47.2, α2=46.6, α3=0.51*
*ώ=-47.2*(0.07, 0, 1) + 46.6(0.09, 0, 1)+ *0.51*(0.14,2,1)=(-50.5,50.8,1.6)=1.6
So, (0<1.6(Hyperplane value) < 2) →Processor
Similarity if [2 < (Hyperplane value) <4]→RAM
if [4< ( Hyperplane value ) <6] →battery

**Figure 5. Example of aspect computation using SVM Classifier**

**Table 1. Example of Aspect Classification**

| id | Reviews | Aspect |
|---|---|---|
| 1. | I liked the speed of the processor. | Processor |
| 2. | The battery life is lacking. | battery |
| 3. | Needs more RAM and it can't be expanded. | RAM |

### 4.5. Sentiment Prediction

For the sentiment prediction, the determination will be made by the used of two lexicons. SentiWordNet and SentiStrength lexicons are loaded to classify the sentiment natures.

**Table 2. Example of Final Result for Laptop Reviews**

| id | Reviews | Aspect | Sentiwordnet | Senti-Strength |
|---|---|---|---|---|
| 1. | I liked the speed of the processor. | Processor | Positive | Positive |
| 2. | The battery life is lacking. | battery | Negative | Negative |
| 3. | Needs more RAM and it can't be expanded. | RAM | Neutral | Neutral |

## 5. Evaluation Metrics

Five evaluation metrics, which are precision, recall, F-measure, accuracy and failure-ratio, are used to evaluate the effectiveness of the system. These are calculated by using Eq. (5.1)-(5.5) respectively.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F- measure} = \frac{Precision * Recall}{Precision+Recall} \quad (4)$$

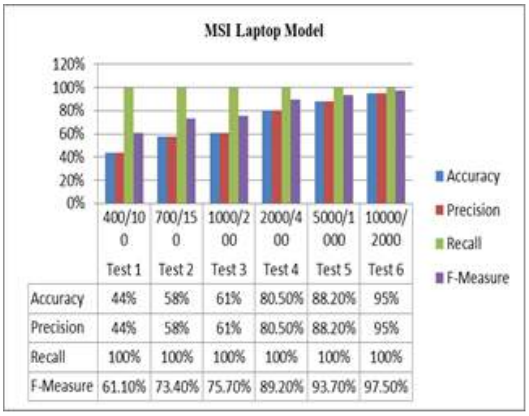
where:

- TP refers to the number of true positive reviews.
- TN refers to the number of true negative reviews.
- FP refers to the number of false positive reviews.
- FN refers to the number of false negative reviews.
- Number of Misclassified Reviews refers to the reviews labelled to the class label which was not included in the actual class labels.
- Total Number of Reviews refers to the number of all reviews.

## 5.1. Experimental Result

**Experimental Evaluation results for Aspect Classification**

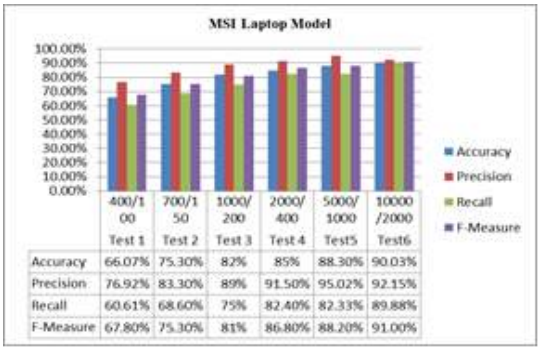
Six distinct training dataset and testing dataset pairs are employed for aspect classification in each analysis.




| | 400/100 Test 1 | 700/150 Test 2 | 1000/200 Test 3 | 2000/400 Test 4 | 5000/1000 Test 5 | 10000/2000 Test 6 |
|---|---|---|---|---|---|---|
| Accuracy | 44% | 58% | 61% | 80.50% | 88.20% | 95% |
| Precision | 44% | 58% | 61% | 80.50% | 88.20% | 95% |
| Recall | 100% | 100% | 100% | 100% | 100% | 100% |
| F-Measure | 61.10% | 73.40% | 75.70% | 89.20% | 93.70% | 97.50% |

**Figure 6. Aspect Classification Experiments**

**Experimental Evaluation results for SentiWordNet**



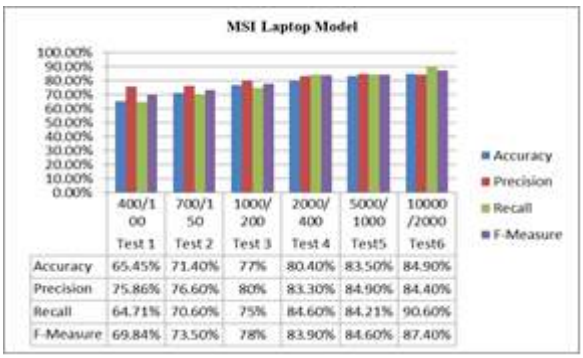

| | 400/100 Test 1 | 700/150 Test 2 | 1000/200 Test 3 | 2000/400 Test 4 | 5000/1000 Test5 | 10000/2000 Test6 |
|---|---|---|---|---|---|---|
| Accuracy | 66.07% | 75.30% | 82% | 85% | 88.30% | 90.03% |
| Precision | 76.92% | 83.30% | 89% | 91.50% | 95.02% | 92.15% |
| Recall | 60.61% | 68.60% | 75% | 82.40% | 82.33% | 89.88% |
| F-Measure | 67.80% | 75.30% | 81% | 86.80% | 88.20% | 91.00% |

**Figure 7. Sentiment Classification Experiment (SentiWordNet)**

SentiWordNet Lexicon is employed in each study using 6 separate training dataset and testing dataset pairs.

**Experimental Evaluation results for SentiStrength**

SentiStrength Lexicon is utilized in each study with 6 separate training dataset and testing dataset pairs.




| | 400/100 Test 1 | 700/150 Test 2 | 1000/200 Test 3 | 2000/400 Test 4 | 5000/1000 Test5 | 10000/2000 Test6 |
|---|---|---|---|---|---|---|
| Accuracy | 65.45% | 71.40% | 77% | 80.40% | 83.50% | 84.90% |
| Precision | 75.86% | 76.60% | 80% | 83.30% | 84.90% | 84.40% |
| Recall | 64.71% | 70.60% | 75% | 84.60% | 84.21% | 90.60% |
| F-Measure | 69.84% | 73.50% | 78% | 83.90% | 84.60% | 87.40% |

**Figure 8. Sentiment Classification Experiment (SentiStrength)**

From the experimental results obtained, it can be seen that SentiWordNet lexicon gave the highest test accuracy better than SentiStrength lexicon over Support Vector Machine Classifier.

## 6. Conclusion

In the commercial domain, aspect-based opinion mining is mostly used to recognize consumer opinion. Utilizing customer feedback, this system generates an opinionated summary. The duty of aspect and opinion extraction was the key accomplishment. With the use of customer reviews of laptop products, this system would deliver "aspect-based opinionated summary." This system's primary goal is to compare SentiWordNet and SentiStrength to the Support Vector Machine (SVM) classification algorithm in order to choose the method that will provide the most accuracy. The experimental results of this work show that sentiwordnet performs better than the sentistrength lexicon. According to the experimental findings of this work, the SVM method classification accuracy for reviews of MSI laptop models was 95%.

## References

[1] C Sindhu, Bihanga Som, Samar Pratap Singh, "Aspect Based Opinion Mining Leveraging Weighted BiGRU And CNN Module in Parallel", Institute of Science & Technology, Kattankulathur, India, Ieee Xplore: 04 August 2021.

[2] H.A. Caldera, I. K. C. U. Perera, "Aspect based opinion mining on restaurant reviews", University of Colombo School of Computing, Colombo, Sri Lanka, 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA).

[3] Hay Mar Su Aung, Win Pa Pa," Analysis of Word Vector Representation Techniques with Machine-Learning Classifiers for Sentiment Analysis of Public Facebook Page's Comments in Myanmar Text", ICCA 2020 Conference, Myanmar,2020.

[4] Kim Schouten, Onne Van Der Weijde, Flavius Frasincar, and Rommert Dekker, "Supervised and Unsupervised Aspect Category Detection for Sentiment Analysis with Co-occurrence Data", IEEE TRANSACTIONS ON CYBERNETICS, Netherlands: April 2017.

[5] Kanishk Verma, Brian Davis, "Implicit Aspect-Based Opinion Mining and Analysis of Airline Industry Based On User-Generated Reviews", Springer Nature Computer Science, Volume 2, Article number: 286 (2021).

[6] Pratima More, Archana Ghotkar, "A Study of Different Approaches to Aspect-based Opinion Mining", International Journal of Computer Applications (0975-8887) Volume 145-no.6, India, July 2016.

[7] Raisa Varghese, Jayasree M," Aspect Based Sentiment Analysis using Support Vector Machine Classifier", International Conference on Advances in Computing, Communications and Informatics (ICACCI),2013.

[8] Sonal Meenu Singh, Nidhi Mishra, " Aspect based Opinion Mining for Mobile Phones", IEEE, India, October 2016.

[9] Sheng-Lung Peng, Le Hoang Son, G.Suseendran, D.Balaganesh Editor, "Intelligent Computing and Innovation on Data Science", Springer, 2019.

[10] Venkateswarlu Bonta, Nandhini Kumaresh and N. Janardhan,"A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis", Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 1-6 © The Research Publication, www.trp.org.in,March 2019.

# Myanmar Entity Identification for Natural Language Understanding Using Bidirectional LSTM

Saung Thazin Phway, Win Pa Pa
*University of Computer Studies, Yangon*
*saungthazinphway@ucspathein.edu.mm*

## Abstract

*Entity Identification (EI) refers to as entity chunking, extraction, recognizing and categorizing key information (entities) in text. An entity can be one word or series of words which constantly indicates to the same thing. Every detected entity is classified into a predetermined category. Developments are performed by applying deep neural network architectures on word level Myanmar sentences. Entity-annotated corpus for Myanmar language is manually constructed. Entity-annotated corpus in sentences supported from Asian Language Treebank (ALT) parallel corpus are also used in developing. EI supports easily to identify the clue elements in text , like names of person, location, brands, monetary values, time and more. In this paper, using Bidirectional Long Short Term Memory (BiLSTM) network outcomes the highest F-score value with 83.65%.*

**Keywords:** BiLSTM, Myanmar Language, Entity Identification, Deep Learning

## 1. Introduction

Entity Identification is a crucial thing in all Natural Language Processing (NLP). EI is indispensable for Myanmar NLP. Appropriate entity classifications and categorizations are highly subjective and a major challenge for NLP researchers. It is EI' s responsibility to find and sort within a text that already has a predefined category such as person, corporation, situation, date_time_month and number of units, etc. Constructing the experiment using BiLSTM model to reveal the effect of deep learning of Myanmar sentences. This study applies very work neural network to work in Myanmar EI.

## 2. Related Work

Significant amount of work has been done in the name entity recognition task in another language. We presented for this paper that will emphasis on the identification of Myanmar sentence so it is calmed extened in Myanmar NLP. AP.C. J. Chiu, and E. Nichols an et.al [2] surveyed the literature of NERwith Bidirectional LSTM-CNNs. In neural networks, partial lexicons match by encoding 'novel method. This model improves the performance of CoNLL2003 and On to Notes datasets. From publicly-available sources, the F score of 91.62 on CoNLL-2003 and 86.28 on Onto Notes constructed by applying two lexicons.

Huanzhong Duan and Yan Zheng, [3] presented that feature templates varies in window size and the fitting of sequence level are so essential for Chinese names entities. The contribution of F score measure predicted for CRF by adding Chinese characters, part of speech, prefix and suffix. The results show that choosing proper factor templates and succession the sets of label may develop the accuracy of CNER, abbreviate the process of model training and pare the consumption of system source.

Hesheng Xu and Bin Hu [4] proposed LSTM-CRF deep learning Model. The analysis outcomes that word sequence labeling corpus on the named entity of the model training value is elevated than the word sequence labeling corpus for 88.13%. The names of place and organization is obtained by the Bi-LSTM-CRF model using word segmentation are the F1 value is 67.60%. By using character segmentation on this model for F1 value is 89.45%. The use of Bi-LSTM-CRF model on word segmentation is recognizing for more acceptable extended entities.

In Z. Huang, W. Xu, and K. Yu. [5], researched that the based models for sequence tagging using all kinds of LSTM. By comprising

the networks of LSTM and BI-LSTM, LSTM - CRF and BI-LSTM-CRF, sequence tagging is constructed. At first, NLP criterion sequence tagging data set is used by BiLSTM-CRF model that can construct closing to performance on POS, extraction and NER datasets. In sum, that is vigorous and has less vulnerability in embedding of word as matched to before observations.

## 3. Myanmar Language

The official language of the Republic Union of Myanmar is historically also known as Burmese also called Myanmar language. A cordial of tonal language is used by more than 50 million people. Generally, Myanmar writing script includes absolutely 75 characters. These characters can be more classified into 12 groups such as 34 consonants, 4 medial letters, 8 dependent vowels, one Sign Virama and one Sign Asat [1]. One group determines three independent vowels, three independent various signs and the characters in these group can perform by standing alone syllables and 10 Myanmar digits and 2 punctuation marks. Other group is hammered with four independent vowels and one Myanmar Symbol preceding. By adding to those characters, white space is applied between phrases and also no obvious rule to apply it. Myanmar-Numerals are decimal-based, and displays zero to nine in series. Thousands of separators are not used; in spite of, spaces are frequently applied among digits for clear reading documents. The punctuation marks part in a related style to the coma (,) and the duration in other languages like English, respectively [1].

### 3.1. Challenges of Entity Identification for Myanmar Language

The work of recognizing entities in Myanmar sentence is complicated correlated to other languages considering countless arguments. The paucity of resources such as entity-annotated corpus, entity lists a thesaurus is the reasons of Myanmar Language. Myanmar language is source- strained language. No capitalization and distinctly characteristic is the major symbol of useful entities for another language. Its writing structure is complimentary order and generates entities a complicated process. This uncertainly of entities may start to issues in identifying entities into pretend categories. The identification and interpretation in unstructured text for appropriate entities are dissimilar. In perception, the provocation of entities that some expressions are difficult to realize applying the rule of NLP because they have to appear the unrolled class. So, the invention of unbounded variety and new expressions are continuously being there.

## 4. Methodology

### 4.1. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is suitable concerning with figures and it also has accomplished outcomes in computer perception. In recent times, it has been used in NLP tasks and it can be launched that it exceeded historic architecture like bag of words, n-grams, etc.

### 4.2. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is capable for training sequential instructions. RNN makes the related application frequently on current of data. RNN appears with the acceptance which the output is being reliant on the earlier calculations whilst a conventional neural architecture accepts entire inputs and outputs are self-reliant of one-by-one another.

### 4.3. Bidirectional LSTM

In sequence tagging task, a moderation of the LSTM is designed to represent information and has ingress to not only path but also future contexts. In bidirectional, our input goes along to two directions, making a BiLSTM dissimilar from regular LSTM. In reverse order, the forward LSTM takes the sequence and makes hidden states the backward. It generates two individual hidden states for each sequence so the information of sequences from bidi rections is recollected that is behind the idea. By linking together, the number of two hidden states, the output of final is established. The BiLSTM architecture, where the input layer is ‘xt’, the output generated sequence is ‘yt’, tag sequence, the cell state is c, and the hidden state is ‘h’. In entity identification, BiLSTM can enhance its ability to take into sequential information during EI process. The example of Myanmar sentence is described as follows.

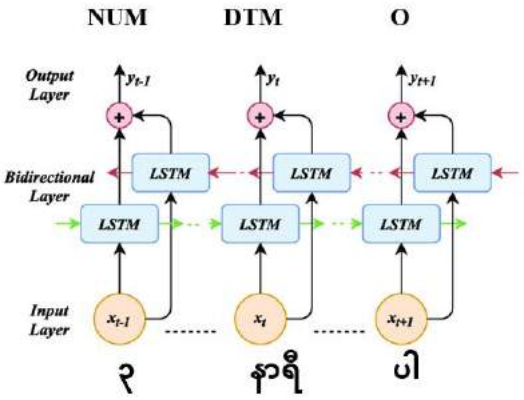



**Figure 1. Bidirectional Long Short-Term Memory**

## 5. Our Work

The extraction of datasets in the ALT corpus, that is tagged, annotated, and constructed specifically to train the identifier to predict entities such as name, location etc. Data set also includes additional features for POS that can be used in identification and manual word segmentation. Each predicted tag and real tag are checked and f-score is generated. Only with one feature sentence are processing. Experiment is performed by using BiLSTM to eliminate the need for most feature engineering.

### 5.1. Training Setup

Experiment is carried out as word-based sequence labeling, in which the BiLSTM neural network is trained. To convert the entities into a sequence labeling problem, a label is tagged for word-level to make the entities in training and test data. Word segmentation algorithm "word-break" of Myanmar sentences for word-level data representation and word-break labeling.

### 5.2. Data Preparation

Entity-annotated corpus for the Myanmar language is restricted and distributed. No other acceptable Myanmar entity-annotated corpus that has more data as entity-annotated corpus. For entity-annotated corpus constructing, Myanmar sentences are chosen from ALT parallel corpus which are manually segmented according to predefined entity tags. In recently,3375 sentences in total, 773 sentences for development, 240 sentences for testing and 2362 sentences for training are developed. Sixteen types of entity tags for manual annotation: PER, LOC, ORG, DTM, TIME, FW, NUM, QTY, CUR, N, V, O, PRON, FOOD, TITLE and PUNC.

In addition, location entities include man-made structures like airports, highways, streets, factories and monuments, etc., 'ORG' tag is defined to annotate names of organizations (government and non-government organizations, institutions, corporations, companies and other groups of people defined by an established organizational structure). 'TITLE' identifies the name of the start (e.g Mrs., Ms., Mr.). FOOD classifies Cuisines, various type of food such as Chinese Food, India food, Myanmar food etc. Number identifies numeric number. QTY recognizes quantity, distance count. DTM identifies date, time, month, and year. N represents a thing. V represents the action or state in a sentence. PRON refers to identify someone. CUR identifies the system of money general use in a particular country. TIME identifies indefinite continued progress of existence or events. FW means foreign word of a country or language. Table 1 lists the entity division in house entity corpus and Table 2 lists of entities distribution in house entity corpus.

**Table 1. Define Entity List**

| Define Entity types | Description |
|---|---|
| PER | Person name or family (ဒေးဗစ်စမစ်၊ဝမ်၊လီ) |
| ORG | governmental or cooperate name (အကအဖွဲ့) |
| LOC | Location name of publicly or graphically decided location (တရုတ်) |
| Title | မစ်၊ မစ္စတာ၊ မစ္စ |
| FOOD | Various kinds of food such as Chinese, India, Myanmar food etc. (ပေါင်မုန့်မီးကင်) |
| NUM | Numeric numbers (တစ်၊ နှစ်) |
| QTY | Distance, Money, Quantity, Count (ခု၊မိုင်) |
| DTM | Time, year, Month, days, and period (တစ်နှစ်၊ နှစ်နှစ်) |

| | |
|---|---|
| N | Represent a thing (အိတ်) |
| V | Action or state in a sentence (သွား၊ စား) |
| O | Undefined categories (ဒါ့အပြင်) |
| PRON | A way to identify or refer a someone (ကျွန်တော်) |
| CUR | money in typically apply in a country (ဘတ်၊ ဒေါ်လာ) |
| TIME | Indefinite continued progress of existence and events (လွန်ခဲ့တဲ့၊အခု) |
| FW | Not Myanmar words (natural) |
| PUNC | "။" |

**Table 2. Corpus Data Statistics**

| Data | Total No of Entities | Occurrence of one entity (%) in Entity Tagged Corpus |
|---|---|---|
| Sentence line number of entity | 31630 | 26187 |
| PER | 210 | 26 |
| ORG | 75 | 9 |
| LOC | 2091 | 89 |
| Title | 60 | 8 |
| FOOD | 736 | 94 |
| NUM | 635 | 40 |
| QTY | 383 | 45 |
| DTM | 1069 | 39 |
| N | 3510 | 219 |
| V | 4459 | 268 |
| O | 8581 | 620 |
| PRON | 1135 | 112 |
| CUR | 207 | 30 |
| TIME | 174 | 15 |
| FW | 505 | 6 |
| PUNC | 3357 | 255 |

### 5.3. Preparation of Entity Annotated Dataset

**Table 3. Example of Annotated Dataset**

၁။ ခင်ဗျား @PRON| ရဲ့ @O| အာလူးမီးဖုတ် @FOOD| မကြာခင် @TIME| ရတော့မှာ @V| ပါ @O| ။ @PUNC|

၂။ ကိုကာကိုလာ @FOOD| တစ် @NUM| ဗူး @QTY| နဲ့ @O| တခြား @N| က @O| တော့ @O| လိမ္မော်ရည် @FOOD| တစ် @NUM| ခွက် @QTY| လောက် @O| ။ @PUNC|

၃။ ကျွန်တော့် @PRON| ကြက်သားကြော် @FOOD|@O|စား @V| ဖို့ @O| မာကြော @V| လွန်း@O| တယ် @O| ။ @PUNC|

၄။ အာလူးမီးဖုတ် @FOOD | နဲ့ @O | အချဉ်နှစ် @FOOD | နဲ့ @O | ပေး @V | ပါ @O | ။ @PUNC |

၅။စားပွဲထိုး @N | တစ်@NUM |ယောက် @QTY | က @O | နေမကောင်း @V | လို့ @O | လွန်ခဲ့တဲ့ @TIME | တစ်နာရီ @DTM | လောက်က @N | ပြန်သွားခဲ့လို့ပါ @V | ။ @PUNC |

### 5.4. Work Flow of Myanmar Entity Identification System

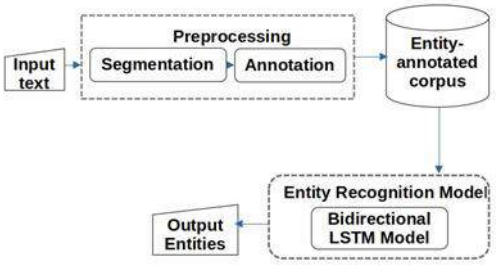



**Figure 2. System Architecture of Myanmar Entity Identification System**

A sentence has been tagged for tag sets will be accepted the system. The tagged values of the current and surrounding tokens as features will be used. In this system, input text is segmented or annotated for preprocessing to get the entity-annotated data. Entity-segmented data is trained using Entity Identification Model (BiLSTM) is generated output entities. This model will define the pronoun, verb and other categories in a sentence to classify entities with their relevant classes such as People, Organization, Title, Food, Number, Quantity, Currency, Date_time_month, Punctuation, etc. If the system found Noun, Verb, Other tag with their concerned classes by applying the arrange rules

### 5.5. Experiments with BiLSTM

To appliance the neural network architecture in developing, the neurological libraries supported with the PyTorch framework are practiced so it supports malleable selections of selected-feature inputs and output sentences. The developments are trained on Google Colab on Jupyter Notebook. The sentence of input layer into the model divides words level presentations of word embedding. Adam algorithm is tried. The initial training rate was seeds 0.0001, Batch sizes was setup 24 for the Adam algorithm, all along developing, established on the achievement on recognition generates, previously blocking was applied in order that it could delay as well the finest dropout was also used all along the developing. It is seen that dropout developing is important for well abstraction achievement and encouraging upon setting up the developing action. The invisible length was seeded into 128 in the entire development.

In addition, that developing model for Myanmar entity identification is inclined a Myanmar sentences. Word break is checked as the basic developing entity for sequence label tagging. The input presentation of character arrangement is essentially studied by using BiLSTM model within each segmented tagged input. It shows that BiLSTM work the best feature engineering is carefully prepared.

**Table 4. Total measure of precision, recall and f-score for Myanmar Entity Identification**

| Model | Precision | Recall | F-score |
|---|---|---|---|
| BiLSTM | 83.64 | 82.52 | 82.35 |

## 6. Conclusions

A wide range of the natural language applications Entity Identification is a long-term dependency study of technology. Rich resource languages with so high accuracies have developed for EI systems. Myanmar language is a resource-poor language that builds for EI system is very challenging due to unable of suitable resources. Myanmar has no upper and lowercase structure, agglutinative nature, ambiguity. The indicator of proper names for some other languages as English is easily seen. If we perform Entity Identification, BiLSTM can provide the ways to improve the accuracy and the performance metrics. To give a good quality EI model for Myanmar language, EI model is trained and proposed. By trying the neural EI model to provide entity tools for Myanmar language develops Myanmar EI system. NLP research works for Myanmar language will be convenient the result from this Myanmar EI system.

## References

[1] H.M.Mo, K.M.Soe.“Syllable -Based Neural Named Entity Recognition for Myanmar Language”, International Journal on Natural Language Computing (IJNLC) vol.8, No.1, February 2019.

[2] P.C.J.Chiu ,and E. Nichols, “Named Entity Recognition with Bidirectional LSTM-CNNs”, Transactions of the Association for Computational Linguistics , vol. 4 (2016):pp.357-370,2016.

[3] Huanzhong Dun and Yan Zheng, “A Study on Features of the CRFs-based Chinese Named Entity Recognition”, International Journal of Advanced Intelligence. Volume 3, Number 2, July, 2011, pp.287- 294.

[4] Hesheng Xu and Bin Hu“Legal text recognition with LSTM-CRF deep learning Model”, vol. 2022, March 17, 2022.

[5] Z. Huang, W. Xu, and K. Yu. “Bidirectional LSTM-CRF Models for Sequence Tagging,” vol.1508.01991. 2015 Aug 9.

[6] D. Bonadiman, A. Severyn, and A. Moschitti, "Deep Neural Networks for Named Entity Recognition in Italian," CLiC it 51 (2015).

# Networking and Security

# Securing Critical Data Using Hybrid Cryptosystem

Koung Hsu Wai, Myat Thu Zar
*University of Computer Studies, Pyay*
khsuwai290@gmail.com, myatthuzar86@gmail.com

## Abstract

*Data security play an important role in every organization. Therefore, many organizations need to protect their information or data when transmission over the network. This system intends to implement the combination of the AES and RSA algorithm in order to more secure than single algorithm. PEC's Student marks file is sent from Pyay Education College to Naypyidaw office for examination result. Student's data are stored in excel files format and place in the computer. In this system and AES algorithm uses to encrypt the student's mark file, RSA algorithm uses to encrypt the AES's key. HMAC SHA-256 algorithm is applied to generate the hash value and to authenticate transferred information between the two sides (that share a secret key). Advantages of RSA algorithm and AES algorithm is combined in this system. The proposed system is implemented using PHP programming language. The experimental results show that hybrid AES-RSA cryptography system has been performed along with data integrity.*

**Keywords:** Cryptography, Data Encryption, AES RSA and HMAC

## 1. Introduction

Two types of encryptions such as symmetric and asymmetric key encryption is defined in crypto system [4]. In symmetric key cryptosystem, only single key is used for both encryption and decryption process. Asymmetric key cryptosystem has two keys, one for public key and one for private key.

To transfer message or data transmission, electronic mail (e-mail) is the medium over the network [5]. The combination of the benefits of AES algorithm encryption and decryption speed and advantages of key management of RSA algorithm [6].

## 2. Related Works

A. Guru, A. Ambhaikar explained about AES and RSA- based Hybrid Algorithm for message encryption & Decryption [1]. This paper utilizes the speed advantage of the AES algorithm in the encryption operation and the stability and key management advantage of the RSA algorithm, and incorporates the encryption power of both to encrypt the code. This paper suggests the implementation of file encryption of the AES and RSA hybrid encryption algorithm, and analyses their advantages and drawbacks.

B.Rana, S. Wankhade analyzed Hybrid Cryptographic Algorithm for Enhancing Security of text at International Conference on emanations in Modern Technology and Engineering, 2017 [2]. The hybrid model uses a combination of three symmetric algorithms AES, DES and IDEA. This paper presents the comparison of different parameter of result analysis for different.The proposed algorithm uses three different keys of different length for encryption and decryption process.

Salini Dev P V, A.P. Jose, J.Joseph, show that hybrid encryption algorithm for data transmission over public network [6]. This paper discusses the combination of 3DES and RSA algorithms and combination of AES and ABE algorithms. This paper compares the facts of DES, 3DES and AES. This paper presents the effective method to resolve the problem of safe transmission in Internet.

## 3. Background Theory

### 3.1. Advanced Encryption Standard (AES)

The Advanced Encryption Standard (AES) is the popular encryption algorithm in cryptographic system. This is start established by the National Institute of Standard and Technology (NIST) in 2001 [7]. There are two types of cipher in cryptography such as stream cipher and block

cipher. In block cipher encryption, plaintext messages or data are divided into fixed size blocks before transforming these messages into ciphertext using a key. AES is the symmetric block cipher algorithm with a block size of 128 bits. Three types of block ciphers are composed in AES such as AES-128, AES-192 and AES-256. This means AES-128 uses 128-bit key length, AES-192 uses 192-bit key length and AES-256 uses256-bit key length for encryption and decryption a block of message. In AES, data is encrypted and decrypted for each block in 128 bits using the key size of 128, 192 and 256 bits, respectively. The number of rounds depends on the key length as follows:

- 10 rounds for 128 bit key
- 12 rounds for 192 bit key
- 14 rounds for 256 bit key

Four main processes are composed in AES encryption such as Substitution byte, shift row, Mix-column and Addround key. These four operations are operated as inverse in decryption process. Data are operated on bytes rather than bits in AES. Therefore 128-bit block size of input data process 128 bits (16 bytes) at a time.

### i. Byte Substitution (Sub Bytes)

SubByte perform the substitution process. Each byte is substituted by another byte and this operation is made using a lookup table also called S-box. In this substitution a byte is never substituted by itself and do not substitute by another the compliment of the current byte. The result consists of 16-byte (4×4) matrix.

### ii. Shift Row

Left shift is made by shifting each of four rows of the matrix until all entries are reinserted on the right side of row. In this step, the shift process is done as follows:

- First row of the matrix is not shifted.
- One byte position of second row is shifted to the left.

Two position of third row is shifted to the left.

- Three position of fourth row is shifted to the left.
- The new matrix contain with the same 16 bytes.

### iii. Mix Columns

Matrix multiplication is applied to transform each byte of a column in order to get a new matrix value. The result contains new matrix with 16 new bytes. However, this step is not done in the last round.

### iv. Add Round Key

Although each matrix contains 16 bytes, these bytes are examined as 128 bits in this step. These bits are XOR with the round key 128 bits. The output of the last round is the ciphertext. This is described in Figure 1.

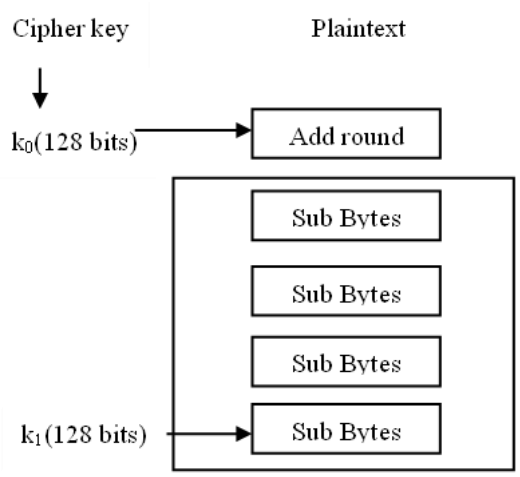



**Figure 1. Process of add round key**

## 3.2. Asymmetric Key Standard (RSA)

RSA algorithm is the asymmetric key cryptosystem with public and private key pair. RSA algorithm consists of three parts:

1. Public key and private key generation
2. Encryption process
3. Decryption process

### 3.2.1. The public key and private key generation

Steps of private key and public key generation for RSA algorithm are as follow:

- Two unequal large prime numbers p and q are randomly chosen.
- And then calculate n of p and q, $n= p \times q$
- $Ø(n) = (p-1)(q-1)$ is calculated by Euler function.
- Positive integer **e** is randomly selected between $1 < e < Ø(n)$ and gcd $(e, Ø(n)) = 1$.
- By applying the formula of $ed = 1 \bmod Ø(n)$, and then private key value d is obtained where $0 < d < n$.

- The public key **e** is get by applying the formula of PU = {e, N}.
- PR = {d, p, q}, the private key is saved, where d is the private key

#### 3.2.2. RSA Encryption

Public key is needed to encrypt the file at the sender side.

$$C = P^e \bmod n$$

Where, **C** is the ciphertext, **P** is the Plaintext, **e** is the public key and **n** is the modulus.

#### 3.2.3 RSA Decryption

At the receiver side, decryption need the private key related with the public key used for encryption.

$$P = C^d \bmod n$$

Where, **P** is the Plaintext, **C** is the ciphertext, **d** is the private key and **n** is used as the modulus.

### 3.3. HMAC SHA-256 Algorithm

Accurate and reliable data is very important for organization to make decision on these data. Therefore, user need to measure data integrity and authentication for this system. HMAC SHA-256 algorithm is used to provide data integrity for this proposed system. HMAC is a message authentication code get by processing a cryptographic has function SHA 256 over the data [3]. HMAC SHA-256 algorithm apply both data integrity and authentication due to the both use of key and hash function. The work of HMAC SHA-256 is shown in Figure 2.

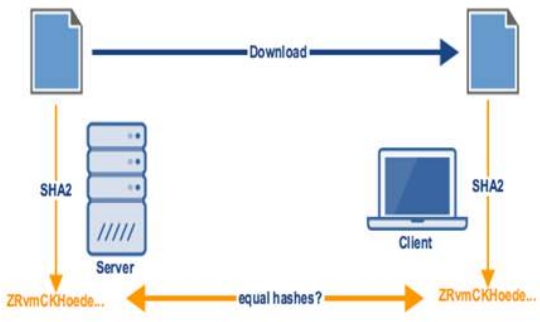



**Figure 2. Work of HMAC algorithm**

When calculating the hash value of HMAC, one type of cryptographic hash function may use such as SHA-2 or SHA-3 and so on. The output MAC algorithm is called HMAC-X, where X is the number of hash function used (e.g. HMAC-SHA256 or HMAC-SHA 512).

## 4. Overview of the Proposed System

PEC student's mark file is encrypted using AES algorithm. And then AES's Key is encrypted by using RSA's public key. HMAC SHA-256 algorithm is applied to generate the hash value. Hash value is used for authentication. The encrypted file, encrypted key and hash value are sent to the receiver. The receiver decrypts the encrypted key by using the RSA's private key. And then, the receiver also decrypts the encrypted file by using the AES's decryption algorithm. In the receiver side, SHA-256 algorithm is also applied to the result file to generate hash key value. And then compare this hash key value with the receive hash key value. If the key value is matched with the sender and receiver, this file is authenticated and if not, assume the receive file is not correct and the decryption process will be exit. Overview of the proposed system is shown in Figure 3.

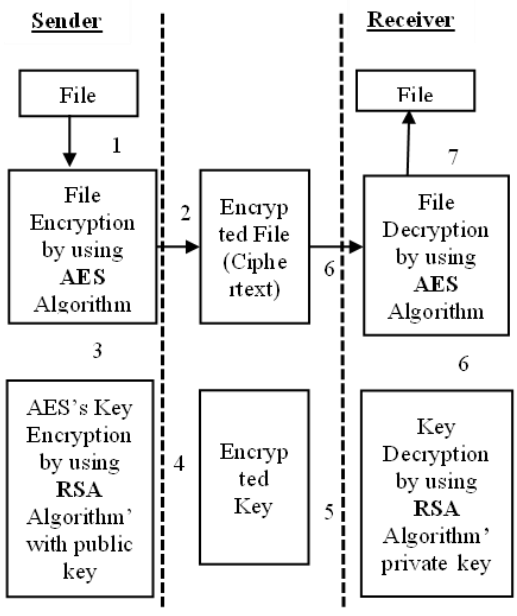



**Figure 3. Overview of the Proposed System**

### 4.1. Sender side of the Proposed System

In the sender side, the user choose the plaintext PEC student's mark file in order to

make encryption with AES key. These mark files are stored in computer drive with excel file. This process is shown in Figure 4. After selecting the file from computer drive, give the input key to encrypt the file. Input key must contain one lowercase letter, one capital letter, one number and at least minimum 8 character for one password creation.

AES algorithm have three types of key length size such as 128 bit, 192 bits and 256 bits. Therefore, user must choose one type of the key length size. After encryption, the input plaintext file is changed to encrypted value file. HMAC SHA-256 algorithm is used to generate the hash key value. Hash key value is used to verify data integrity and authenticate the data for this proposed system. In the sender side, after encrypting the PEC student's mark file with AES key, the system go with RSA algorithm to complete the encryption process. Encrypted file, encrypted key and hash value are sent to the receiver.

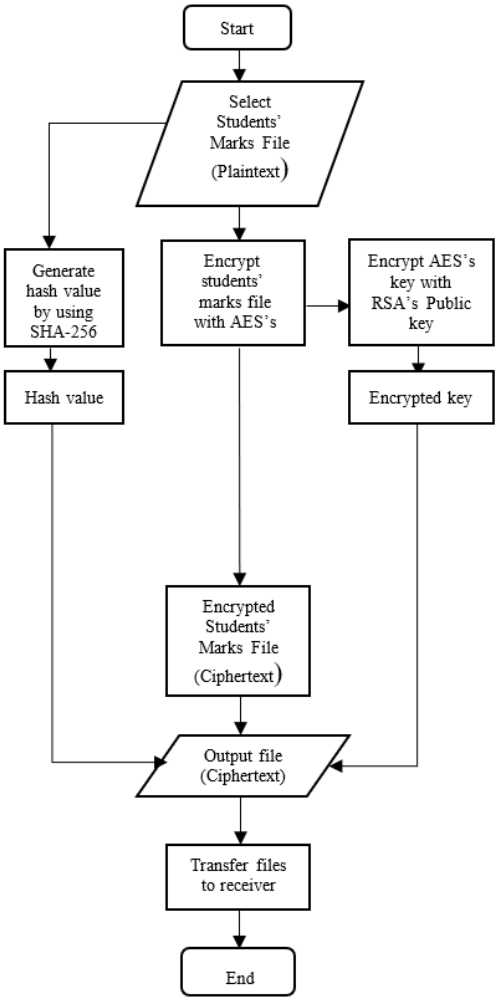



**Figure 4. Flow diagram of the sender side**

### 4.2. Receiver side of the Proposed System

Encrypted key, encrypted (ciphertext) and hash value are received at the receiver site. Encrypted FILE is decrypted with the RSA private key. And then the encrypted key is decrypted with AES key. Hash value is calculated with HMAC SHA-256 algorithm on the resulted plaintext file and compare with the receive hash value from the sender side. If the hash value is matched, the output file is validated and then proceed operation such as student result win list, fail list, credit list and so on. The process of decryption in receiver side is shown in Figure 5.

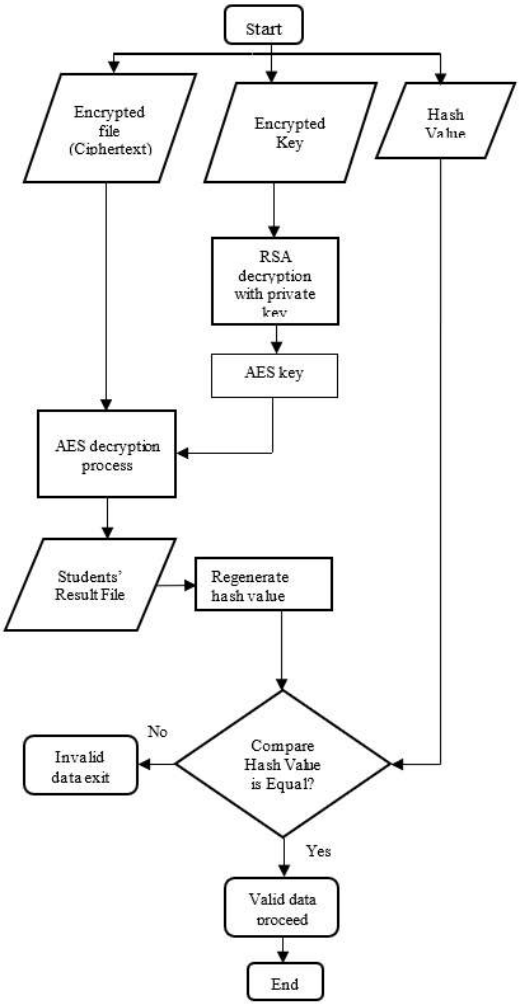



**Figure 5. Decryption process of receiver side**

## 5. Experiment and Result Analysis

In this proposed system, compare the AES, RSA and hybrid AES_RSA encryption and decryption time is made for the experimental analysis. Various length of key size such as 128 bit, 192 bit and 256 bit are applied in AES

encryption. By combining the advantages of RSA and AES algorithm, the system more secure and hoped to be applied more secure tool for the file. The experimental result for AES-128 bit is shown in Figure 6.

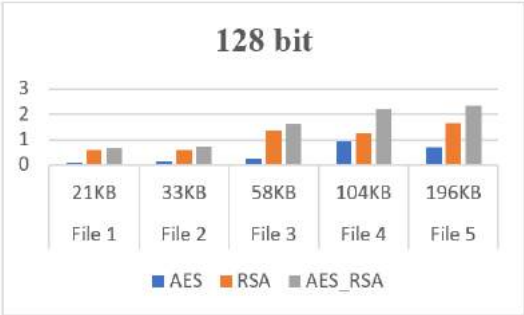



**Figure 6. Result with AES 128 bits**

The experimental result for AES-192 bit is shown in Figure 7.

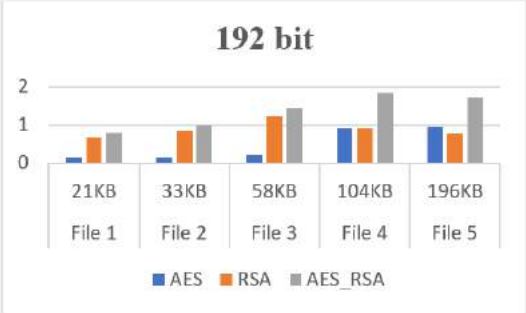



**Figure 7. Result with AES 192 bits**

The experimental result for AES-256 bit is shown in Figure 8.


**Figure 8. Result with AES-256 bits**

## 6. Conclusion

In this proposed system, there are three types of encryption have been cooperated in order to utilize the advantages of each one to build a high security system. Cryptographic encryption methods, one can prevent a third party from understanding transmitted raw data over unsecured channel during transmission. The purpose of creating this hybrid algorithm is to provide better security. The hybrid algorithm provides more security than individual algorithm for the plaintext. By using hybrid cryptographic approaches, cryptographic goals such as confidentiality, integrity and authenticity can achieve. This system provides safe mechanism for data transmission over the network. This mechanism provides dual protection by taking the advantages of the algorithms used, so the data transmission in the network will be more secure. The proposed system can be applied to other educational organization not only to University infrastructure but also Basic Education High School (BEHS).

## References

[1] Abhishek Guru and Asha Ambhaikar , " AES and RSA based Hybrid Encryption Algorithm for Message Encryption and Decryption", IT in Industry, March,2021.

[2] B.Rana, S. Wankhade, "Hybrid Cryptographic Algorithm for Enhancing Security of Text", International Conference on Emanations in Modern Technology and Engineering, 2017.

[3] E. S. Ibrahim Harba, "Secure Data Encryption Through a Combination of AES, RSA and HMAC", Engineering, Technology & Applied Science Research, 2017.

[4] G.V.S Pavan Mallik1, Y Saranya Bala2, "Securing Email using Hybrid Encryption System", International Research Journal of Engineering and Technology (IRJET), July 2020.

[5] P. Prajapati, N. Patel, R. Macwan, N. Kachhiya, P. Shah, "Comparative Analysis of DES, AES, RSA Encryption Algorithms", International Journal of Engineering and Management Research, Vol. 4, No. 1, pp. 292-294, 2014.

[6] Salini Dev P V, A.P. Jose, J.Joseph, "Hybrid Encryption Algorithm for Data Transmission over public network", IJARIIE-ISSN(O)-2395-4396, 2017.

[7] T.Monoth and N. Francis, "An Analysis of Hybrid Cryptographic Approaches for Information Security", International Journal of Applied Engineering Research, 2018.

# Secure Messaging System Using RC4-2S

Hnin Hsu Hlaing, Cho Cho San
*University of Computer Studies, Yangon*
*hsulaychocolate402@gmail.com, chochosan@ucsy.edu.mm*

## Abstract

*Nowadays, telecommunication technologies are developing rapidly and many organizations are mostly using these technologies. The well-known telecommunication technology is the short message service (SMS). It plays a vital role in the business area such as mobile banking, organizational marketing system, etc. A simple SMS system doesn't have the built-in procedure to offer data security and attackers can easily intercept it. This problem can be solved by using the encrypted message. This system developed a secure SMS application for an android smartphone with an RC4-2S algorithm for SMS data confidentiality on mobile networks. For the performance evaluation, the proposed system tested the experimental result on NIST statistical test. And the proposed system also provides confusion and diffusion properties of the experiment.*

**Keywords:** SMS, Cryptography, Randomness, Encryption, Decryption

## 1. Introduction

Short Message Service (SMS) is the top application in the telecommunication technology environment. SMS is a communication method between mobile phones or personal computers. An SMS message can accept only 160 characters as it maximum size. A simple SMS communication system over the GSM networks is not secure, thus confidential messages of the SMS can easily intercept by unauthorized parties. Therefore, the security in SMS is very important. To improve the security, a simple SMS system needs to combine with the cryptosystem. By using a cryptosystem, the message that the sender is encrypted into the cipher text. If the attackers are intercepted the SMS system, they can't change the cipher text to the plain text without using the correct key. Therefore, the proposed system developed the mobile phone application to encrypt the SMS message into cipher text as a result the content of the SMS message can't know by unauthorized parties. When secret information is sent by using the original SMS system, it is not safe and secure. To solve these problems, a secure messaging service has been proposed in this system by encoding the plain text message.

## 2. Related Work

The author in [5] developed the android offline application using the RC4 algorithm. This system is focused on the security of short message service using the RC4 stream cipher algorithm. This system needs two applications, one for the sender application and another for the receiver application. And the users type the encryption key manually.

The authors in [10] also proposed the SMS android application using the RC4 algorithm. The author claimed that the encryption and decryption time depended on the characters, the number of the SMS message, and the key along with the smartphone specification. The authors mentioned that the correlation value was only affected by the above three characters.

The author in [8] proposed a secure SMS system by using the hybrid architecture with RC4 and Affine Cipher. RC4 key stream is used to encrypt messages and the Affine cipher algorithm is used to store the key of the RC4.

This section describes the recent related work for a secure messaging system. The proposed system uses RC4-2S to implement the secure messaging application.

## 3. Background Theory

This section describes the cyber security goals, cryptography, and NIST two tests for the randomness key-stream.

### 3.1. Information security Goals

The information security goal is to defend the data from intercepting by unauthorized people. Three types of security goals are confidentiality, integrity, and availability [1].

**Data confidentiality** is to protect the information disclosure from unauthorized parties. It is provided to access the data if the user is authorized. Otherwise, the data needs to be protected from unauthorized users.

**Data integrity** is ensuring the authenticity of the data or information that does not alter between sender and receiver. Data integrity is the availability of data created by the source that does not alter by an unauthorized person in an unauthorized way.

**Data Availability** concerns both the accessibility and continuity of information. Information needs to be available to an authorized entity when it is needed.

### 3.2. Cryptography

Cryptography is the art of protecting data security by turning it into another form that is human unreadable form. It can provide secure communication by preventing malicious threads, the attackers can't access the information. Cryptographic systems change the plain text into cipher text by applying an encryption algorithm when transmitting the data. Cryptography provides three techniques: Symmetric Key Encryption, Asymmetric Key Encryption, and Hashing Algorithm.

**Symmetric key encryption** can be defined as a secure algorithm. This used only one key for both the encryption and decryption process. The symmetric key algorithm can be considered very secure. RC4 and RC4-2S are symmetric key algorithms [4].

**Asymmetric key encryption** algorithm is also known as the public key encryption algorithm. It uses two keys; one for public and another for private. The private key cannot be calculated through the use of public-key because they are related cryptographically. Rivest Shamir Adleman (RSA), Digital Signature Standard (DSS), and Diffie-Hellman exchange methods are well-known asymmetric key algorithms [4].

**The hashing algorithm** is also known as a hash function. A hash function can take an arbitrary amount of input data, and produce the fixed-length cipher text by applying a mathematical formula. Message-digest algorithm (MD5), and Secure Hashing Algorithm (SHA) are well-known algorithms.

## 3.3 RC4-2S (Rivest Cipher 4 with two state tables)

RC4 is one of the most symmetric stream cipher algorithms in real-time security. Variable key length can be used between 1 and 256 bits. However, the minimum number of the key length must be 128 bits to keep the security of the data. RC4 stream cipher is widely used because the encryption time is faster than the other stream cipher algorithm. This has two phases: Key Scheduling Algorithm (KSA) and Pseudo-Random Generation Algorithm (PRGA) [6]. RC4 algorithm has some weaknesses including a correlation problem.

RC4-2S is an updated version of the original RC4 to improve data security. The generated RC4-2S's key-stream is more random than the RC4. RC4-2S also contains the KSA phase and PRGA phase.

**KSA** phase of the RC4-2S, swap elements by using two arrays. Figure 3.1 describe the process of the KSA stage. The first array $S_1$ is filled with the elements from 0 to 127and the second array is filled with the remaining elements 128 to 255. The initial key-value K is set with another array. After swapping the operations, the generated $S_1$ and $S_2$ become two secret random inputs for seconds phase PRGA [6].

```
Input k, m
Output: S1, S2
For i from 0 to N/2-1
        S1 [i] = i
For i from N/2 to N-1
        S2 [i-N/2] = i
j = 0
For i from 0 to N/2 -1
        j = (j + S2 [i] + k [i mod l]) mod N/2
        Swap S1 [i] with S1 [j]
j =0
For i from 0 to N/2 -1
        j = (j+ S2 [i] + k [i mod l]) mod N/2
        Swap S2 [i] with S2 [j]
Return (S1, S2)
```

**Figure 3.1. Key Scheduling Algorithm**

In the **PRGA** phase, swap the elements of $S_1$ and $S_2$ by using the three-pointers. Each loop cycle produces two keys. The encryption process is done that creates the cipher text by XOR-ing the plaintext and the generated key-stream [6]. Figure 3.2 describes the process of the PRGA phase.

The key generation time of the RC4-2S is faster than the original RC4 because it generates a pair of key-stream after taking two swaps and five modulo functions. This takes a loop cycle until half of the plain text sequence is reached. RC4 can generate only one key stream after one swapping and taking three modulo functions.

```
Input: S1, S2
Output: key Sequence Kseq
i, j1, j2 =0
While not end of half sequence Do
        i = (i +1) mod N/2
        j1 = (j1 + S1 [i]) mod N/2
        Swap S1 [i] with S2 [j1]
        t1 = S1 [(S1 [i] + S1 [j1])] mod N/2
        j2 = j2 + S2 [i] mod N/2
        Swap (S2 [S2 [i] + S2 [j2]) mod N/2
        Kseq = [t1, t2]
Return Kseq
```

**Figure 3.2. Pseudo-Random Generation Algorithm**

## 3.4. Firebase

Firebase is an app development platform developed by Google that enables developers to develop mobile applications and web applications. This platform offers a number of services: Analysis, Authentication, Cloud messaging, Real-time database, Crashlytics, Performance, and Test lab. This proposed system is used an authentication feature and a real-time database feature.

**Firebase Authentication Feature** provides backend services and is easy to use with support authentication using the password, phone numbers, Google, and more. This feature can provide the user registration and login process and then this also handles sending the password reset email.

**Firebase Real-time Database feature** is a cloud-hosted database. Data is stored in JSON format and synchronized in real-time to every connected client. All of the clients share the real-time database instance and automatically receive updates with the newest data.

## 3.5. NIST Randomness Test

The quality of the data confidentiality relies on the randomness of pseudo randomness numbers. The binary generated sequences are tested using the National Institute of Standard and Technology (NIST) test suite, which is a statistical package for random number generation tests. There are 15 NIST statistical tests. This proposed system is tested two NIST test. Among the 15 NIST statistical tests, frequency test and run test are the fastest because they process each bit of the bit stream. All subsequence tests are depended on the passing of these two tests.

The first is **Frequency Monobit Test [2],** the focus of the test is to decide whether the number of zeros and one arrangement are just about equal and would be predictable for an actual sequence.

$$P_{value} = erfc\left(\frac{S_{obs}}{\sqrt{2}}\right) \quad (1)$$

The second is **Run Test [2],** the focus of this test is to define whether the oscillation between such zeros and one is too fast or slow.

$$P_{value} = erfc\left(\frac{V_{n(obs)-2n\pi(1-\pi)}}{\sqrt[2]{2n}\,\pi(1-\pi)}\right) \quad (2)$$

If the computed P-value is less than 0.01, the sequence is not random. Otherwise, assume the sequence is random.

## 3.6. Confusion and Diffusion

Confusion and diffusion properties are important to secure the message in cryptography. Both Confusion and diffusion are used to prevent the encryption key from its deduction or ultimately to prevent the original message [9].

**Confusion** hides the relationship between the cipher text and the key but they do not depend on each other. If one bit of the key is modified, the generated cipher text also has many changes this is the confusion property.

**Diffusion** hides the relationship between the pain text and the cipher text. That is if one bit of the plain text is modified, the cipher text also has many changes.

## 4. Proposed System

This system develops an online chat application by using the end-to-end encryption mechanism therefore, the user need to access the internet. The user can accept this application anywhere and anytime over the internet. The messages are directly sent to the receiver and use the Firebase real-time database feature instead of short message service center therefore, the user do not need to pay the sending SMS fees like a normal SMS system but the user need the SIM for user registration phase. Simple SMS systems is not secure for messaging confidential data. Thus, the RC4-2S encryption algorithm has been applied to ensure the message security in the original SMS service. Figure 4.1 is described about the flow diagram of the proposed system. The proposed system contains four phases: User registration and Login Authentication, Key Stream Generation, Encrypt Message, and Decrypt Message.

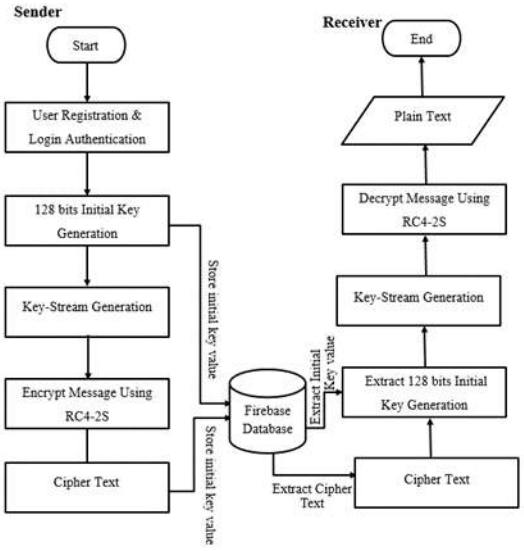



**Figure 4.1 Proposed System Flow Diagram**

### 4.1. User Registration and login Authentication

When the user registers this app, the user must not leave any empty text box, fill with your email address, password, and the phone number that starts with "09" and does not exceed 11 numbers. And then the password must have at least 6 characters.

### 4.2. Key Stream Generation

The initial secret key is automatically generated by the system that has 128 bits. This secret key and the plain text are the input for the KSA phase and take many times swapping operations based on the secret key. After the operation is successfully complete, two arrays are generated. PRGA phase is used these two arrays as input. Then the permutation and swapping operations are performed based on the three-pointers. After one loop cycle that generates a pair of keys. The random keystream is generated after taking loop cycle until the half of the plain text sequence.

### 4.3. Encrypt Message

When the sender send the message, the encryption process of the plain text is performed by system internally. The plain text message is encrypted by XOR-ing with the generated key-stream. The cipher text is stored at the firebase database and forward these cipher text to the receiver.

### 4.4. Decrypt Message

At the receiver side, the receiver receives the cipher text. When the user tap this cipher text, the decryption process are performed internally. Cipher text and initial secret are extracted from the database and perform the KSA phase and PRGA phase based on these two inputs. After the permutation processes are successfully complete that generate the plain text.

### 4.5 Experiment Result and Performance Evaluation

The following Figure 4.2 are the experimental result of the proposed system. The first figure shows the message at sender side. The second figure shows the received ciphertext at the receiver side. The encryption and decryption processes are performed at the background internally.

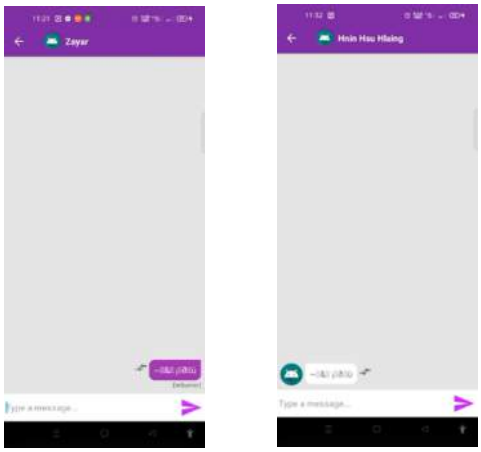


**Figure 4.2 User Interface of the Application**

**Table 4.1 Performance Evaluation Result**

| | Experiment 1 | Experiment 2 |
|---|---|---|
| Plain Text | Master of Computer Science (M.C.Sc) | Master of Computer Science (M.C.sc) |
| Initial Key | ZhW%iwlVN EAwlST8 | WZOmtwr#C5 copZlc |
| Cipher Text | N•M™$Zå®3‹ ÖÏ%ÅU¬Iô9• ¼}ÆW«2ÉQì $ | ž 5¦à- ðLïA¶;ë{_v,E Þ Û?ñ+• 9¹<mw ‰>c,,’ |
| Frequency Test | 0.9803693443 109074 | 0.8663855960 81871 |
| Run Test | 0.9799999999 904022 | 0.9910872973 90982 |

The proposed system provides two types of evaluation results. These tests used the plain text with a small change but the generated cipher text has many changes that provide the diffusion properties. The generated cipher text is not related to the key thus the system also provides the confusion properties. Based on the NIST test results, the generated key stream can be considered randomization. Therefore, the system provides a secure messaging application for the users than simple SMS.

## 5. Conclusion

SMS data security has become critical, and access control plays an important role in data security. The proposed system is used RC4-2S algorithm to improve the security of SMS. This system proposed online chat application therefore the user can access this application anywhere the can access the internet. The user can use any language (English or Myanmar). The proposed system stored the key and cipher text in the firebase real-time database so it is secure until the firebase database is not breakdown. The secret key distribution has not been performed in this system. Therefore, the key distribution will conduct in future work.

## References

[1] A.Kahate, 2013. "Cryptography and network security"

[2] A.Mousa and A.Hamad, 2006, "Evaluation of the RC4 algorithm for data encryption"

[3] A.Rukhin, J.Soto, J.Nechvatal, M.Smid and E.Barker,2001." A statistical test suite for random and pseudorandom number generators for cryptographic applications "

[4] E.N.Ekwonwune and V.C.Enyinnaya, 2020, "Design and Implementation of End to End Encrypted Short Message Service (SMS) Using Hybrid Cipher Algorithm"

[5] K.H.Myint, 2019. “A Data Confidentiality Approach to Short Message Service (SMS) on Android”

[6] M.M.Hammood, K.Yoshigoe and A.M.Sagheer,2013."RC4-2S: RC4 stream cipher with two state tables”

[7] M.S.Novelan, A.M.Husein, M.Harahap and S.Aisyah, 2018. "Sms security system on mobile devices using tiny encryption algorithm".

[8] O.S.Sitompul, N.H.Pasaribu and E.B.Nababan, 2018."Hybrid RC4 and Affine Ciphers to Secure Short Message Service on Android."

[9] P.Praveenkumar,R.Amirtharajan, K.Thenmozhi and J.B.B.Rayappan, 2017."Fusion of confusion and diffusion:a novel image encryption approach"

[10]R.Rifki, A.Septiarini and H.R.Hatta, 2018,"Cryptography using random Rc4 stream cipher on SMS for android-based smartphones"

# Security Control By Ticket-based Address Resolution Protocol

Chit Hnin Wai, Si Si Mar Win
*University of Computer Studies, Yangon*
*chithninwaiucst@gmail.com, sisimarwim@ucsy.edu.mm*

## Abstract

*Nowadays, communication is becoming more and more important to keep in touch with family and friends. Computer networks play a key role in this process. To make this process facilitating, Network engineers have used protocols for exchanging messages between computers. Many protocols are optimized to simplify the process of initializing these sites. However, it is still need to take security in some areas. This paper presents some of the vulnerability that exists in the Address Resolution Protocol (ARP) protocol and implements the Ticket based Address Resolution Protocol (TARP) by creating some spoofing attacks such as Man-in-the-Middle (MITM) attack and DoS attack to deceive a victim's machine and a router for exploiting the weaknesses of ARP protocol. In the experiments, TARP is implemented for ARP spoofing by distributing centrally secured mapping of MAC/IP address through existing ARP messages. This system explored some of operational vulnerability related with ARP security of deploying and administering. In this system, window operating system is chosen as to implement the attack as well as the defense creation.*

**Keywords**: LAN, ARP, DoS, Man-in-the-middle, TARP

## 1. Introduction

IP over Ethernet networks are the most well-known Local Area Networks (LANs) these days. They used the Address Resolution Protocol, ARP to determine IP addresses into equipment, or MAC (Medium Access Controllers), addresses [12]. Every one of the hosts in the LAN backup a reserve of settled addresses. ARP goal is conjured when another IP address must be settled or on the other hand a passage in the store lapses. The ARP harming assault comprises of noxiously adjusting the relationship between an IP address and its comparing MAC address. Different devices accessible on the Internet permit to play out the refined ARP harming assault [11].

Albeit this is the most famous rendition, ARP harming isn't associated with Ethernet organizations. Layer 2 exchanged LANs, 802.11b organizations, and cryptographically safeguarded associations are additionally helpless. In [3], different situations are portrayed where a remote aggressor harms two wired casualties, a remote casualty and a wired one, or two remote casualties, either through various passages or a solitary one. With respect to cryptographically safeguarded networks, the utilization of cryptography at network layer, e.g., through Secure Sockets Layer (SSL) or Secure Shell (SSH) [4], doesn't safeguard against.

ARP harming is an assault performed at the layer beneath. By performing ARP harming, an assailant powers a host to send bundles to a MAC address not quite the same as the one of the expected objective, which might permit her to listen in on the correspondence, change its substance (e.g., modifying it, infusing orders or malignant code), commander the association. Besides, when performed on two unique hosts simultaneously, ARP harming empowers a foe to send off a man in the center. In MITM assaults, traffic between two hosts is diverted through a third one, which goes about as the man in the center, without the two knowing it. The MITM may essentially transfer the traffic subsequent to reviewing it or adjust it prior to resending it.

MITM assaults are conceivable at different OSI stack layers. ARP harming permits to perform such an assault at information connect layer. At network layer, the assault takes advantage of DNS harming [5]. The aggressor first alters the DNS tables to relate its own IP address with the emblematic names of both casualty has. Hence, when the casualties will inquiry the DNS requesting the each other's IP

address, they will get the aggressor's IP address. Right now, all the traffic between the two hosts will initially be gotten by the assailant that will advance it to the particular objective, after potentially adjusting it.

This work implements the Ticket-based Address Resolution Protocol (TARP) convention. Covering executes security by circulating midway produced MAC/IP address planning validations. By sending tickets to clients, they are able to join the organization and are thusly circulated through existing ARP messages. Covering is a plausible methodology for the different grouping of existing organization proficient gadgets. We give an itemized depiction of the convention plan and its execution inside the window working framework.

## 2. Related Work

A few endeavors have been made to address the above security issues through strategies outer to the ARP convention. Tables of ARP have statically arranged as an example [1]. Obviously, this would cause an enormous managerial above and is to a great extent immovable for dynamic conditions. Alternately, the port security [11] highlights accessible in ongoing switches limit the utilization of actual ports to designed MAC addresses. This approach just forestalls specific sorts of MAC capturing, however never really forestalls MITM assaults. Thus, it is just a fractional (and in numerous ways restricted) arrangement.

Different arrangements endeavor to distinguish trouble making, as opposed to forestall it. ARP Watch [2], an organization level discovery gadget, distinguishes noxious ARP parcels by checking MAC/IP address pairings happening on a subnet. Alternately, have level location administrations contrast in that each host on the organization endeavors to distinguish malignant messages showing up at the nearby connection point [8]. This is accomplished by distinguishing copy as well as spontaneous ARP bundles. Identification methods are reformatory by definition, and consequently are of restricted utility in numerous conditions.

Various cryptographic conventions have designated issues in the ARP security. In the Secure Link Layer (SLL), everything join layer traffic is verified and scrambled. While this keeps approved has from infusing vindictive messages, it doesn't forestall approved, yet deceitful hosts, from infusing vindictive messages. In one more methodology, the Secure Address Resolution Protocol was proposed by the authors in [6]. A solid server in this convention imparts secret keys to every host on a subnet. The mappings of address that is refreshed intermittently through correspondence with each host. All ARP demands and answers happen between a host and the server, and answers are confirmed utilizing the common pair keys. Note that the server addresses a particular weak spot and blockage, which make it an unfortunate counterpart for most organizations. IP addresses are utilized as open keys in ABK. Nonetheless, contemporary character based frameworks require at least one heavyweight cryptographic activities for each signature or approval. Thus, their expense is restrictive for the majority asset unfortunate gadgets.

The most well-known ARP security convention, S-ARP [10], likewise utilizes hilter kilter cryptography. In any case, dissimilar to ABK, has utilize self-made public/confidential key pair confirmed by a nearby confided in party. S-ARP demands continue as ordinary ARP demands. Nonetheless, S-ARP answers are endorsed by the source's confidential key. After getting an answer, the mark is checked utilizing the shipper's public key. In the event that the collector doesn't have the shipper's public key, or on the other hand on the off chance that the mark can't be confirmed by the keys presently in its AKD’s key ring. AKD sends it to the mentioning host in a marked message. Assuming the new open key confirms the signature, the answer is acknowledged and the public key is stored; in any case, it is dismissed. To stay away from replay assaults, messages are time-stepped and synchronization messages are traded with the AKD.

At least, SARP expected a solitary mark age and confirmation per address goal. This cost can be huge. None of these arrangements at the same time address both the similarity and execution necessities of current organizations. As we will show in the accompanying segment, TARP effectively accomplishes strength to reserve harming and similarity with ARP, at essentially no expense.

## 3. Background Theory

ARP reserve harming is the procedure by which an aggressor perniciously changes the planning of an IP address to its comparing MAC address in the ARP store of one more host by sending satirize ARP answer. So this method is likewise called ARP mocking. In Figure 1, the assailant is Host C. It executes the ARP Cache Poisoning assault by sending a mock ARP answer to Host A that 'IP address of Host B guides to MAC address of Host C' and a satirize ARP answer to Host B saying that 'IP address of Host A guides to the MAC address of Host C'. ARP is a stateless convention and answers are not checked against forthcoming solicitations. Consequently Host A and Host B will refresh their ARP reserve with the planning got in the ARP answers. ARP reserve harming is an assault having areas of strength for an in LANs (i.e., all hosts associated with a similar organization) that of a pernicious host is powerless against this assault.

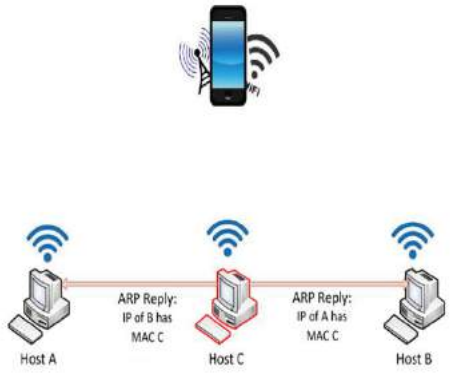



**Figure 1. Host C performing ARP cache poisoning attack on Host A and Host B**

### 3.1. Man-in-the-Middle (MITM) Attack

MITM is the attack that the current conversation or data transfer is intercepted by the attackers. Figure 2 shows how MITM can take in the network traffics. When the ARP reserves of Host A and Host B are harmed, all the traffic bound for Host B will be sent by Host A, to Host C. Likewise Host A will be sent all traffic bound by Host B, to Host C. Now Host C can peruse all the traffic between Host A and Host B.


**Figure 2. Man-in-the-center (MITM) Attack**

In the event that Host C advances the parcels, in the wake of understanding them, to the real objective machine, then Host A and Host B won't actually recognize that they are being gone after. This is a Man-in-the-Middle assault by which the aggressor can redirect the traffic passing between two machines to pass through him. In the above Figure, the assailant is Host C. Have C can redirect the traffic passing between two machines to pass by means of him.

### 3.2. Denial-of-Service (DoS) Attack

A Denial-of-Service attack is an endeavor to make a PC asset inaccessible to its expected clients.

It for the most part comprises of the coordinated endeavors of an individual, or numerous individuals to proficiently keep the help from working. It is somewhat not quite the same as MITM assault, when the aggressor doesn't advance the bundles, in the wake of understanding them, to the genuine objective machine. In the accompanying figure, the assailant is Host C. Have C doesn't advance the parcels, in the wake of understanding them, to the real objective machine.


**Figure 3. Denial-of-Service (DoS) attack**

## 4. A Ticket-Based Approach

The significant imperfection in ARP is the absence of message validation. Until the end of this paper, we arrange ARP weaknesses as tending to be categorized as one of the two following classifications:

- answer satirizing: manufacturing an ARP answer to infuse another location relationship into the casualty's store
- section harming: manufacturing an ARP answer to supplant a location relationship in the casualty's store

This framework address these weaknesses through the Ticket-based Address Resolution Protocol (TARP). Covering executes security by disseminating midway created validations [5]. Tickets validate the relationship among MAC and IP tends to through explanations endorsed by the neighborhood Local Ticket Agent (LTA). Each ticket encodes a legitimacy period as a termination time. Obviously, the utilization of lapse times accept a few type of free clock synchronization between the backer LTA and the approving clients. Such synchronization is a typical necessity for numerous conventions, and gadgets for its authorization are notable. To safely perform address goal utilizing TARP, a host communicates an ARP demand. The host with the mentioned IP address sends an answer, joining recently got ticket. The mark on the ticket demonstrates that the LTA gave it, i.e., the MAC to IP address planning is legitimate. The mentioning host gets the ticket, approving it with the LTA's public key. Assuming the mark is legitimate, location affiliation is acknowledged; in any case, it is disregarded.

### 4.1. The TARP Protocol

The implication by which a ticket is made and disseminated subject to whether the IP address tasks are static or dynamic. Shown in Figure 4, at whatever point a host is added to a static task organization, it is designed with the organization public key, an IP address, and a ticket. Since the affiliations are probably not going to change as often as possible, setting long ticket lifetimes might be satisfactory. In any case, there are security, execution, and authoritative contemplations connected with the determination of ticket lifetimes.

In IP organizations, has are relegated IP locations and design boundaries by a setup server. Each host gets a rent on an IP address and sends a recharging demand upon lapse. As needs be, the ticket terminates alongside the IP rent. Note that tickets are by definition public, in this manner a solid correspondence channel is pointless. Having the DHCP server assume the part of LTA dispenses with the requirement for extra ticket conveyance messages, consequently keeping up with straightforwardness of convention plan. A host requires the public key of LTA to confirm tickets. Key conveyance is generally secure whenever performed out of band. While less secure, this appropriation could likewise be performed through declaration and client acknowledgment, like that in the Secure Shell (SSH) convention [9]. Tragically, this permits an enemy new strategies for assault. For this paper, we just think about manual dissemination of the public key of LTA.

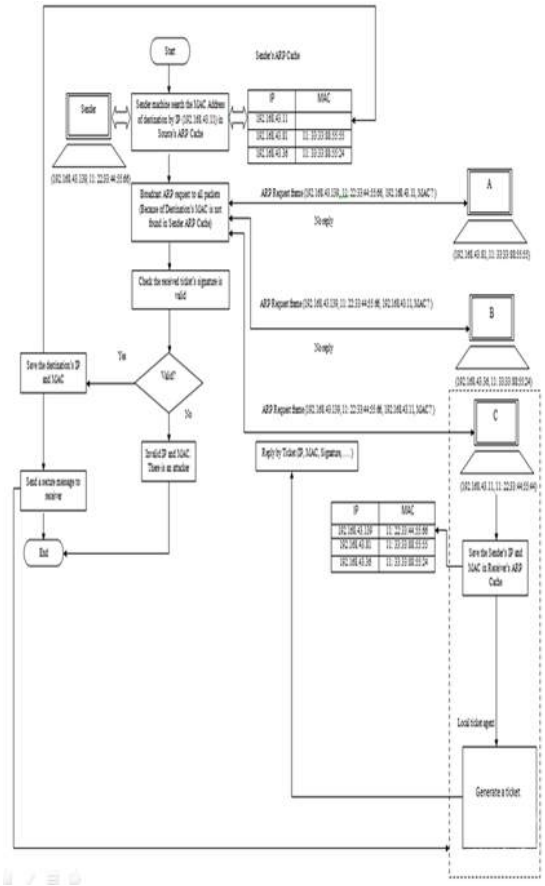



**Figure 4. The System Flow Diagram**

The activity of the ticket goal continues indistinguishable in the static as well as the dynamic cases whenever tickets have been conveyed to each host. Canvas message stream is like the ARP, with the special case that the ticket is attached to each answers as characterized in the first segment.

The flow of message in TARP is the same with ARP but the ticket is added to each replies. The flow of ARP request and the reply for the request with ticket is shown in Table 1. The format of ticket appended in the reply is described in the next section.

**Table 1. ARP Request and Reply with Ticket**

| Message Type | Sender IP | Sender MAC | Receiver IP | Receiver MAC |
|---|---|---|---|---|
| ARP Request (Sender→Receiver) | 192.168.43.11 | AA:BB:CC:11:22:33 | 192.168.43.22 | ? |
| ARP Reply (Receiver →Sender) | 192.168.43.11 | AA:BB:CC:11:22:33 | 192.168.43.22 | (AA:22:BB:33:DD:EF, With Secure Ticket) |

### 4.2. Ticket Format

Keeping up with in reverse similarity with ARP is vital for the reception of any improved location goal convention. Similarity is accomplished by coordinating the ticket into the ARP answer; no progressions need happen to the solicitation. As displayed in Figure 5, the ticket is affixed as a variable length payload, with the ticket header changed likewise.

The Magic field of the ticket header is utilized to answer recognize the new answer from an ARP. In the event that it is a TARP answer, 0x789a0102 is set to the enchanted field. Since TARP has just a single message type, the Type field really assigns the cryptographic algorithm. SigLen demonstrated the mark length. The leftover fields contain data expected to guarantee legitimate activity. The MAC as well as the IP Addresses make the location affiliation. The ticket legitimate is called the expiration time. Issue Timestamp shows when the ticket was produced and is utilized for ticket disavowal.

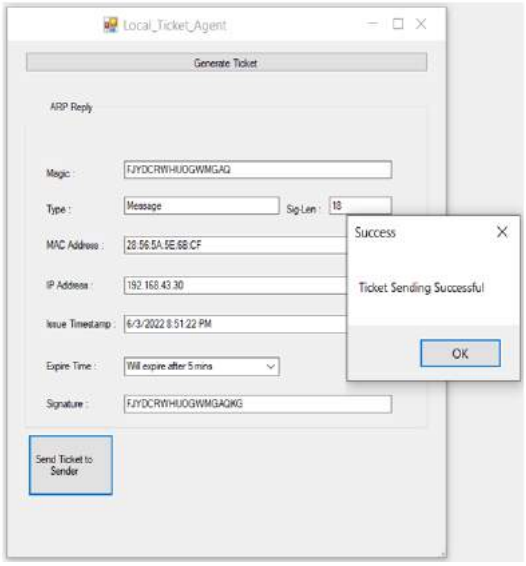



**Figure 5. Ticket Format**

## 5. Conclusion

Remote organizations have turned into a basic piece of the present organizations. The simplicity of arrangement, minimal expense, portability and high information rates have contributed altogether to their prominence. The mode of information transmission in remote organizations makes them intrinsically less secure than wired networks. For remote organizations to get to the Internet they should be associated with a wired organization by means of an Access Point or a remote switch. This has driven remote organization hardware makers to execute remote Access Points and remote switches with an implicit switch for wired clients and a WiFi passage for remote clients. The set up inside the gear is with the end goal that the wired and remote organizations are inside connected together to such an extent that they are in a solitary Local Area Network (LAN). This blend of wired and remote organizations represents another class of assaults on wired networks by means of uncertain remote LANs. One such class of assault is the Address Resolution Protocol (ARP) Cache Poisoning assault. Contingent upon the remote LAN set-up, beforehand secure wired organizations might become defenseless against assaults from remote clients in a similar LAN as the wired client.

This system implemented TARP for network security. Worked as an augmentation to ARP, TARP accomplishes flexibility to reserve harming. Canvas decreases cost by as much as

two significant degrees over existing conventions (e.g., SARP, ESARP).

### 5.1. Limitation and Further Extension

ARP weaknesses will stay a serious organization security issue until a good option is acknowledged. This system is demonstrated TARP to be more reliable, yet much work stays before our execution can be comprehensively utilized. Expansions including support for dynamic conditions are imperative. At long last, we look for additional functional experience; a more profound comprehension of the expenses and impediments of the methodology must be gathered from field testing.

## References


[1] Anatomy of an arp harming assault. http://www.watchguard.com/infocenter/article/135324.asp, got to June 2015.

[2] C. Adams and R. Zuccherato. A General, FlexibleApproach to Certificate Revocation, June 1998.http://www.entrust.com/securityzone/whitepapers.htm.

[3] W. Aiello, J. Ioannidis, and P. McDaniel. Beginning Authentication in Interdomain Routing. In Proceedings of tenth ACM Conference on Computer and Communications Security, pages 165-178. ACM, October 2013. Washington, DC.

[4] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. RFC 4034, Resource Records for the DNS Security Extensions. Web Engineering Task Force, March 2015.

[5] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. RFC 4035, Protocol Modifications for the DNS Security Extensions. Web Engineering Task Force, March 2005.

[6] S. M. Bellovin. Security issues in the tcp/ip convention suite. PC Communications Review, 2(19):32-48, April 2009.

[7] S. M. Bellovin. A glance back at "security issues in the tcp/ip convention suite". In twentieth Annual Computer Security Application Conference (ACSAC), pages 229-249, December 2014.

[8] D. Bruschi, A. Orgnaghi, and E. Rosti. S-arp: a protected location goal convention. 2013.

[9] Cisco Systems. Impetus 4500 Series Switch Cisco IOS Software Configuration Guide, 12.1(19) EW.

[10] R. Droms. Dynamic host setup convention. RFC 2131, March 2017.

[11] R. Droms and W. Arbaugh. Validation for dhcp messages. RFC 3118, June 2011. http://www.ietf.org/rfc/rfc3118.txt?number=3118.

[12] B. Bit and J. Dimov. Remote passages and arp harming: Wireless weaknesses that uncover the wired organization. http://downloads.securityfocus.com/library/arppoison.pdf.

[13] J. Galvin. Public Key Distribution with Secure DNS. In Proceedings of the sixth USENIX Security Symposium, pages 161-170, July 2016.

[14] M. Gouda. furthermore, C. Huang. A solid location goal convention. PC Networks, 41:860-921, January 2013.

[15] C. A. Gunter and T. Jim. Summed up authentication denial. In POPL '00: Proceedings of the 27th ACM SIGPLANSIGACT discussion on Principles of programming dialects, pages 316-329, New York, NY, USA, 2010. ACM Press.

[16] R. Housley, W. Portage, W. Polk, and D. Solo. RFC 2459, Internet X.509 Public Key Infrastructure Certificate and CRL Profile. Web Engineering Task Force, January 2019.

# SQL Injection Detection Using Pattern Matching Algorithm for Library System

Mar Mar Than[1], Nwe Zin Oo[2], Tin Thein Thwel[3]
*University of Computer Studies, Yangon[1,3] , University of Computer Studies, Myeik[2]*
*marmarthan2017@gmail.com[1], nwezinoo2022.psu@gmail.com[2],*
*tintheinthwel@ucsy.edu.mm[3]*

## Abstract

*Security concerned vulnerabilities are frequently detected and exploited in modern library system. Intruders obtain unrestricted access to the information stored in the library system by exploiting security vulnerabilities. Traditional web-based library can unable to detect malicious users from SQL injection attacks. In this paper, an effective library system is proposed to detect 6 common types of SQL injection attacks by using static pattern matching algorithm. The proposed system makes use of an effective token mapping and validation with the static pattern lists whether the authenticated user or not for the library system. It can update a new anomaly pattern to the existing static pattern list whether any form of new anomaly occurs. Moreover, the matching percentage of the attacks can be calculated after detection. We modified matching algorithm which checks how many percentages based on the defined threshold and it is applied to evaluate the accuracy of the system when SQL Infections are attacked.*

**Keywords:** library system, static pattern matching, SQL injection.

## 1. Introduction

With the widespread use of web application, web-based library system can easily provide information on literature and academic areas [4]. SQL Injection is a type of web application security vulnerability in which an attacker is able to submit a SQL command in order to extract or update information in the library database that they are not authorized to access. One of the most frequent web-based application vulnerabilities, SQL injection focuses on the form of incoming SQL queries and allows users to access restricted data, get beyond authentication, and execute unwanted data manipulation language [10]. SQL Injection Attacks can be identified and avoided using a variety of methods, including encryption, extensible markup language (XML), pattern matching, parsing, and machine learning. These techniques can address login, URL, and search vulnerabilities processes by handling input type checking, pattern matching, and input encoding assaults [6].

In proposed system, the user generated SQL Queries are checked whether they are SQL injected or not by applying static pattern matching algorithm. If any form of new anomaly occurs, then a new anomaly pattern will be updated to the existing static pattern list. This work serves as a pattern matching based library system for experimenting with different SQL injection attacks and calculates the matching percentage of the attacks.

The paper is organized as follows. In the next section, the related work for pattern matching based library system is described. In Section 3, discusses the background theory of SQL injection attacks and static pattern matching algorithm for the development of library system. The proposed system design and pattern matching approach is presented in Section 4 and the experimental results and analysis are discussed in Section 5. The conclusion of this paper is summarized in the last section.

## 2. Related Work

For many years, SQL injection has been a problem, and numerous tools and methods have been created to address vulnerability [2]. M. A. Prabakar et al. [5] proposed a detection and prevention technique for preventing SQL Injection Attack (SQLIA) using Aho–Corasick pattern matching algorithm. They showed that the proposed algorithm worked well against the SQL

Injection Attack based on some sample of standard attack patterns. N. Patel and N. Shekokar [7] developed a detection and prevention technique for SQL Injection Attack using modified Aho–Corasick pattern matching algorithm. The system checked the user generated SQL queries by applying static pattern matching algorithm whether these are SQL injected or not by using SQLMAP tool and AIIDA-SQL techniques.

The proper research had been done to pinpoint the flaws, exploits, and defenses against SQL injection attacks made use of these imperfections. The researchers presented a neural network-based solution for high accuracy SQL injection detection in [9]. The system acquired authentic user URL access log data from the Internet Service Provider (ISP). The statistical research was conducted on normal data and SQL injection data. Based on statistical findings, their experimental results showed that accuracy was over 99 percent.

A study by P. Javali and S. V. Chougule [8] applied Aho-Corasick pattern matching algorithm to detect and prevent SQL injections on Bank Application. To keep user information, they employed Apache Tomcat, and MySQL. The findings demonstrate that the pattern matching technique successfully identifies and secures websites from five different forms of attacks (tautologies, illegal or illogical requests, union, piggy-backed, and alternate encodings).

## 3. SQL Injection Attacks Categories

A prominent attack method is SQL injection, which manipulates back-end databases to retrieve data that was not intended to be displayed. When harmful code is introduced as user input, it is processed by the system as a SQL query and then the malicious code is triggered to run. It has the ability to access data and either erase it or steal information [1],[11],[12]. An attacker who gains access to data and assumes the identity of a database administrator can then utilize the transferred credentials to get access to the entire system. Types of SQL injection attack are described as follow.

1. **Tautology** - It is a kind of attack in which condition becomes always true [3].
2. **Piggy-Backed Queries**- Malicious queries are inserted into an original injected query [3].
3. **Union Query** [12] -UNION keyword is used to get information by joining the injected query with safe query [3].
4. **Stored Procedure**- It is a kind of attack in which built-in stored procedure is used with malicious SQL injection codes [3].
5. **Logically incorrect query**-This attack lets an attacker to get information about the back-end database of a Web application using error message [3].
6. **Alternate Encodings** - It is a kind of attack which is used to encode the attack strings to avoid the filtering from the programmer (e.g., by using hexadecimal, ASCII and Unicode character set) [3].
7. **Inference** – In-Blind injection, hackers obtain database information by submitting a server's true/false questions and the replies from this page gives leading information that will be exploited further [3].

## 4. Proposed Library System

In the proposed library system, static pattern matching algorithm is used to identify and detect any anomaly queries by using static pattern analysis. Figure. 1 illustrates the ER diagram of the proposed system. Relational database is used to keep the information of the admin, users, books and static known patterns.

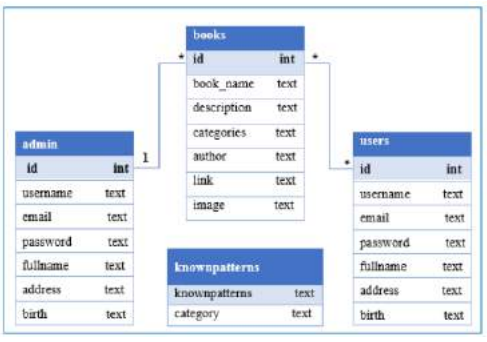



**Figure 1. ER Diagram of the Proposed System**

The step-by-step procedure of pattern matching algorithm is presented in Table 1. The core of a proposed system consists of the following design process.

| Table 1. Pattern Matching Procedure | |
|---|---|
| Input: | user input query |
| Output: | Pattern matched or not matched |
| Step 1: | User input SQL query is tokenized and sent to the Pattern Matching algorithm. |
| Step 2: | User query is compared with stored pattern in existing static database. |
| Step 3: | If it is equal to static pattern in back-end database, SQL injection attack is detected and exist from the system. |
| Step 4: | If it is not equal, then detected query is accepted. After mapping, it is added into the existing static database of library system to protect for the next SQL injections. |

As indicated in Table 2, the purpose of the proposed library system is to examine incoming SQL queries and to identify injection attacks.

| Table 2. System Design Process | |
|---|---|
| Step 1: | User generated SQL query is sent to the proposed system. |
| Step 2: | Then the queries are validated. |
| Step 3: | The procedure of pattern matching algorithm is shown in Table 1. |
| Step 4: | If the user generated query does not match each pattern in DB patterns, such a user will be validated with library user data to identify authenticate user. |
| Step 5: | Otherwise, if the pattern is match with one of the stored patterns in the anomaly pattern list, this query is identified as entering the library system with a SQL injection attack. |
| Step 6: | The query is considered as malicious user and then reject the query. |
| Step 7: | Then, alert is sent to the admin about SQL injection attack, and then pattern mapping is performed. |
| Step 8: | The anomaly pattern that are not in the static pattern list are inserted and updated to the pattern list of the proposed library system to prevent further SQL injection attack. |
| Step 9: | The evaluation is performed using the Bayes Classifier. |
| Step 10: | The output result is showed, the percentage of what kind of SQL attack is injected to the library system. |

Algorithm 1 shows the detail procedure of the static pattern matching algorithm.

Algorithm 2 shows a modified matching algorithm which is applied in step 3 of the static pattern matching algorithm.

**Algorithm 1. Static Pattern Matching Algorithm**

1: Procedure SPMA (Query, SPL[ ])
INPUT: Query ← User Generated Query
SPL[ ] ← Static Pattern List with m Anomaly Pattern
2: For j=1 to m do
3: If (MA (Query, String Length (Query), SPL[j][0]) == Ø) then

4: $Anomaly_{score} = \frac{Matching_{value}(Query, SPL[j]) \times 100}{String\ Length(SPL[j])}$

5: If ($Anomaly_{score} \geq Threshold_{value}$)
6: then
7: Return Alarm →Admin
8: Else
9: Return Query→ Accepted
10: End If
11: Else
12: Return Query→ Rejected
13: End If
14: End For
15: End Procedure

**Algorithm 2. Matching Algorithm**

```
1: Procedure MA (Query,
                 String Length (Query), SPL[j][0])

   INPUT:   SPL[j][0] ← Known Pattern
            Query ← SQL Query Statement
            n← Length of string, String Length
(Query)
2: m← Length of pattern, prevpattern ← pattern,
   pattern← pattern. Split (" "),
   Query ← Query. Split (" ") matched ← 0,
   lenofmatched ← 0
3: For i=1 to n do
4:    For j=1 to m do
5:       If pattern[j] in Query[i] &&
            pattern[j] not in matched then
6:             matched.append(pattern[j])
7:             lenofmatched += len(pattern[j])
8:       End if
9:       return (lenofmatched/len(prevpattern)) * 100
10:   End for
11: End for
12: End Procedure
```

The flow chart of the proposed library system is show in Figure 2. The user generated SQL query is firstly validated with malicious pattern list in Database (DB). The total number static patterns which are stored in DB is 24648. If there is no match in the malicious pattern list, it will continue and test with the existing user in registration lists. It there is a match, set it as a authenticate user and allow the library system to access data. If there is no match, identify that it is not the specified user and map the incoming query and no one can access the proposed library system.

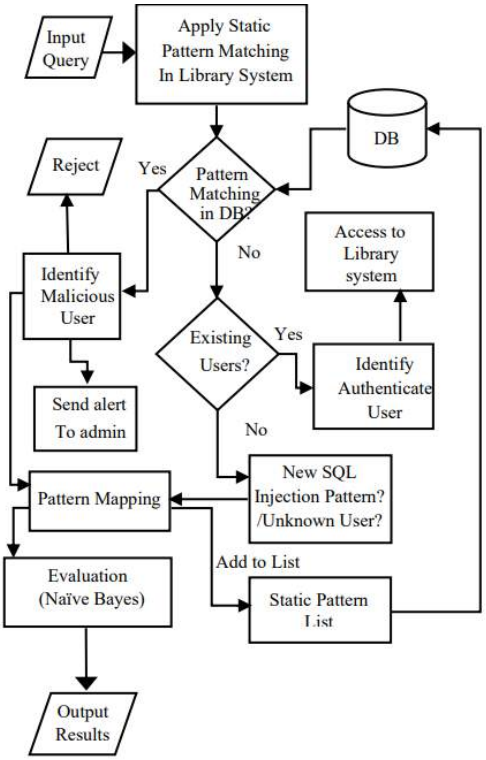



**Figure 2. Flow Chart of Proposed Library System**

If a malicious user enters an anomaly ID such as 1' or ' 1=1'—, the query becomes SELECT * FROM user WHERE id = '1' or ' 1=1'—' AND password = '1111'. The verification process is completed because the remainder of the text after—becomes a comment and "1=1" is always true and that leads to SQL injection attack.

If the matching percentage is greater than or equal to the threshold value, it is Yes.

**Input is: OR 1=1;**
100 % matched pattern is.... OR 1=1
Detected Pattern is: OR 1=1
SQL Injection detected by 100%
Alerting to admin! Attacker's IP address is 127.0.0.1 and Category is Tautology

Otherwise,

**Input is: '1' ORDERED BY marmarthan--+**
Alerting to admin! Attacker's IP address is 127.0.0.1 and Category is Tautology

### 4.1. Naive Bayes Classifier

The Naive Bayes approach is employed in the proposed system to compute the probability of a SQL injection attack in which user generated queries are made against the pattern matching algorithm in a static database system. When calculating the probability of a distinct attack based on numerous occurrences, Naive Bayes outperforms the accuracy in circumstances when computing the probability of attack occurred. Let *A* represent the static database where the SQL injection attack was discovered. Let ***B*** be the emergence of the SQL injection attack.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where *P(A|B)* is the probability of event *A* occurring given that event *B* has occurred. *P(B|A)* is the probability of event *B* occurring given that event a has occurred. *P(A*) and *P(B)* are the probabilities of observing *A* and *B* independently of each other.

## 5. Experiment and Evaluation

The experimental setup consisted of a standard desktop computer with Intel(R) Core (TM) i3-7100U CPU @ 2.40GHz and 2.4 GHz. Python 3.1 and VS code. This section presents the evaluation of the proposed system to detect SQL injection attack experiments, namely, Tautology, Union, Logically incorrect, Piggy-Backed, Alternate Encodings and Inference. Testing threshold values is 80 for increasing accuracy.

**Table 3. Attack Types Evaluation**

| Types of Attacks | Total | SQL Injection Attack | |
|---|---|---|---|
| | | **≥ threshold (Yes)** | **< threshold (No)** |
| Tautology | 6 | 5 | 1 |
| Union | 6 | 4 | 2 |
| Logically incorrect | 6 | 4 | 2 |
| Piggy-Backed | 6 | 3 | 3 |
| Alternate Encodings | 6 | 3 | 3 |
| Inference | 6 | 4 | 2 |
| **Total Tests** | **36** | **23** | **13** |

| **SQL Injection Attack** | | | |
|---|---|---|---|
| Types of Attacks | **(Yes)** | **(No)** | **Probability** |
| Tautology | 5 | 1 | 6/36 |
| Union | 4 | 2 | 6/36 |
| Logically incorrect | 4 | 2 | 6/36 |
| Piggy-Backed | 3 | 3 | 6/36 |
| Alternate Encodings | 3 | 3 | 6/36 |
| Inference | 4 | 2 | 6/36 |
| | **23/36** | **13/36** | 6/36 |

Table 3 illustrates the types of attacks and evaluation. We assumed that if the matching percentage is greater than or equal to the threshold value, the probability of fully attacked is Yes. Otherwise, assuming is No.

Likelihood of *Yes* given Tautology attack is

P (Tautology |Yes) = 5/23 = 0.217,
P(Tautology) = 6/36 = 0.167
P(Yes) = 23/36 = 0.639

$$P(Yes|Tautology) = \frac{P(Tautology|Yes) * P(Yes)}{P(Tautology)}$$

$$P(Yes|Tautology) = \frac{0.217 * 0.639}{0.167}$$

$$P(Yes|Tautology) = 0.830$$

Likelihood of *No* given Tautology attack is

P (Tautology |No) = 1/13 = 0.076
P(Tautology) = 6/36 = 0.167
P(No) = 13/36 = 0.361

$$P(No|Tautology) = \frac{P(Tautology|No) * P(No)}{P(Tautology)}$$

$$P(No|Tautology) = \frac{0.076 * 0.361}{0.167}$$

$$P(No|Tautology) = 0.164$$

Likelihood of *Yes* given Union attack is

P (Union| Yes) = 4/23 = 0.174,
P(Tautology) = 6/36 = 0.167
P(Yes) = 23/36 = 0.639

$$P(Yes|\text{Union}) = \frac{P(Union|Yes) * P(Yes)}{P(\text{Union})}$$

$$P(Yes|\text{Union}) = \frac{0.174 * 0.639}{0.167}$$

$$P(Yes|\text{Union}) = 0.666$$

Likelihood of No given Union attack is

P (Union |No) = 1/13 = 0.076
P(Union) = 6/36 = 0.167
P(No) = 13/36 = 0.361

$$P(No|\text{Union}) = \frac{P(\text{Union}|No) * P(No)}{P(\text{Union})}$$

$$P(No|\text{Union}) = \frac{0.154 * 0.361}{0.167}$$

$$P(No|\text{Union}) = 0.333$$

Likelihood of *Yes* given Logically incorrect attack is

P (Logically | Yes) = 4/23 = 0.174,
P (Logically) = 6/36 = 0.167
P(Yes) = 23/36 = 0.639

$$P(Yes|\text{Logically}) = \frac{P(\text{Logically}|Yes) * P(Yes)}{P(\text{Logically})}$$

$$P(Yes|\text{Logically}) = \frac{0.174 * 0.639}{0.167}$$

$$P(Yes|\text{Logically}) = 0.666$$

Likelihood of No given Logically incorrect attack is

P (Logically |No) = 1/13 = 0.076
P (Logically) = 6/36 = 0.167
P(No) = 13/36 = 0.361

$$P(No|\text{Logically}) = \frac{P(\text{Logically}|No) * P(No)}{P(\text{Logically})}$$

$$P(No|\text{Logically}) = \frac{0.154 * 0.361}{0.167}$$

$$P(No|\text{Logically}) = 0.333$$

Likelihood of *Yes* given Piggy-Backed attack is

P (Piggy| Yes) = 3/23 = 0.131
P(Piggy) = 6/36 = 0.167
P(Yes) = 23/36 = 0.639

$$P(Yes|\text{Piggy}) = \frac{P(\text{Piggy}|Yes) * P(Yes)}{P(\text{Piggy})}$$

$$P(Yes|\text{Piggy}) = \frac{0.131 * 0.639}{0.167}$$

$$P(Yes|\text{Piggy}) = 0.501$$

Likelihood of No given Piggy-Backed attack is

P (Piggy |No) = 3/13 = 0.231
P (Piggy) = 6/36 = 0.167
P (Piggy) = 13/36 = 0.361

$$P(No|\text{Piggy}) = \frac{P(\text{Piggy}|No) * P(No)}{P(\text{Union})}$$

$$P(No|\text{Piggy}) = \frac{0.231 * 0.361}{0.167}$$

$$P(No|\text{Piggy}) = 0.499$$

Likelihood of Yes given Alternate Encoding attack is

P (Alternate| Yes) = 3/23 = 0.131
P (Alternate) = 6/36 = 0.167
P(Yes) = 23/36 = 0.639

$$P(Yes|\text{Alternate}) = \frac{P(\text{Alternate}|Yes) * P(Yes)}{P(\text{Alternate})}$$

$$P(Yes|\text{Alternate}) = \frac{0.131 * 0.639}{0.167}$$

$$P(Yes|\text{Alternate}) = 0.501$$

Likelihood of No given Alternate Encoding attack is

P (Alternate |No) = 1/13 = 0.076
P (Alternate) = 6/36 = 0.167
P(No) = 13/36 = 0.361

$$P(No|\text{Alternate}) = \frac{P(\text{Alternate}|No) * P(No)}{P(\text{Alternate})}$$

$$P(No|\text{Alternate}) = \frac{0.231 * 0.361}{0.167}$$

$$P(No|\text{Alternate}) = 0.499$$

Likelihood of Yes given Inference attack is

P (Inference| Yes) = 4/23 = 0.174
P(Inference) = 6/36 = 0.167
P(Yes) = 23/36 = 0.639

$$P(Yes|\text{Inference}) = \frac{P(\text{Inference}|Yes) * P(Yes)}{P(\text{Inference})}$$

$$P(Yes|\text{Inference}) = \frac{0.174 * 0.639}{0.167}$$

$$P(Yes|\text{Inference}) = 0.666$$

Likelihood of No given Inference is

P (Inference |No) = 2/13 = 0.154
P (Inference) = 6/36 = 0.167
P (*No*) = 13/36 = 0.361

$$P(No|\text{Inference}) = \frac{P(\text{Inference}|No) * P(No)}{P(\text{Inference})}$$

$$P(No|\text{Inference}) = \frac{0.154 * 0.361}{0.167}$$

$$P(No|\text{Inference}) = 0.333$$

Table 4 show the performance evaluation of the proposed library system in terms of accuracy.

**Table 4. Performance Evaluation**

| Types of Attacks | Total Attacks | Probability of Yes | Probability of No |
|---|---|---|---|
| Tautology | 6 | 0.830 | 0.164 |
| Union | 6 | 0.666 | 0.333 |
| Logically incorrect | 6 | 0.666 | 0.333 |
| Piggy-Backed | 6 | 0.501 | 0.499 |
| Alternate Encodings | 6 | 0.501 | 0.499 |
| Inference | 6 | 0.666 | 0.333 |
| Total | 36 | 3.828 | 2.161 |

$$P(Yes|\text{Attacks}) = \frac{P(\text{Attacks}|Yes) * P(Yes)}{P(\text{Attacks})}$$

$$= \frac{9.404 * 3.828}{36}$$

Accuracy = 0.999%

The experimental results show that the proposed system achieves above 99% accuracy in classifying the injected SQL statements for 6

common types of SQL attack. The implementation of the proposed technique effectively detects and blocks all types of SQL Injection attacks. Therefore, the propose library system can identify and detect SQL injections, according to the experimental results various injection attacks.

## 5. Conclusion

SQL injection attacks and web-based attack are major issues in the security of financial, health, and other critical data, and this problem only increases in importance to protect the malicious queries. This paper proposes a library system that can detect against 6 common types of SQL injection attacks when log-in authentication stage. In addition, it can detect and blocks code SQL injection vulnerabilities effectively using modified pattern matching technique. The experimental results provide that the proposed algorithm handles malicious queries effectively matching and prevents unauthenticated users for library system. Some of the limitation of this study is that the proposed system can only update the existing static pattern list if a new absolutely attack patterns has been attacked. In future work, the detail experiment of a new malicious pattern entering to the proposed system will be performed.

## References


[1] Blind SQL Injection. https://owasp.org/ www-community/attacks/ Blind_SQL_Injection

[2] D. Kar, S. Panigrahi and S. Sundararajan, "SQLiGoT: Detecting SQL Injection Attacks using Graph of Tokens and SVM, Computers & Security, 2016. Doi: 10.1016/j.cose.2016.04.005.

[3] I. Lee, S. Jeong, S. Yeo and J. Moon, "A novel method for SQL injection attack detection based on removing SQL query attribute values". International Journal of Mathematical and Computer Modelling, Vol. 55, Issues 1–2, January 2012, pp. 58-68.

[4] K. Zhang, "A Machine Learning based Approach to Identify SQL Injection Vulnerabilities". In Proceeding of 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 1286-1288, pp.75-91,2019.

[5] M. A. Prabakar, M. K. Keyan and K. Marimuthu, "An Efficient Technique for Preventing Sql Injection Attack Using Pattern Matching Algorithm". In Proceeding of IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013), pp.503-506.

[6] M. Hasan, Z. Balbahaith, M. Tarique, "Detection of SQL Injection Attacks: A Machine Learning Approach", In Proceeding of IEEE International Conference on Electrical and Computing Technologies and Applications (ICECTA),2019.

[7] N. Patel, N. Shekokar, "Implementation of pattern matching algorithm to defend SQLIA", In Proceeding of International Conference on Advanced Computing Technologies and Applications (ICACTA), 2015.

[8] P. Javali, S.V. Chougule, "SQL Injection Detection and Prevention using Pattern Matching Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016.

[9] P. Tang, W. Qiu, Z. Huang, ""Detection of SQL Injection Based on Artificial Neural Network", Journal Knowledge-Based Systems, vol.190,2020. Doi: 10.1016/j.knosys.2020.105528.

[10] T. K. George, K. P. Jacob and R. K. James, "Token based Detection and Neural Network based Reconstruction framework against code injection vulnerabilities", International Journal of Information and Application, vol.41, 2018. Doi: 10.1016/j.jisa.2018.05.005

[11] SQLi. https://www.acunetix.com/blog/articles/sqli-part-4-in-band-sqli-classic-sqli/

[12] SQLi.https://www.acunetix.com/blog/articles/sqli-part-6-out-of-band-sqli/

# Prevention of Cross-Site Request Forgery Using Anti-CSRF Token

Phyu Phyu Win, Yi Mon Thet
*University of Computer Studies, Yangon*
*phyu95368@gmail.com, yimonthet@ucsy.edu.mm*

## Abstract

*Cross-Site Request Forgery (CSRF) is an attack that forces an end user to execute unwanted actions on a web application in which they're currently authenticated. CSRF attacks specifically target state-changing requests, not theft of data. This attacks target functionality that causes a state change on the server such as changing the victim's email address, password or purchasing something. The proposed system illustrates the secure transaction in banking system and provides the data security of the customer's critical transmission data. This system is implemented to prevent the CSRF attack using anti-csrf token. The Blum Blum Shub algorithm is used to generate the Anti-csrf token that is a secret, unique and unpredictable value in order to protect CSRF vulnerable resources. The tokens are generated and submitted by the server-side application and SHA-256 hash is used when sending to the client site. This system prevents the csrf attack using the anti-csrf token when making transactions in banking system.*

**Keywords**: CSRF, HMAC (sha-256), Blum Blum Shub

## 1. Introduction

Banking is popular today and uses to deposit/withdraw the cash [2]. This is the place where customers feel the sense of safety for their property. This is why banking becomes an important role in lives. In the act of using web applications, people give personal information to the organization, and then store sensitive information on them. On the other hand, some attackers who are unethical and selfish exploit the web application to gain unauthorized access and do other things such as identity theft, privacy violation, and other cyber-attacks. These illegal points allow the attackers to make whatever they want through the weaknesses of the web application.

Vulnerability is the weak point of the web application caused by unawareness of the developers who cannot be handled validation the user inputs, appropriate validation methods, and so on. Because of those facts, detection of vulnerability is needed more. There are so many different kinds of vulnerabilities but, it is indicated to OWASP in 2013 that CSRF attack reaches number eight vulnerabilities [8]. Cross-site request forgery (CSRF) is a web security vulnerability that allows an attacker to induce users to perform actions that they do not intend to perform [9]. In a successful CSRF attack, the attacker causes the victim user to carry out an action unintentionally. This might be to change the email address on their account, to change their password, or to make a funds transfer.

## 2. Related Works

The system intends to support the admins who require to obtain secure transaction without vulnerabilities and to prevent CSRF attacks from the attacker. The system uses anti-csrf token and also is generated by a random number generator. In this section, we discuss the previous studies of preventing CSRF vulnerabilities concerning needed to protect against.

In 2014, Sentamilselvan. K [12] Assistant that describes the implementation of cross site request forgery method using tools[]By employing parsing techniques to identify the attacking spots prior to the attackers' attack, their experiment provides suitable solutions for the cross-site request forgery attack. It takes a long time and doesn't use any more memory.

In 2006, Emil Semastin [13] implemented to identify the available solutions to prevent CSRF attacks using tools Pinata, CSRF tester, Burp Suite and OWASP ZAP. Tests against the exploitation of the vulnerabilities were conducted after implementing the solutions into the web

application to check the efficacy of each of the solutions. The suggested solution is a combination of the most effective existing technique and the second-best option. By implementing this, a double validation takes place at the server side of the web application to ensure the prevention of CSRF attacks.

In 2018, Sami Azam [4] introduced the 'Preventive Measures for Cross Site Request Forgery Attacks on Web-based Applications' that identify the available solutions to prevent CSRF attacks. By analyzing the techniques employed in each of the solutions, the optimal tool can be identified. Tests against the exploitation of the vulnerabilities were conducted after implementing the solutions into the web application to check the efficacy of each of the solutions. The research also proposes a combined solution that integrates the passing of an unpredictable token through a hidden field and validating it on the server side with the passing of token through URL.

In 2008, Adam Barth and Collin Jackson [1] that examine the scope and diversity of CSRF vulnerabilities, study existing defenses, and describe incremental and new defenses based on headers and web application firewall rules. We introduce login cross-site request forgery attacks, which are currently widely possible, damaging, and under-appreciated. There are three widely used techniques for defending against CSRF attacks: validating a secret request token, validating the HTTP Referrer header, and validating custom headers attached to XML Http Requests.

The research methodology is expressed in section (3), Cross Site Request Forgery Attack, Blum Blum Shub Algorithm and Anti-CSRF Token. Section (4) explains proposed work that includes System Flow Diagram of the Proposed System, Algorithm of Proposed System and HMAC (SHA-256). Then, in section (5), discuses the conclusion of the research.

## 3. Background Theory

The Open Web Application Security Project (OWASP) is a nonprofit foundation that works to improve the security of software. It represents a broad consensus about the most critical security risks to web applications. The cross-site request forgery attack is the popular attack in OWASP top ten attacks.

### 3.1. Cross Site Request Forgery Attack

An end user can be tricked into loading or submitting data to a web application in a number of different ways [8]. Before we can carry out an attack, we need to know how to create a legitimate malicious request for the victim to carry out. Consider the following illustration: Alice intends to use the CSRF-vulnerable bank.com web application to send Bob $100.Maria is an attacker who wants Alice to send the money to her instead of Maria. The following steps will be included in the attack:

- building an exploit URL
- tricking Alice into executing the action with social engineering

### 3.2. GET scenario

GET
http://bank.com/transfer.do?acct=BOB&amount=100 HTTP/1.1.

Maria now makes a decision to exploit this web application vulnerability the usage of Alice as her sufferer. Maria first constructs the subsequent make the most URL so one can switch $100,000 from Alice's account to her account. She takes the original command URL and replaces the beneficiary's name with herself, raising the transfer amount significantly at the same time:

http://bank.com/transfer.do?acct=MARIA&amount=100000

The social engineering aspect of the attack tricks Alice into loading this URL when she's logged into the bank application. This is usually done with one of the following techniques:

- sending an unsolicited email with HTML content
- planting an exploit URL on pages

The exploit URL can be disguised as an ordinary link, encouraging the victim to click it:

<ahref="http://bank.com/transfer.do?acct=MARIA&amount=100000">View my Pictures! </a>

Or

Fake image:

<img src="http://bank.com/transfer.do?acct=MARIA&amount=100000" width="0" height="0" border ="0">

Alice would not see anything if the email contained this image tag. However, the browser will still send the request to bank.com despite the fact that there is no visible sign that the transfer has occurred.

### 3.3. POST scenario

Let's assume the bank now uses POST and the vulnerable request looks like this:
POST http://bank.com/transfer.do HTTP/1.1 acct= BOB & amount=100

Such a request cannot be delivered using standard A or IMG tags, but can be delivered using a FORM tag:

<form action="<nowiki>http://bank.com/transfer.do </nowiki>" method="POST">
<input type="hidden" name="acct" value="MARIA"/>
<input type="hidden" name="amount" value="100000"/>
<input type="submit" value="View my pictures"/> </form>

This form will require the user to click on the submit button, but this can be also executed automatically using JavaScript:
<body onload="document. Forms [0].submit ()">

### 3.4. Blum Blum Shub Algorithm for generating anti-csrf token

The Blum Blum Shub is proposed in 1986 by Lenore Blum, Manuel Blum and Michael Shub [7]. The generator BBS works as follow: Select two sufficiently large secret random (and different) prime integers p and q,

Begin
Compute n = pq.
Select a random integer 0 <S < n (the seed) such that gcd (S, n) =1
Compute $y = S^2 \bmod n$
For i from 1 to N do the following:
$y_i = y_{i-1}^2 \bmod n$
$x_i = y_i \bmod 2$ the least significant bit of $y_i$
The output sequence is $x_1, x_2, \ldots, x_i$.
End

Example generation
Let- two-primes number
Let p=11, q=23, s=3
Calculate M = p.q
= 11.23
= 253
Calculate $x_0$ $= s^2 \bmod M$
$= 3^2 \bmod 256$
= 9
For i from 1 to 5
$x_i = x^2_{i-1} \bmod M$
$y_i$ = at least significant bit ($x_i$)
The output sequence is $y_{1\ldots,} y_5$
= 81, 236, 36, 31,202

### 3.5. Anti CSRF-token

A CSRF token is a secure random token that is used to prevent CSRF attacks [2]. The server-side application compares the two tokens in the person consultation and the request after the request has been made. The request is rejected, the user session is ended, and the event is recorded as a possible CSRF attack if the token is missing or does not match the value during the session.

This system introduces the token that is generated by the sever-side application using random number generator. This system that is normally effective is to transmit the token to the client within a hidden field of the user submit form. The token will then be included as a request parameter-

<input type="hidden" name="csrf-token" value= "CIwNZNlR4XbisJF39I8 yWnWX9wX4WFoz" />

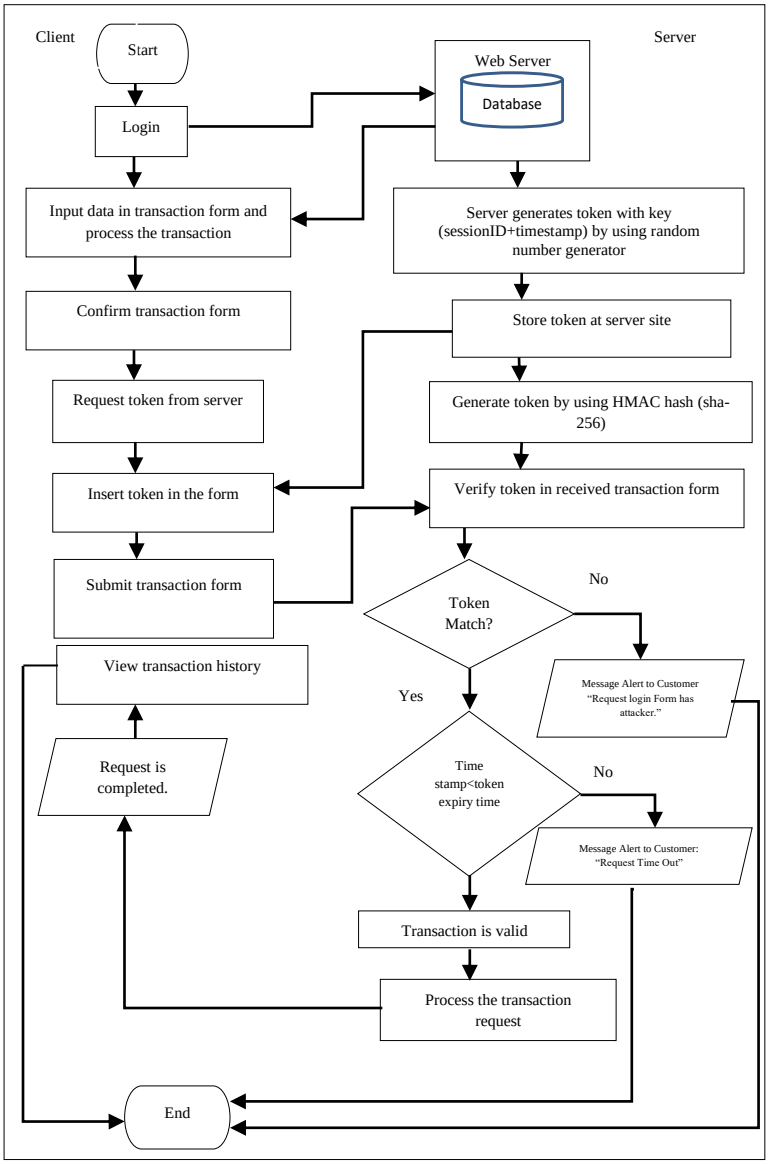



**Figure 1. System Flow Diagram of the Proposed System**

## 4. Proposed Work

This system aims to prevent csrf attack using anti-csrf token.

### 4.1. Algorithm of Proposed System

BEGIN

**Step1**: Create CSRF token

1.1 Start a session on the server;

1.2 Generate a user session ID (using random number generator);

1.3 Keep Token Expire Time;

1.4 Generate a CSRF Token using key K

1.4.1 Generate HMAC (user session ID + timestamp)

1.4.2 Append the same timestamp value to it

**Step2**: Include the token in the form (i.e., HMAC + timestamp)

2.1 Inject the token into the hidden field of the User Submit form;

**Step3**: Validate the token

3.1 Regenerated the token with the same key K (parameter are session ID from the request and timestamp in the received token).

3.2 If ("If the HMAC in the received token and the one generated in this step match") {

If (Timestamp received is less than defined token expiry time) {
Request is treated as legitimate and can be allowed;
} Else {
Request Time Out;
}
End If
Else If ("If the HMAC in the received token and the one generated in this step not match")
{
Reject the process;
}
End If
End If
END

### 4.2. CSRF Attack System Flow

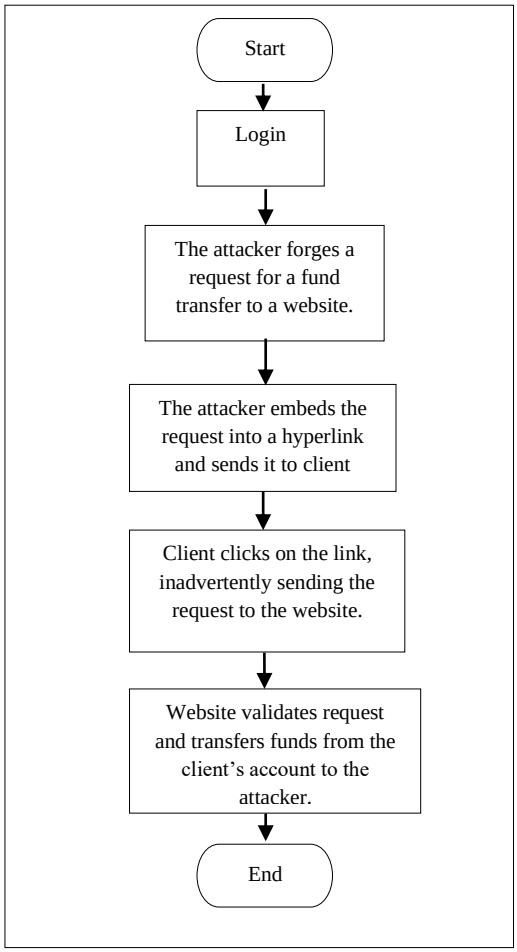



**Figure 2. CSRF Attack System Flow**

### 4.3. HMAC (SHA-256) using transmitting token to the client

SHA-256 stands for Secure Hash Algorithm 256-bit and it's used for cryptographic security [10]. A hash is not 'encryption' – it cannot be decrypted back to the original text. It is almost impossible to reconstruct the initial data from the hash value. To crack a hash, you need all 64 of the digits to match. It would take a long time to crack a SHA256 hash using the entire network.

HMAC stands for Keyed-Hashing for Message Authentication Code [5]. It's a message authentication code obtained by running a cryptographic hash function over the data (to be authenticated).

$HMAC(K, m) = H((K' \oplus opad) \| H((K' \oplus ipad) \| m))$
K' = H (K) K is a larger than block size, K otherwise

Where:
- H is a cryptographic hash function
- m is the message to be authenticated
- K is the secret key K' is a block-sized key derived from the secret key,
- || denotes concatenation
- ⊕ denotes bitwise exclusive or (XOR)
- opad is the block-sized outer padding
- ipad is the block-sized inner padding

HMAC (key, msg, hash_func) -> hash

This system uses a secure hash function (sha-256) in HMAC method when transmitting token into the user submits form.

### 4.4. Evaluation of the Proposed System

**Table 1. Attack Detection**

| Detection | Attacked (%) | Attack Affected (%) | Defence (%) |
|---|---|---|---|
| Without Anti CSRF Token | 100 | 100 | 0 |
| With Anti CSRF Token | 100 | 0 | 100 |

The proposed system was evaluated in terms of in percentage. In detection attack, without anti-csrf token in 100 times, the attack affected 100 %

and defense in 0%. In detection attack, with anti-csrf token in 100 times, the attack affected 0% and defense in 100%. In the system, the detection of the attack is showed with the percentage.

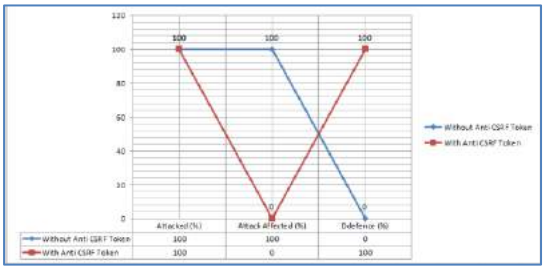



**Figure 3. Attack Detection Rate with Graph**

## 5. Conclusion

CSRF attacks exploit a vulnerability which is included in the structure of websites and the behavior of the browser by design. The random number generator is used to generate key when creating a token. The token is essential to prevent the attack for state changing functionalities in web applications.

The token is a unique ID and transmits the transaction in a secure hash (sha-256). SHA-256 is used when generating a hash code to transmit the token to the client side. The token checks at the server side. When the token is not matched, the system alerts message to the customer "Login Form has the attack". This system is to show the secure transaction, the attack transaction, record transaction history and prevents the csrf attack using the anti-csrf token.

## References

[1] A.Barth, C.Jackson, and J.C.Mitchell. "Robust defenses for cross site request forgery". In Proc. ACM Conference on Computer and Communications Security (CCS), Oct, 2008.

[2] Anti-csrf token, https://blog.insiderattack.net/anti-csrf-tokens-to-prevent-cross-site-request-forgery-csrf.

[3] Banking System prevent from the Attack, https://blog.nettitude.com/how-can-banks-protect-themselves-from-cyber-attacks.

[4] Emil Semastin, Sami Azam,Preventive Measures for Cross Site Request Forgery Attacks on Web-based Applications, College of Engineering, IT and Environment, Charles Darwin University, Australia, 2018.

[5] HMAC, hash-based message authentication code, https://en.wikipedia.org/wiki/HMAC

[6] Nenad Jovanovic, Engin Kirda, and Christopher Kruegel. "Preventing cross site request forgery attacks".In IEEE International Conference on Security and Privacy in Communication Networks (SecureComm), 2006.

[7] Lenore Blum, Manuel Blum and Michael Shub,https://en.wikipedia.org/wiki/Blum_Blum_Shub,1986.

[8] OWASP. Top ten most critical web applications securityvulnerabilities.https://www.owasp.org/index.php/Top_10_2013Top_10.Forgeries.www.securityfocus.com/archive/1/19S90,2001.

[9] OWASP. (2017), CSRF prevention cheat sheet. https://www.owasp.org/index.php/CrossSite_Request_Forgery_(CSRF)_Prevention_Cheat_Sheet.

[10] Ozgur Yildirim, Daniel Gerep, SHA-256 Algorithm, 2020. https://blog.boot.dev/cryptography/how-sha-2-works-step-by-step-sha-256/

[11] Sooel Son, "Prevent Cross site Request Forgery PCRF"userweb.cs.utexas.edu/~samuel/PCRF/Final_PCRF_paper.pdf.

[12] Sentamilselvan. K, Lakshmana Pandian. S ": Preventive Measures", Assistant Professor Kongu Engineering College Perundurai, Tamilnadu, 2014.

[13] W. Zeller and E. W. Felten, "Cross-Site Request Forgeries: Exploitation and Prevention," Technical Report, Princeton University, 2008.

# Internal Revenue Department (IRD) Data System by Using Blowfish Algorithm

Pyae Sandar Win, Yu Wai Hlaing
*University of Computer Studies (Yangon)*
*pyaesandarwin@ucsy.edu.mm*

## Abstract

*The document storage system has become the most popular for electronic communication in the world. Documents are used in every organization day by day and most of these documents' data need security system. A cryptographic system (or a cipher system) is a method of securing information and communication through the use of code so that only authorized person can read and process it. A cryptographic system consists of keys, algorithms, and key management facilities. In this system, Blowfish Algorithm is applied to encrypt and decrypt the files which is used in Internal Revenue Department (IRD) system. Blowfish is a variable-length key with sixteenth rounds and with each block sizes of 64 bits. It can encrypt and decrypt the document by selecting the file on symmetric keys.*

**Keywords**: Cryptography, Blowfish, Data, Encryption, Decryption, Block Cipher.

## 1. Introduction

Security is one of the biggest concerns in communications and electronic applications. When exchanging data according to the electronic system, information security is becoming necessary. Due to the growth of multimedia applications, security has become an important issue for communication and data storage. There are many reasons why data/information is crucial for organizations all over the world. Data security is a very important task for the today enterprise. Today, organizations are increasingly responsible to protect customer and user data from being lost or stolen.

Nowadays, most of the communication is done by using electronic media. Most of the data/information is also collected and stored on the server or database through different networks. There is a need to protect from various access. Data security plays an important role in such communication. One of the most important methods for ensuring data secrecy is cryptography. Cryptography is the process of hiding or coding information. It not only protects against data theft or alteration, but can also be used for user authentication.

In this paper, Blowfish algorithm is used in data storing and processing. The main purpose behind the design of this system is to get the best security and performance. In our system, Blowfish algorithm is applied to encrypt and decrypt IRD system' files.

The organization of the paper is as follows. Related work is presented in section 2. Section 3 presents IRD system, section 4 describes performance evaluation of the system, limitations of the system is described in section 5 and final conclusions are drawn in section 6.

## 2. Related Work

In recent years, there has been a need for better technologies for securely transmitting and storing information. Cryptography has become a way to solve some of these requirements and has been focused on a growing research effort. The main of this field is to effectively identify cryptography algorithms in software and/or hardware.

In [6], Blowfish algorithm is proposed for encoding or decoding of text data while sending data to the cloud. The data processing section of encoding and decoding is used by android software. In this work, mock the data processing section of encoding and decoding using Android software. The message/text has been considered as input to the Blowfish algorithm. On behalf of that a key which is also used as input for the encryption. Result shows the original text, encrypted data, and the decrypted data. The encrypted data by using this algorithm will be

sent to cloud. The proposed algorithm is meant using android(java).

Pia Singh [8] was proposed for encryption and decryption of images using a secret-key cryptography. They have simulated the image processing part of encryption and decryption in MATLAB software. Here taking an image. Firstly, obtaining the matrix and pixels of the chosen image & then encrypting the image matrix using blowfish algorithm. The result shows the original image, encrypted image, and the decrypted image.

Sarita Kumari [9] was proposed of cryptography encryption and compression techniques. This paper gives a brief idea on different encryption and decryption techniques associated with Symmetric & Asymmetric Key Cryptography algorithms. Along with this it also provides ideas on different compression technique that is used for transmission. Basically, compression is a technique which reduces the size of data which in turn saves space. Cryptography enables us to confidentially transmit the data such that no data is altered. Data is a type of stored information it can be digital or physical. Security means providing protection to assets. Data security means security of data privacy to prevent unauthorized access to computers, personal databases etc. Cryptography protects users by acting as a proxy or firewall allowing only the legitimate users access the data & blocking fake users from access the data. Cryptography is a famous way of sending confidential data in a secret way.

H. Fathima, K.S.R. Matriculation & K.S.R. Kalvi nagar [4] was proposed for comparative study of symmetric key algorithm. This paper presents an analysis in the field of cryptographic algorithms, focusing on the private key ciphers which are used for bulk data and link encryption. This paper's main aim is to study different kinds of cryptographic algorithms that are used today and provide differences between them as comparative study in form of literature survey. The study further represents performance of each encryption algorithm and analyze security issues of each algorithm. A detailed study of various symmetric key encryption algorithms such as DES, & AES is provided in this paper.

# 3. Internal Revenue Department (IRD) System

## 3.1. System Design

The system consists of two roles: admin side and taxpayer side. In this system, data encoding and decoding process is managed by Blowfish algorithm.

At the taxpayer side, he/she will need to login with email and password to use system. When the email and password are correct, he/she can make upload and download action. The taxpayer must enter a secret key before the return file is uploaded. In the background of the application the program takes the information, such as key and return file, work through the operation steps of the Blowfish algorithm. After these steps, the return file is successfully uploaded and shown on e-filling page.

The taxpayer can also download the tax amount file uploaded by the admin on the e-filling page. When he/she downloads the tax amount file, he/she will need to enter the secret key to generate the original tax amount file. In the background of the application, the program takes the secret key, work through the operation steps of the Blowfish algorithm. After these steps, the program outputs the original tax amount file will appear. System flow for taxpayer is shown in figure 1.

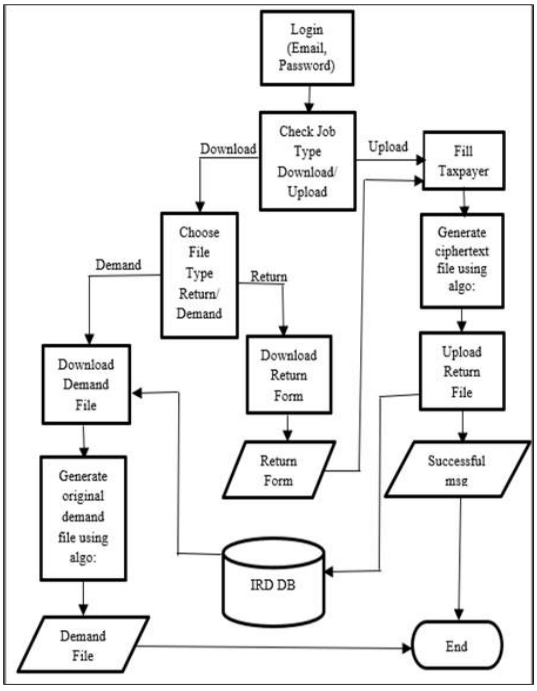



**Figure 1. System Flow for Taxpayer**

At the admin side, he/she will need to login with email and password to use the system. When the email and password are correct, he/she can upload and download. If the admin wants to view return file uploaded by the taxpayer, he/she downloads the return file from e-filling page by using secret key to generate the original return file. In the background of the application, the program takes the secret key, and then works through the operation steps of the Blowfish algorithm. After these steps, the program outputs the original return file will appear.

The admin looks at the return file downloaded and calculates the tax amount to be paid. And the admin uploads the calculated tax amount file to e-filling page. The admin must enter a secret key before the calculated tax amount file is uploaded. In the background of the application, the program takes the information, such as key and calculated tax amount file, work through the operation steps of the Blowfish algorithm. After these steps, the tax amount file is successfully uploaded and shown on an e-filling page. In figure 2, admin's system flow is shown.

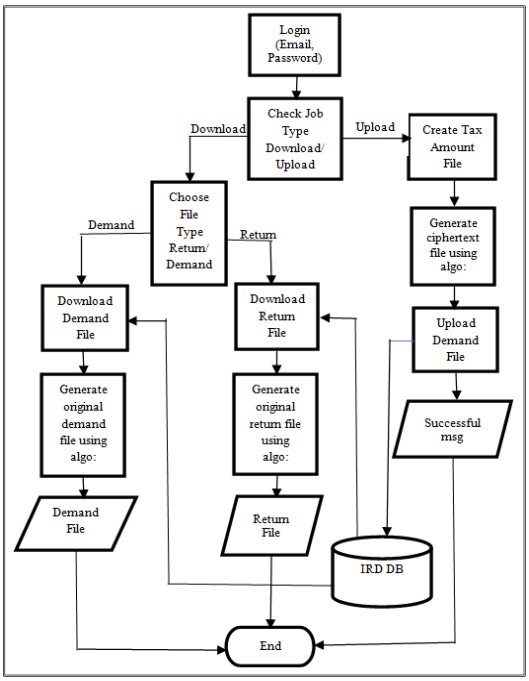



**Figure 2. System Flow for Admin**

## 3.2. Blowfish Structure

The Blowfish algorithm is a symmetric block cipher which uses the same key at the time of encryption and decryption. This algorithm was developed by Bruce Schneier in 1993.This algorithm is otherwise called Blowfish Cipher or Blowfish Encryption. The Blowfish algorithm is license free and available free for all users.

Blowfish is a 64-bit block cipher and uses a variable-length key from 32 to 448-bits (14 bytes). The algorithm was developed to efficiently and securely convert from 64-bit plaintext to 64-bit ciphertext. Blowfish algorithm consists of a 16 rounds Feistel-iteration for encryption and decryption. Each iteration operates on a 64-bit block that's divided into two blocks of each 32-bit. Every block works the same process in the blowfish's encryption standard.

The Blowfish algorithm consists of two main procedures: key expansion and data encryption. The operators selected for the Blowfish algorithm were table lookup, addition and bitwise exclusive-or. The table lookup consists of four S-boxes [each having 256 entries] and a P-array [18 each of having 32-bits].

Data encryption function consists of two parts: round and post-processing. During each round of blowfish, the 32-bits data on the right and left sides are exchanged to become the input of the next round. In post-processing, the last two unused P-box entries are XOR with plaintext from previous round. All of these steps are completed, 64-bits ciphertext will be available. The function "F" is shown in figure 3.

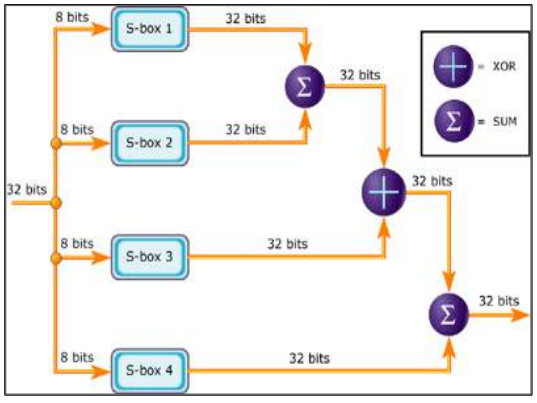



**Figure 3. Workflow Diagram of Function "F"**

- Key Size: 32-bits to 448-bits
- Number of subkeys: 18 subkeys
- Number of rounds: 16 rounds
- Number of substitution boxes: 4 [each having 256 entries of 32-bits each]
- **Step1**: Initial step of this algorithm is to fix a key array with the size of 18 and 4 substitutional boxes with a maximum of 256

entries and in size of 32 bit are initialized for each p-array. Each Key string should be in the form of hexadecimal digits.

- **Step2**: For the first iteration, input data of size 64 bits is divided into two 32-bits data as LX and RX. Then LX is XOR with P1 key and then the output LX1 is given as input to the F function.
- **Step3**: Inside F function we have four s boxes of size of 32 bit each,8-bit input is given to each four s box and 32-bits output generated by each S box, output of S1 box is added with S2 box and then S3 and S4 boxes are also added finally these two added outputs are XOR to get final encrypted data.
- **Step4**: The output value of "F" function is XOR with another 32-bits input R X to get out put RX1.
- **Step5**: The Output LX1 and RX1 is swapped and given as the input for next iteration
- **Step6**: These steps are completed after 18 rounds of iteration and finally 64-bits ciphertext is received from the output.
- Decryption is identical to encryption, with the exception that the P-array is used in reverse order.

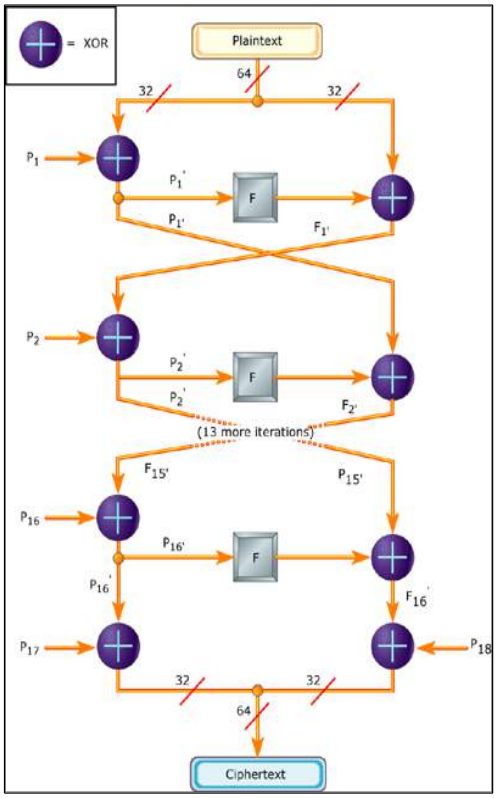



**Figure 4. Workflow Diagram of Blowfish Algorithm**

## 4. Performance Evaluation

In this paper, the encryption and decryption time is tested on English language with different workloads of PDF file (15KB, 35KB, 55KB, 75KB and 95KB). The processing time is measured by capturing the time difference between the starting point and ending point of the algorithm. Any change to input file in transit will result in a different time. This result is implemented using C# programming language and the execution of the developed tool on a personal computer equipped with an Intel ® corei7 1.8 GHz CPU, 8G RAM, Window 10 operating system. Figure 5. shows in runtime of encryption at various file size.

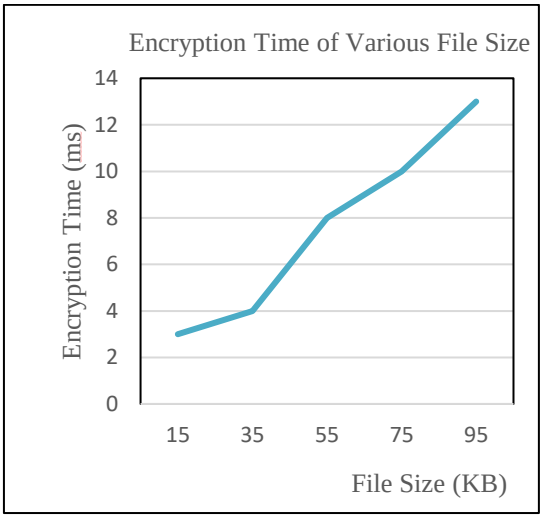



**Figure 5. Encryption Time of Various File Size**

Runtime of decryption at various file size is shown in figure 6.

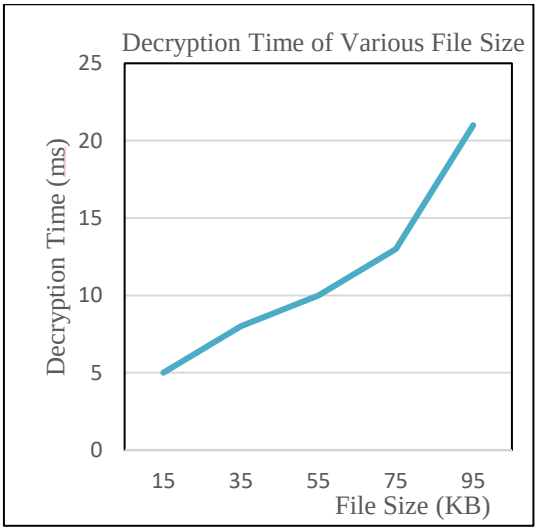



**Figure 6. Decryption Time of Various File Size**

Table 1 shows various file sizes which is used in experiment and figure 7 shows our experimental results on encryption and decryption time at various file sizes.

**Table 1. Encryption and Decryption Time of Various File Sizes**

| File Size (KB) | Encryption Time (Milliseconds) | Decryption Time (Milliseconds) |
|---|---|---|
| 15KB | 3 | 5 |
| 35KB | 4 | 8 |
| 55KB | 8 | 10 |
| 75KB | 10 | 13 |
| 95KB | 13 | 21 |

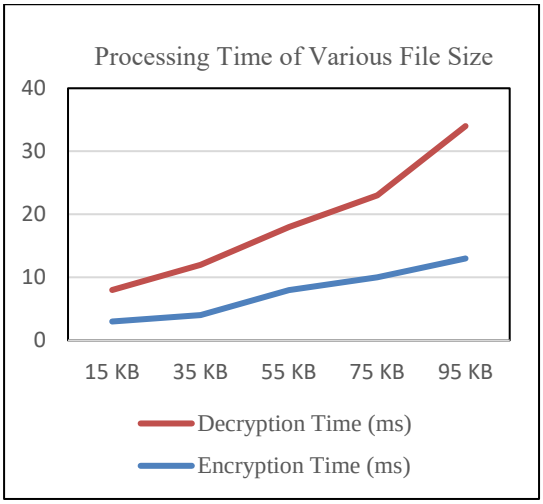



**Figure 7. Experimental Results on Encryption and Decryption Times**

According to this figure, we can conclude that the encryption time of data file is faster than the decryption time in various file sizes.

## 5. Limitation

The system can only support the cryptography portion of security of the Blowfish algorithm. The system can only support the encryption of PDF file. The system does not consider other types of files such as image or song or video files. The user must have a device that runs on application.

## 6. Conclusion

This paper presents a secure data crypto system that allows two peers to exchange encrypted data by using Blowfish algorithm. Blowfish is short program that will run on most machines and encipher safely. It is safe because of the numbers of cycles in encoding/decoding and length of key. Blowfish is a very efficient data encryption algorithm. It creates 64-bit keys, which are extremely efficient. Blowfish encryption is used today because it can encrypt and decrypt large amounts of data quickly, and it's easy to implement.

## References

[1] Anis Cherid, "Asymmetric and Symmetric Cryptography to Secure Social Network Media Communication: The Case of Android-Based E-Learning Software", Proceeding in International Research Journal of Computer Science (IRJCS), Issue 01, Volume 5, January 2018.

[2] Chatterjee, Rishav, Sharmistha Roy, and U.G. Scholar, "Cryptography in cloud computing: a basic approach to ensure security in cloud", Proceeding in International Journal of Engineering Science 11818 (2017).

[3] Chaudhari, Maulik P., and Sanjay R. Patel "A survey on cryptography algorithms" Proceeding in "International Journal of Advance Research in Computer Science and Management Studies 2", no.3, March 2014.

[4] Fathima, H., and K. S. R. Matriculation, Comparative Study of Symmetric Key Algorithms-Des, AES and Blowfish", In Proceeding of Global Journal of Computer Science and Technology ISSN: 0975-4350, Volume 17, Issue 2, 2017.

[5] G.C.Kessler "An Overiew of Cryptography" Published by Auerbach "Handbook on Local Area Networks", September 1998.

[6] J.Unni Kiran, P. Sai Kiran, "Secure Communication with Blowfish Cryptography for Data Sharing on Cloud using Android Devices", In Proceeding of International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9, Issue-6, April 2020.

[7] Parihar, Veena, and Mr Aishwary Kulshrestha, "Blowfish algorithm: a detailed study", Proceeding in International Journal for Technological Research In Engineering 3, no. 9 (2016).

[8] P. Singh, Prof. K. Singh, "Image Encryption and Decryption Using Blowfish Algorithm in MATLAB", In Proceeding of International Journal of Scientific & Engineering Research, ISSN: 2229-5518, Volume 4, Issue 7, July 2013.

[9] Sarita Kumari, "A research paper on cryptography encryption and compression techniques", In Proceeding of International Journal of Engineering and Computer Science 6, no. 4, April (2017).

[10] Vinod D. Rajput, Kajal D. Jaisinghani, "Security in Cloud Computing Using Blowfish Algorithm", In Proceeding of International Journal of Mechanical Engineering, ISSN: 0974-5823, Volume7, Special Issue 5, April-May 2022.

# Security Control in Private Message Sending System Using AES Algorithm and Secure Key Sharing with RSA Algorithm

Mar Lar Thinn, Cho Cho San
*University of Computer Studies, Yangon*
*marlar.thinn@gmail.com, chochosan@ucsy.edu.mm*

## Abstract

*The protecting of file security for sensitive information plays critical role in many organizations. It needs to be secured from unauthorized person. Confidential information disclosure to unauthorized person might lead to a serious problem or risk for organization. Therefore, the issue of confidentiality in messaging still arises. There are many techniques in cryptography for enciphering the message to be secured. Thus, the important message needs to be secured with the use of encryption algorithm first as purpose of security before sending to the receiver over internet. Hence, secured messaging system will be implemented using AES Algorithm to guarantee the content of message completely inaccessible to anyone who does not have the decryption key. Moreover, RSA algorithm will also be used to encrypt the AES key to ensure the security of key exchanged. The proposed system aims to help the government or non-government organizations to secure the confidential information from information disclosure accidentally or intentionally. This system will be implemented by using C#.Net Language on Microsoft Visual Studio 2015 with Microsoft SQL server 2017 Express Database Engine.*

**Keywords**: Cryptography, RSA, AES, confidential message

## 1. Introduction

Messaging is the exchange of messages to a messaging server with the use of files such as text file or Microsoft office documents. It is transmission of file messaging system between multiple users through Internet. The message includes text and files such as **.txt .xlsx and .doc or .docx**. Today, messages need to be secured from illegal access. Some of the data have important information that should not give access permission to unknown users instead of the authorized users. Thus, the information need to encrypt first to protect the confidentiality of original message content. It will make the message more secured before sending to the receiver through network. The authorized person can only read the message by decrypting with the cipher key. It will prevent the intruders from modifying or deleting the content of message. The original information cannot be created without possessing the key to obtain the information. The proposed system implements a system to send the information more securely between the users by using Advance Encryption Standard (AES) symmetric algorithm and RSA algorithm. In AES, this system uses symmetric keys for all 128 bits, 192 bits, and 256 bits.

The problem statements are described in Section 2 and Section 3 also discussed the recent work of message encryption systems. Section 4 highlights the theory background for encryption system. The proposed system together with performance evaluation of experiment results are described in Section 5 and conclusion and future work provide in Section 6.

## 2. Problem Statements

Nowadays, confidential information disclosure has been occurred accidentally or intentionally in many organizations. By accidentally forwarding the confidential information to unauthorized person who is not from organization might lead to a serious problem or risk. Message transmission through network will face many threats such as unauthorized access and snooping. The attacker may steal or change the information during transmission. These risks can minimize by using the encryption algorithms before transmission the information. Therefore, secured messaging is a significant issue to meet the goals of

cryptography, CIA (Confidentiality, Integrity, and Availability).

## 3. Related Work

This section describes some of the recent techniques and approaches for message encryption. A file security system is carried out by using AES algorithm. In testing the system, a trial is performed on all files with different file sizes and for the results of the encryption process (cipher text) in the form of files with the file format with the *. encrypted extension [2].

This system focuses on the systematic analysis of these issues and summarizes AES algorithm implementation, comprehensive application and algorithm comparison with other existing methods. To examine the performance of their proposed system and to make full use of the advantages of AES encryption algorithm, one needs to reduce round key and improve the key schedule, as well as organically integrate with RSA algorithm. The encryption system combining AES and RSA algorithm makes full use of the advantages of symmetric key and asymmetric key. The session key used in the file is encrypted by RSA, and the encryption of data file is encrypted by AES [5].

## 4. Background Theory

The most common two types of cryptographic encryption systems are described as followed:

**Asymmetric key encryption**: Asymmetric encryption involves two keys such as public and private keys to encrypt the information. The user needs to generate two keys, one for public key and one for private key, then user needs to send the public key to others., who then use it to encrypt data. Then the ciphertext is obtained and it is sent back to the original user. The authorized user can then decrypt the data with the private key that he/she only known. This kind of encryption algorithms are relatively slow because the mathematical operations have been applied to perform the encryption, and they are suitable for small files or data encryption.

**Symmetric key encryption**: It is also referred to secret key cryptography or single key encryption because it uses only one secret key. The secret key needs to be shared between the users. Thus, if one of the parties are cheat, the secret key might be unsecured. It is suitable for large message encryption. The well know secure encryption algorithms are 3DES, and AES algorithms.

In this system, the RSA public key crypto system has been applied to encrypt the AES key for secure key transmission.

### 4.1. AES (Advanced Encryption Standard)

It is one of the most secure algorithms for symmetric key encipherment. It is fast in both software and hardware implementation. The design and strength of all key lengths of the AES algorithm are sufficient to protect confidential information. It is a block cipher with the block size of 128 bits using 128-, 192- and 256-bits key size. It uses substitution box (S-box) and permutation box (P-box), which is a series of mathematical operations.

The details processes of AES algorithm are described in Figure 1 and Figure 2 as below:

1) Key Expansion
2) Initial round
    a. Add-Round-Key
3) Nr -1 Round
    a. Sub-Bytes
    b. Shift-Rows
    c. Mix-Columns
    d. Add-Round-Key
4) Final Round
    a. Sub-Bytes
    b. Shift-Rows
    c. Add-Round-Key

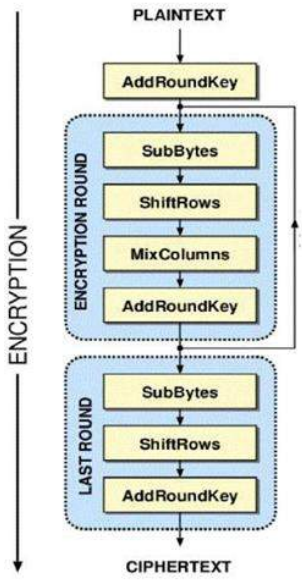



**Figure 1. AES Encryption**

The number of rounds (Nr) in AES are 10 rounds for 128 bits, 12 rounds for 192 bits, and 14 rounds for 256 bits. Each round contains

- Sub Bytes,
- Shift Row,
- Mix Column and
- Add Round Key processes.

Safety aside, AES encryption is very appealing to those who work with it. Why? Because the encryption process of AES is relatively easy to understand. This allows for easy implementation, as well as really fast encryption and decryption times. In addition, AES requires less memory than many other types of encryptions (like DES), which makes it a true winner when it comes to choosing your preferred encryption method. Finally, when an action requires an extra layer of safety, you can combine AES with various security protocols like WPA2 or even other types of encryptions like SSL.

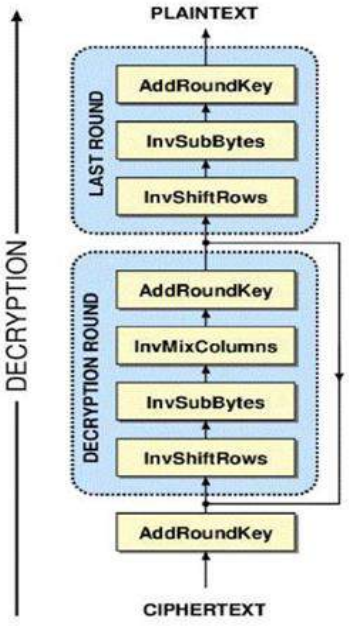



**Figure 2. AES Decryption**

### 4.2. RSA Cryptosystem

**R**on Rivest, Adi **S**hamir, and Leonard **A**dleman (RSA) are the creators of the RSA cryptosystem. RSA cryptosystem is an asymmetric key or public key cryptography system. It can be used for both public key encryption and digital signature. The key generation for public and private keys, encryption and decryption are performed in RSA. Prime numbers are used for key generation process in RSA encryption method. In RSA, p and q must be at least 512 bits; n must be at least 1024 bits.

RSA's security is the assumption that factoring a big number (n into p, and q) is hard. Generation of random prime numbers give the algorithm extra strength and efficient. RSA is useful for short messages. Public key encryption system solves the problem of key distribution.

RSA security relies on the computational difficulty of factoring large integers. As computing power increases and more efficient factoring algorithms are discovered, the ability to factor larger and larger numbers also increases.

Encryption strength is directly tied to key size. Doubling key length can deliver an exponential increase in strength, although it does impair performance. RSA keys are typically 1024- or 2048-bits long, but experts believe that 1024-bit keys are no longer fully secure against all attacks. This is why the government and some industries are moving to a minimum key length of 2048-bits.

## 5. The Proposed System

Nowadays, information transmission through unsecure network is the most important concerned for every organization. To ensure the information is secured while sending through unsecure network, various techniques can be used. One of the common techniques used widely is cryptographic technique. By taking into account the extent to which the data contained in the private information can be misused (whether working online or offline) providing security, both to online as well as offline email usage is of prime importance. Thus, protecting the data contained in the private information is important for every organization. Therefore, this system proposes the secure information system for government and non-government organizations.

### 5.1. Encryption at Sender

This section describes the information encryption process at the sender. The proposed system flow of encryption system for sender side is described in Figure 3.

Firstly, the user needs to register for user identification process and checks the user is valid

person or not. If the user is authorized, the system will let the user login and then perform the encipherment process. The input message can be text and files such as .txt, document files such as .doc or .docx and excel files.

In the key generation process, the AES keys are generated by using the hexadecimal format. The numbers of HEXA characters are 32 hex characters for 128 bits, 48 hex for 192 bits, and 64 hex for 256 bits in AES. In RSA, the private key and public key are also generated with 1024 bits. Then, the message is encrypted by using AES key and the AES key is also encrypted by using the RSA public key.

Finally, and the encrypted AES key and cipher text are sent to the receiver.

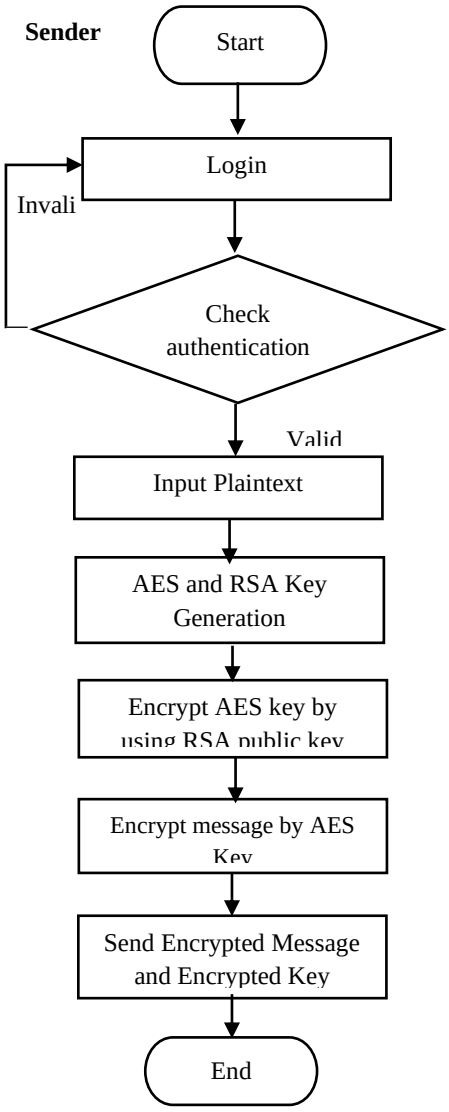



**Figure 3. The Sender System Flow**

## 5.2. Decryption at Receiver

This section describes the information decryption process at the receiver side. The proposed system flow of encryption system for receiver side is described in Figure 4.

At the receiver, the user also needs to register for user identification process and checks the user is valid person or not. If the user is authorized, the system will let the user login and then perform the encipherment process. The received encrypted message and encrypted AES key are needed to decrypt for receiving the original information.

The RSA private key has been used to decrypt the encrypted AES key. Then, the AES key is obtained from the RSA decryption. The receiver now performs the decryption for obtaining the original information by using the AES key. The secure information transmission can be performed by using the proposed system between the senders and receivers.

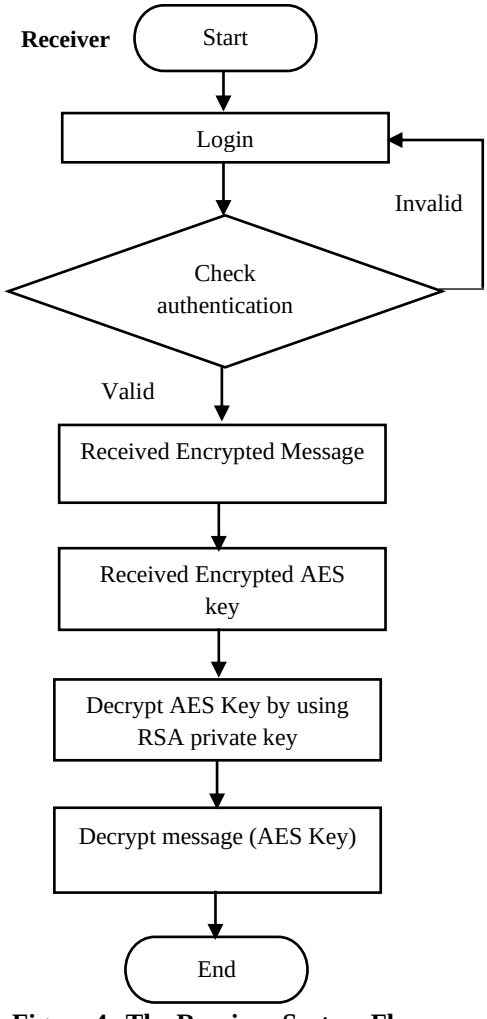



**Figure 4. The Receiver System Flow**

### 5.3. Experiment Result

In this system, before sending/ receiving message, authentication process has been performed between sender and receiver. Then if the authentication checking process is succeed, the user can perform the encipherment process and send the encrypted information to receiver as shown in Figure 5. Then, the sending message encryption of the proposed system is as shown in Figure 6 and the AES's key is encrypted by RSA for secure key sharing is shown in Figure 7.

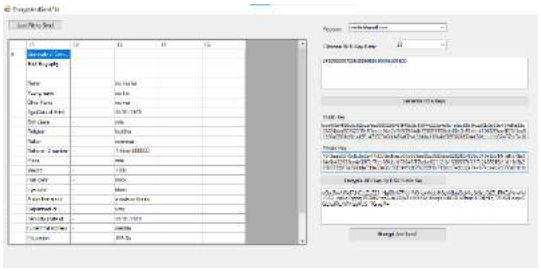


**Figure 5. Message Encryption by AES**

In Figure 5, the encryption process is proceeded by using 128 bits AES key. After encryption process is success, the AES's 128 bits key is encrypted by RSA encryption algorithm as shown in following Figure 6.

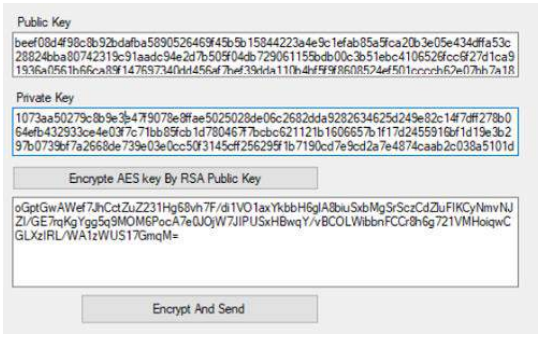



**Figure 6. Encrypt AES Key by RSA public key**

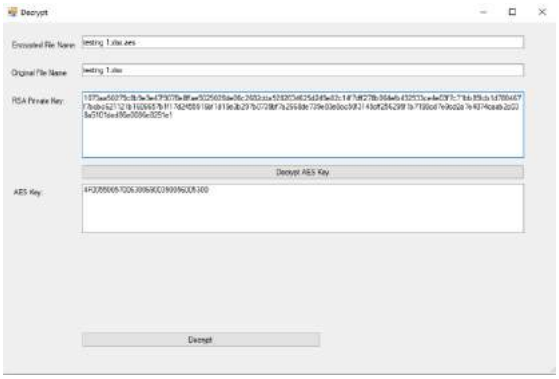


**Figure 7. Decrypt AES Key by RSA private key**

**Figure 8. Message Decryption by AES**

### 5.4. Performance Evaluation

This proposed system evaluated the system performance by using the two important properties of secure cipher text in cryptography such as confusion and diffusion.

Confusion property hides the relationship between the cipher text and the key. Confusion property is if one bit of the key is modified, the cipher text is also modified.

Diffusion property hides the relationship between the plain text and cipher text. Diffusion property is if one bit of the plain text is modified, the cipher is also modified.

This proposed system provides the confusion property because the keys for RSA and AES are generated randomly. Therefore, the key generation process supports the randomness to secure the message. Moreover, due to the randomness of the keys, the same plaintext will produce the different ciphertext. Therefore, this system provides the diffusion property of the cryptography. The proof of performance evaluation is described in the below Figure.

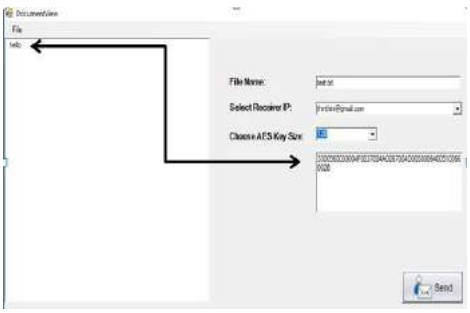


**Figure 9. (a) Same Plaintext with different 128 bit key**

**Figure 9. (b) Same Plaintext with different 128 bit key**

In Figure 9(a) and 9(b); parallel executions are made for same plain text “hello” with same length of different keys (128 bits). As result, the system generates two different ciphertexts. This proves that the system satisfied key randomness by supporting the confusion and diffusion properties.

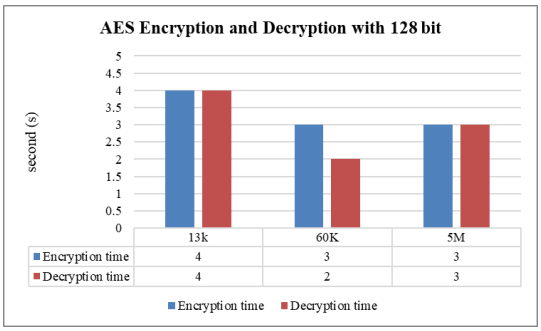



**Figure 10. Analysis of time consuming for 128 bits**

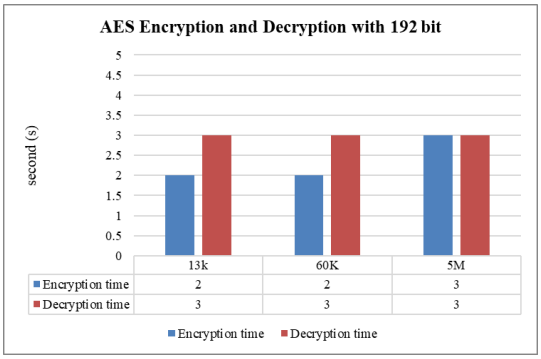



**Figure 11. Analysis of time consuming for 192 bits**

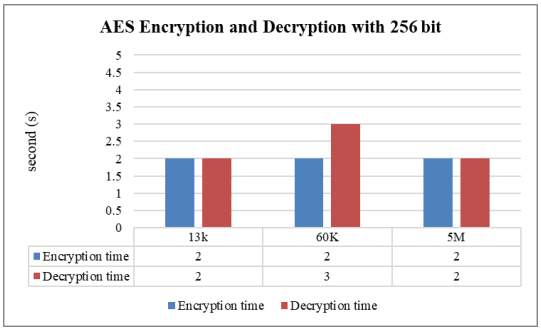



**Figure 12. Analysis of time consuming for 256 bits**

Figure 10, 11 and 12 show the encryption and decryption time analysis is also performed on different key sizes with different file sizes. The above Figures show the encryption and decryption time comparison for 128,192 and 256 bits key size with three different file sizes.

## 6. Conclusion

Confidential information disclosure to unauthorized person might lead to a serious problem or risk for organization. The secure information transmission system is needed to implement because it is vulnerable to attacks. Thus, it has now become play vital role to encrypt the information before sending it. Therefore, this system proposes a secure information transmission system by using AES and RSA algorithms. Then the AES key is encrypted by using RSA to get more benefits due to the hybrid process of symmetric and asymmetric cryptography. It can be secured the information to prevent accessing the sensitive data from unauthorized people or intruders. To provide the better performance and security, different encryption algorithms and key exchange algorithms will be applied in the future work. This encryption system process can be used in government and non-government sectors to support the information security during transmission.

## References

[1] AtulKahate, “computer –based symmetric key cryptographic algorithm”, in cryptography and Network Security, 3th Ed. New Delhi McGraw-Hill, pp.130-141.

[2] Khairul Muttaqin, Jefril Rahmadoni, “Analysis And Design of File Security System AES (Advanced Encryption Standard) Cryptography Based”, Journal Of Applied Engineering And Technological Science Vol 1(2) 2020.

[3] Nikhil An, “Using AES Algorithm Encryption and Decryption of Text File, Image and Audio in Openssl and Time Calculation for Execution”, (School of Electronics Engineering, VIT University, Chennai, India), IOSR Journal of Computer Engineering 2020.

[4] Sourabh Singh, Anurajjain ,(2013, May).”An Enhanced Text To Image Encryption Technique using RGB substitution and AES” , International journal of Engineering Trends and Technology (IJETT) volume -4,issue-5,pp.2108-2112.

[5] Zhimao Lu, Houmed Mohamed, “A Complex Encryption System Design Implemented by AES”, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Journal of Information Security, 2021.

# Security Analysis for ARP Cache Poisoning Attacks Using DS-ARP and S-ARP

Khing Shwe Ye Phu, Tin Tin Htar
*University of Computer Studies, Yangon*
*khaingshweyephu@ucsy.edu.mm*

## Abstract

*Computers employ the Address Resolution Protocol (ARP) to map logical addresses (IP) to physical addresses (MAC). However, because ARP is a stateless, all-trusted protocol, it is vulnerable to a variety of ARP cache poisoning attacks, including Man-in-the-Middle (MITM) and Denial of Service (DoS) attacks. The appeal of using computers to share sensitive data is diminished as a result of these weaknesses, which lead to security breaches. This system provides an explanation of ARP, lists numerous potential ARP cache poisoning attacks, and provides thorough information on several attack scenarios in networks with wireless hosts. Since all of these forms of attacks can be handled by DS-ARP, it also offers a feasible solution. The proposed system is compared with S-ARP protocol according to the detection time.*

**Keywords**: ARP, DoS, Man-in-the-Middle, S-ARP, DS-ARP

## 1. Introduction

The Network layer is where the Address Resolution Protocol (ARP) resides. Each PC in a LAN has a physical (MAC) address as well as an effective (IP) address. The MAC address of the destination machine is normal by the source machine in order to transmit something explicitly from one system to another in the same or different network(s). It is therefore assumed that an arrangement has been made between the IP address and the MAC address in order to obtain the actual MAC address of the source if it is missing from the ARP hold of source. ARP is thus employed. From this, it will be clear that ARP is a stateless protocol and a crucial component of the association layer.

## 2. Related Work

PCs design authentic addresses (IP) to genuine addresses using the Address Resolution Protocol (ARP). ARP can withstand such a wide variety of attacks and is also a helpful design. It is a state-ful display that reduces the potential effects of various ARP attacks by maintaining the Request frame information in the ARP store. By distributing ARP Reply frame around the organization and taking care of pertinent areas in the ARP hold each time correspondence occurs, it is more practical and secure [2].

Undoubtedly, one of the fundamental components of the Internet and the majority of IP networks is the Domain Name System (DNS). Despite the fact that the Domain Name System is enormous, only a small percentage of people have even heard of it. Information confirmation is necessary in the majority of DNS trades. Given its crucial role, DNS is trapped in sophisticated Internet attacks that target both the actual system and other Internet resources. This structure impedes DNS and has flaws, as well as several attacks on the DNS framework [3].

## 3. Background Theory

By delivering a fake ARP reply, an attacker can fraudulently alter the mapping of an IP address to its corresponding MAC address in another host's ARP cache. This technique is known as ARP cache poisoning. Therefore, this method is also known as ARP spoofing. The attacker in the following figure is Host C. By sending a faked ARP reply to Host A stating that IP address of Host B maps to MAC address of Host C and a spoofed ARP reply to Host B stating that IP address of Host A maps to MAC address of Host C, it carries out the ARP Cache Poisoning attack. Since ARP is a stateless protocol, replies are not compared to open requests. A malicious host is vulnerable to the

ARP cache poisoning attack, which has a significant impact on same network.


**Figure 1. ARP Cache Poisoning Attack**

### 3.1. Man-in-the-Middle (MITM) Attack

When Host A and Host B's ARP caches are poisoned, Host A will forward all traffic destined for Host B to Host C. Similar to Host A, Host B will route all traffic to Host C. All of the traffic between hosts A and B can now be read by host C.


**Figure 2. Man-in-the-middle (MITM) Attack**

Host A and Host B won't even be aware that they are being attacked if Host C forwards the packets to the true destination machine after reading them. This is a Man-in-the-Middle attack, where the attacker can force traffic between two machines to go through him instead of the other ay around. Host C is the attacker in Figure 2. The traffic between two machines can be redirected to pass through Host C.

### 3.2. Denial-of-Service (DoS) Attack

A Denial-of-Service attack aims to prevent the intended users from accessing a computer resource. It typically entails the coordinated activities of one or more individuals to hinder the service's ability to operate effectively. When the attacker does not transfer the packets to the actual destination machine after reading them, it differs slightly from an MITM attack. The attacker in the following figure is Host C. After reading the packets, host C does not send them on to the destination node [4].


**Figure 3. Denial-of-Service (DoS) Attack**

## 4. Proposed System

The proposed system firstly creates an attack and send it to the host. This system detects two types of attacks; MITM attack and DoS attack by using holistic approach (DS-ARP) and S-ARP approach on poisoning host. Finally, this system compares two detection approach in execution time. Figure 4 illustrates overall system design.

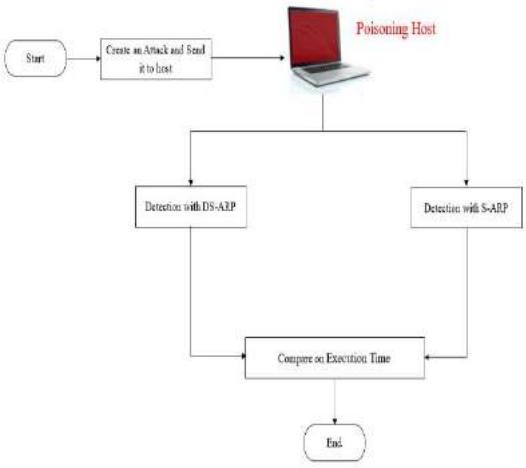



**Figure 4. Overall System Design**

### 4.1. S-ARP

To prevent ARP poisoning attacks, Secure ARP extends ARP with an integrity/ authentication method for ARP replies. S-ARP matches the original ARP specification in terms of message exchange, timeout, and cache because it is built on top of ARP. The authentication data is carried by an extra header that is put at the end

of the protocol standard messages in order to retain compatibility with ARP. Despite the fact that on a secure ARP LAN all hosts should run S-ARP.

S-ARP protocol-running hosts won't accept unauthenticated messages unless they are listed in a list of known hosts. On the other hand, hosts that use the standard ARP protocol will be able to accept messages that have been authenticated. Due to the fact that the portion running standard ARP is still vulnerable to ARP poisoning, a mixed LAN is not advised in a production environment.

Every secured host that has to communicate with an unsecured one must also be given the list of hosts not running S-ARP. Interoperability with the insecure ARP protocol is available only in exceptional circumstances and ought to be avoided at all costs. It is only meant to be used while a LAN transitions to becoming fully S-ARP enabled. Table 1 shows the procedure of S-ARP.

**Table 1. Procedure of S-ARP**

| | |
|---|---|
| Step 1 | If host wants to send to other host, source - construct - connection |
| Step 2 | If hosts - same network, S-ARP distributes certificate to all hosts and look S-ARP cache to locate MAC address of destination host |
| Step 3 | If MAC of destination - present, source - place MAC address from cache |
| Step 4 | Else source - broadcast - S-ARP Request in the network - source MAC, source IP, destination IP |
| Step 5 | Other receive S-ARP Request - send S-ARP Reply. Source - check <IP, MAC> |
| Step 6 | If <IP, MAC> in S-ARP - valid, source -store <IP, MAC> of destination - from S-ARP Reply to S-ARP cache. Source - send message - destination, destination - receive message. Else S-ARP Reply -being hacked |

## 4.2. DS-ARP

The proposed detection scheme for ARP spoofing attack, known as the DS-ARP detection method, uses a routing trace. The agent side and server side of the proposed scheme's architecture can be separated.

The two main technologies at play are detection and protection, as shown in Figure 5. The updated condition of the ARP cache table is periodically monitored by detection. The DS-ARP runs a routing trace to find corresponding ⟨IP, MAC⟩ pair information whenever the ARP cache table is updated. It alerts the server and starts the protection procedure if an ARP spoofing attack is thought to have occurred. Additionally, the appropriate ⟨IP, MAC⟩ pair ARP type is changed from dynamic to static. The procedure of DS-ARP will show in Table 2.

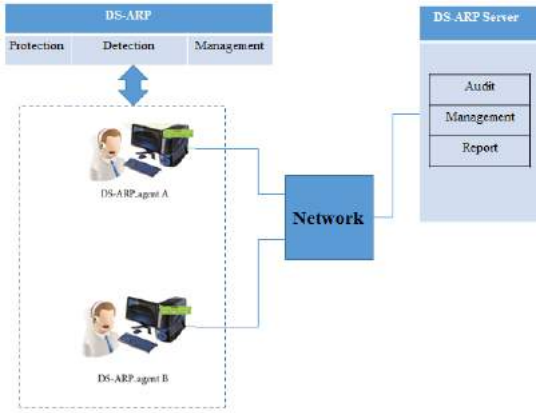



**Figure 5. The System Overview View**

**Table 2. Procedure of DS-ARP**

| | |
|---|---|
| Step 1 | If host send to other, source - construct - connection |
| Step 2 | If hosts - same network, source - look ARP cache to locate MAC of destination |
| Step 3 | If MAC of destination is present, source - place its MAC - from cache to message frame |
| Step 4 | Else source - broadcast ARP Request in network - contains source MAC, source IP, destination IP |
| Step 5 | Other hosts receive - ARP -send ARP Reply. Source - check <IP, MAC> |
| Step 6 | When ARP cache - updated, DS-ARP performs a routing trace - identify corresponding <IP, MAC> pair |
| Step 7 | If <IP, MAC> in ARP Reply - valid, source host will store <IP, MAC> of destination - from ARP Reply to ARP cache. Source - send message - destination and destination host - receive message. Else ARP Reply - being hacked. |

The detection module checks modified entries and periodically keeps the ARP cache table. The DS-ARP uses a routing trace to ascertain whether an ARP spoofing attack has occurred after identifying a change in the ARP cache table.

The protection module converts the previous state of the ⟨IP, MAC⟩ pair information that was altered by the ARP spoofing attack in the ARP cache table list. By switching the link type from a dynamic state to a static state, it avoids ARP spoofing attacks [6]. The sequence diagram of system is shown in Figure 6.

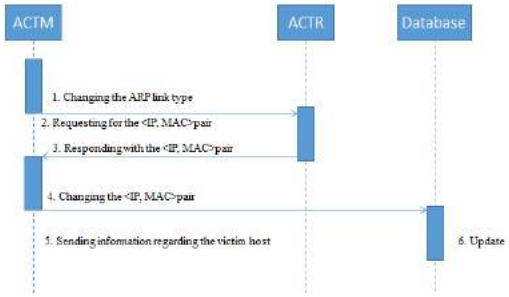



**Figure 6. Sequence Diagram**

## 5. System Implementation

The proposed system implements on sender, receiver and attacker. Two types of attacks, MITM attack and DoS attack will be detected by DS-ARP and S-ARP. This system performs three steps of processes; (1) Broadcast ARP request from sender to receiver, (2) Request Reply form receiver to sender and (3) Detect with DS-ARP and S-ARP when attack occurs.

### 5.1. ARP Request and Receiver's Reply

Before sending ARP Request Frame to the receiver, sender will check the receiver's MAC in its ARP cache. If receiver's MAC is not existed, the request frame must be sent to all members by sender as shown in Figure 7.

**Figure 7. Sending ARP Request to All Members**

The receiver can check the APR Request message in its APR cache after the receiver login authentication is complete. The sender's IP and MAC addresses will be located in the receiver's ARP cache, as shown in Figure 8.

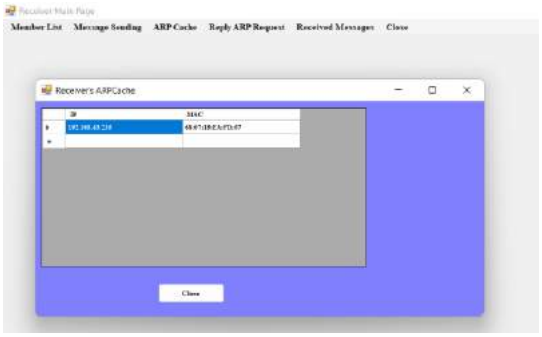


**Figure 8. Check Sender's IP and MAC Addresses in Receiver ARP Cache**

The receiver generates a certificate to prevent ARP spoofing when responding to the sender's ARP request. According to Figure 9, the certificate has the following information: Magic, Type, Sig Len, MAC address, IP address, Issue Timestamp, and Signature. The generated certificate is then returned to the sender by the receiver.

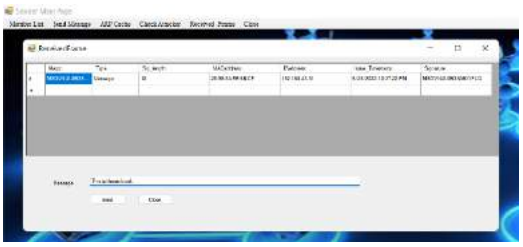


**Figure 9. Sender Check Certificate and Send Message to Receiver**

Figure 9 shows how the sender checks the message before sending it after getting the receiver's certificate. The receiver will view the message as shown in Figure 10 when it is sent. The sender's IP will be included in the message and associated to the receiver.

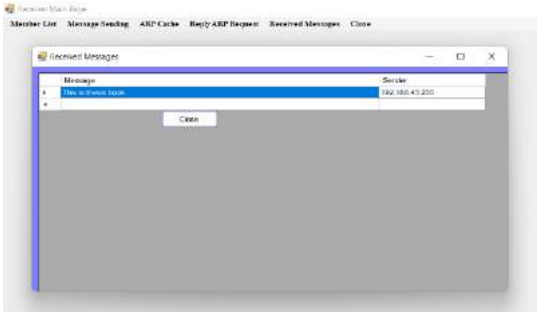


**Figure 10. Receiver Receive Sender's Message**

### 5.2. Replying ARP Request by Attacker to Sender

The attacker login with its IP and MAC addresses. Then, it replies the request of sender as a receiver in Figure 11 and Figure 12.

**Figure 11. MITM Attacker Reply**

**Figure 12. DoS Attacker Reply**

### 5.3. Detecting with D-SARP and S-ARP

When MAC address does not match IP address of receiver, the attack occurs. However, the sender, which detects with DS-ARP and S-ARP approaches, does not accept the reply of attacker and displays notification of attack as shown in Figure 13.

**Figure 13. Noticing the Attacker by Sender**

## 6. Experimental Results

The proposed system provides the execution time when MITM attack and DoS attack will be detected by DS-ARP and S-ARP. The detection time comparisons on ten tests are shown in Figure 14 and Figure 15. Moreover, Table 3 shows the comparisons of DS-ARP and S-ARP detection on existing solutions.

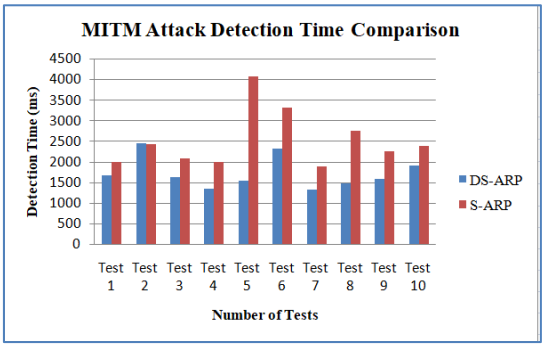



**Figure 14. MITM Attack Detection Time Comparison**

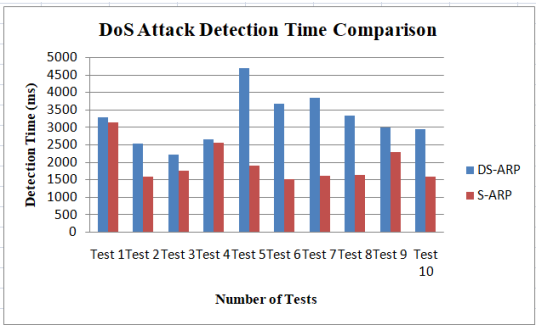



**Figure 15. DoS Attack Detection Time Comparison**

**Table 3. Comparison of Defense Methods of ARP Cache Poisoning Attack**

| Existing Solution | S-ARP | DS-ARP |
|---|---|---|
| Cryptography used | Yes | No |
| Hosts on network | Trusted Host Authoritative Key Distributor (AKD) | DS-ARP Agents |
| New device added to network | Not Allowed | Not Allowed |
| Switches | Not Allowed | Not Allowed |
| Performance Degradation | High | Not Allowed (modified ARP) |
| Mechanism | Signed ARP replies | Stateful protocol and broadcast both ARP request and reply with centrally control of DS-ARP server |

## 7. Conclusion

All of the advantages of the ARP are still present, but its security flaws are closed off. This system provides the detection time in DS-ARP with the comparison of S-ARP. The faster the detection time, the more secure the system. According to the experimental results, DS-ARP is more secure than S-ARP in detecting MITM attack and S-ARP is more secure than DS-ARP in detecting DoS attack.

## References

[1] Corey Nachreiner, "Anatomy of an ARP Poisoning Attack", WatchGuard Life Security Service, 2003. http://www.watchguard.com/ infocenter/editorial/135324.asp.

[2] A Md. Ataullah, Naveen Chauhan, "An Efficient and Secure Solution for the Problems of ARP Cache Poisoning Attacks", World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering, 2012.

[3] Hin Yeung Lo, "Executing Defense System Of Dns Hijacking And Cache Poisoning Attacks In The Domain Name System", Curtin University of Technology, November 2005.

[4] Md. Ataullah, Naveen Chauhan, “An Efficient and Secure Solution for the Problems of ARP Cache Poisoning Attacks”, World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering 2012.

[5] D. Bruschi, A. Orgnaghi, and E. Rosti, “S-ARP: a secure address resolution protocol”, Dipartimento di Informatica e Comunicazione Universita` degli Studi di Milano, Italy 2013.

[6] Min Su Song, “DS-ARP: A New Detection Scheme for ARP Spoofing Attacks Based on Routing Trace for Ubiquitous Environments”, Department of Computer Science and Engineering, Seoul Tech, Republic of Korea, The Scientific World Journal Volume 2014.

# Securing File Sharing Using AES-CBC Authenticated Encryption

Chan Myae Thu, Amy Tun
*University of Computer Studies, Yangon*
*chanmyaetthu@ucsy.edu.mm, amytun@ucsy.edu.mm*

## Abstract

*Today people are widely used internet, electronic records because of their ease of alteration and fast transition. The society is becoming more and more digitalized – therefore Information security is becoming more important than ever. The need for everyone to identify themselves in a digital way has spawned a wide variety of challenges, such as, for example, how to avoid fraud. Data security is main topic while transferring data from one place to other for protection of data from unintended user. Cryptography is essential for data security. Data encryption is an easy means of securing personal or business data protection. Many secure transmission techniques require any encryption. In the proposed system, the user desire files (.doc and .xlsx) can be uploaded and encrypted by AES-CBC in secure data sharing. For secure key sharing purpose, this system will also be used ELGamal encryption algorithm to encrypt the AES-CBC's symmetric key.*

**Keywords**: Cryptography, AES-CBC, ELGamal, data sharing

## 1. Introduction

Files can be arranged into business and non-business, where noncommercial documents can be additionally characterized into private and non-secret. Business and non-business yet private records are delicate which implies that they should be safeguarded from expected assaults or abuses. Such goes after could prompt unapproved exposure (Confidentiality attacks), unauthorized modification (Integrity attacks), or unauthorized withholding (Availability attacks) [5]. These various kinds of assaults can be performed on documents while they are being moved, put away, or utilized by either approved or unapproved clients. Thus, insurances for these kinds of assaults originate from three particular fields of safety which are: Correspondence security which is worried about forestalling various sorts of assaults on information sent over an organization; Edge security which is worried about forestalling assaults on information put away inside a confided in inner organization; and Insider security which is worried about forestalling assaults on information by the people who have been approved with access [2].

In this day, a large portion of the correspondence is finished utilizing electronic media. Information Security assumes a fundamental part in such correspondence. Thus, there is a need to shield information from pernicious assaults. Cryptography is the study of mystery codes, empowering the secrecy of correspondence through an unreliable channel. It safeguards against unapproved parties by forestalling unapproved change of purpose. Cryptographic calculations are vital in data security in which information is scrambled at the shipper side and decoded at the recipient side. PC and correspondences frameworks use cryptography for three expansive purposes — to safeguard the privacy of data (i.e., encryption), to safeguard the uprightness of data, and to confirm the originator or shipper of data [6].

The High level Encryption Standard (AES) calculation is a symmetric block took on by the NIST in 2001. The size of an AES block is 128 pieces, though the size of the encryption key can be 128, 192 or 256 pieces. Methods of activity may likewise give utilization of the block on a flood of plaintext and make the calculation more effective. This proposed secure document sharing framework will be executed by CBC method of AES encryption.

## 2. Related Work

In asset obliged stage, the memory necessities, power utilization and throughput are significant contemplations. This trial results show that AES-CBC accomplishes the higher security

execution contrasted with AES-ECB plot, albeit the speed of encryption debases possibly.

In "Examination and Plan of Record Security Framework AES (High level Encryption Standard) Cryptography Based", the use of AES as a document security framework is completed, where the encryption and decoding process is done on the record. In testing the framework a preliminary is performed on all records with various document sizes and for the consequences of the encryption cycle as records with the document design with the *.encrypted expansion [1].

This framework centers around the precise examination of these issues and sums up AES calculation execution, extensive application and calculation correlation with other existing techniques. To investigate the exhibition of the proposed calculation and to take full advantage of the benefits of AES encryption calculation, one necessities to diminish round key and work on the key timetable, as well as naturally coordinate with RSA calculation. The encryption framework consolidating AES and RSA calculation takes full advantage of the benefits of symmetric key and hilter kilter key. The meeting key utilized in the record is scrambled by RSA, and the encryption of information document is encoded by AES [6].

## 3. Background Theory

In this section, the essential background theory concerned with the system is discussed.

### 3.1. Advanced Encryption Standard (AES)

In AES, block are plans for encryption or decoding where a block of plaintext is utilized to get a block of code text with a similar size [1]. Today, AES (High level Encryption Standard) is one of the most involved calculations for block encryption. It has been normalized by the NIST (Public Foundation of Guidelines and Innovation) in 2001, to supplant DES and 3DES utilized for encryption in that period. The size of an AES block is 128 pieces, though the size of the encryption key can be 128, 192 or 256 pieces. In every one of the phases of encryption, four capabilities are applied: replacement of bytes, stage, number-crunching tasks over limited fields and a XOR activity with the encryption key.

The size of the AES block gives proficiency, yet additionally adequate security. Considering the figuring force of the innovation in the time of the AES normalization and the expected processing power anticipated for the future, it has been thought about that the base size for the encryption key of 128 pieces gives protection from animal power assaults [2]. Additionally, the calculation was built to be impervious to all hinder assaults which were known at that point. Block calculations ought to empower encryption of plaintext with size which is unique in relation to the characterized size of one block too. The method is introduced in [2, 3]. In particular, it is projected that adding a "1" to the plaintext is more modest than the block and adding "0's" cushioning to achieve the necessary size. Utilization a method of activity can be considered for next method.

The method of activity may likewise give use of the block on a surge of plaintext and make the calculation more productive. Then again, the method of activity might change over the block into a stream and furthermore to fortify the impact of the encryption calculation. To meet these prerequisites, in 2001 the NIST normalized five methods of activity: ECB (Electronic Code Book), CBC (Code Block Anchoring), CFB (Code Criticism), OFB (Result Input) and CTR (Counter), which apply to AES [4].

Every method of activity has its own boundaries which are vital to give the fundamental security of the calculation. The five AES methods of activity will be introduced, alongside their particular boundaries that ensure security. The show will initially contain the standards of activity of ECB, CBC, CFB, OFB and CTR methods of activity, as they are portrayed in the writing. For every method of activity, the essential boundaries, their benefits and hindrances, and furthermore their legitimate application, will be introduced. The methods of activity are the most significant for a legitimate AES execution, no matter what its product or equipment execution. An ill-advised execution or utilization of the methods of activity may truly compromise the AES calculation unwavering quality and lead to divulgence of a section or all of the plaintext [6].

The proposed system will be used AES-CBC mode for file encryption.

### 3.2. CBC Method of Activity

To give cryptographic safety, each encryption of the equivalent plaintext ought to result with an alternate code text. The CBC (Code Block Binding) method of activity gives this by utilizing an instatement vector – IV which has the very size as the block that is encoded. Initial, plaintext block (P1) and IV are XORed. Then, an encryption is continued with the key (K). Then, at that point, the aftereffects of the encryption performed on each block (C1, C2, …, CN-1) is utilized in a XOR activity of the following plaintext block PN which brings about CN. Along these lines, when indistinguishable plaintext blocks are scrambled, an alternate outcome is gotten. Likewise, involving an alternate IV for each new encryption, an indistinguishable message will constantly be encoded in an unexpected way. It ought to be underlined that a similar key K is utilized in every one of the encryption blocks.

### 3.3. ELGamal Crypto System

ELGamal encryption calculation was depicted by Taher ELGamal. ELGamal encryption is utilized in the free GNU Security Gatekeeper programming, late renditions of PGP, and other cryptosystems. ELGamal encryption comprises of three parts: the key generator, the encryption algorithm, and the decryption algorithm.

## 4. Methodologies of Proposed System

**AES:** The AES (Advanced Encryption Standard) algorithm is an algorithm for symmetric key encryption. The AES algorithm uses 128-bit, 192-bit, and 256-bit keys of varying lengths. The AES algorithm treats every 128 bits of blocks into a 16 byte segment. Every 16 byte segment gets settled as a 4 and 4 bytes matrix. Number of rounds involved can be controlled by length of the key. It must divide into several blocks if plaintext's length is larger than the block size. To match the block size, he last block of the plaintext must be padded [5].

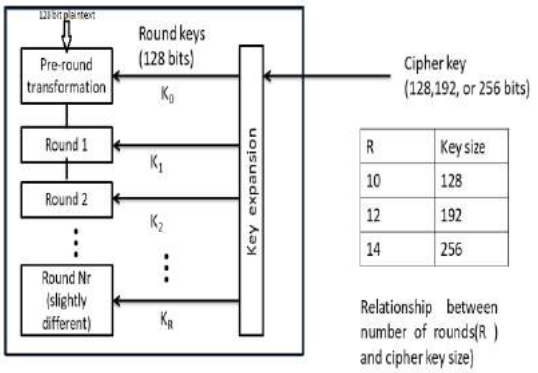



**Figure 1. AES algorithm grouping and encryption diagram**

AES is a substitution-permutation method, a sequences of mathematical processes that use substitutions (S-Box) and permutations (P-Boxes).

Steps of AES algorithm:

1) Key Expansion
2) Initial round
    a. Add-Round-Key
3) Nr -1 Round
    a. Sub-Bytes
    b. Shift-Rows
    c. Mix-Columns
    d. Add-Round-Key
4) Final Round
    a. Sub-Bytes
    b. Shift-Rows

In AES [2], variable number of rounds (Nr) are used according to relationship. During each round, above operations are applied as shown in Figure1.

**Sub-Bytes:** This process is a nonlinear byte substitution using S-box, and changes the byte values.

**Shift-Row:** every row in the 4x4 array is shifted a certain amount to the left depending on the row index.

**Mix-Column:** In this step, mathematical function transforms the values of a given column within a state, acting on the four values at one time.

**Add-Round-Key:** each byte of the state is combined with a round key, a different key for each.

**Cipher block chaining (CBC) Mode:** The CBC method of activity gives the cryptographic security by utilizing an introduction vector-IV. IV has the very size as the block that is scrambled. As a general rule, the IV is normally an irregular

number. In CBC mode, when same plaintext blocks are encoded, an alternate code text blocks are gotten. Likewise involving an alternate IV for each new encryption, an indistinguishable message will continuously be encoded in an unexpected way. If a plaintext or Figuretext block is broken, it will influence all following block. CBC mode is utilized in numerous applications like email or web information. In Figure 2 (a), initial a XOR activity is applied to the plaintext block (P1) with the IV, and afterward an encryption with the key (K) is performed. Then the consequences of the encryption performed on each block (C1, C2, …, CN-1) is utilized in a XOR activity of the following plaintext block PN which brings about CN. Decryption in the CBC Mode is shown in Figure 2 (b).

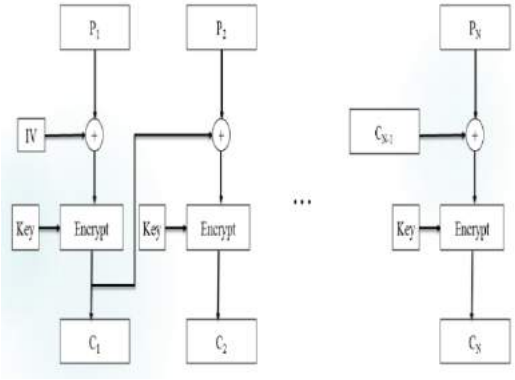



**Figure 2. (a) Encryption in the CBC Mode**

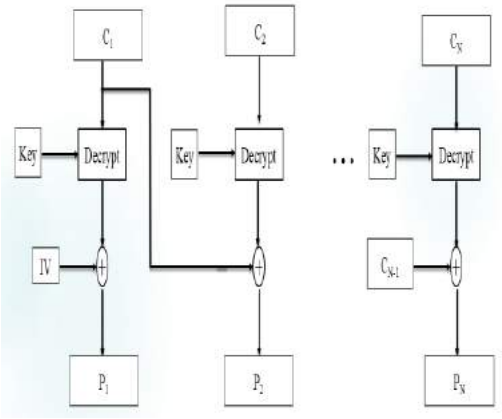



**Figure 2. (b) Decryption in the CBC Mode**

### 4.1. Initialization Vector (IV)

In cryptography, an initialization vector (IV) or starting variable (SV) is a contribution to a cryptographic crude being utilized to give the underlying state. The IV is commonly expected to be arbitrary or pseudorandom, yet now and again an IV just should be capricious or remarkable. Randomization is vital for some encryption plans to accomplish semantic security [3]. This framework will utilized the pseudorandom (Linear Congruential Generator) as IV.

### 4.2. Parts of ELGamal Crypto System

Parts of ELGamal Crypto System are the key generation process, the encryption process, and the decryption process.

**ELGamal Cryptosystem – Key Generation**

{
Select a large prime p;
Select d to be a member of the group G = < $Z_P^*$, X > where 1<= d<= p-2
Select $e_1$ to be a primitive root in the group G = < $Z_P^*$, X >
$e_2 \leftarrow e_1^{\,d} \bmod p$
Public_key ← ($e_1$, $e_2$, p)
Private_key ← d
return Public_key and Private_key
}

**ELGamal Cryptosystem Procedure**

**ElGamal_Encryption** (e1, e2, p, P) // P is the plaintext
{
Select a random integer r in the group G = < $Z_P^*$, X >
$C_1 \leftarrow e_1^{\,r} \bmod p$
$C_2 \leftarrow (P \times e_2^{\,r}) \bmod p$ // $C_1$ and $C_2$ are ciphertexts
return $C_1$ and $C_2$
}

**ElGamal_Decryption** ( d, p, $C_1$, $C_2$ )
{
$P \leftarrow [C_2 (C_1^{\,d})^{-1}] \bmod p$
return P
}

## 5. The Proposed System

In this system, firstly, authentication between sender and receiver is done before sending data. If authentication is successful, the user can continue encryption data and send it to receiver. The sender can create or choose the attach .doc / .xlsx file to send to the receiver. In the encryption

and decryption phase, this system will used AES-CBC mode. In the proposed system, 256 bit key size of AES algorithm is used to do the evaluation of the operation mode of CBC. The AES-CBC key is encrypted by Elgamal for secure key sharing. The flowchart of the system is shown in Figure 3.

In messaging, many transactions and important information transmission are transferred. Therefore, protecting the data is contained in private messaging. Message encryption and sending encrypted message processes of the system are shown in Figure4. AES-CBC's key is encrypted by ELGamal for secure key sharing is shown in Figure 5.

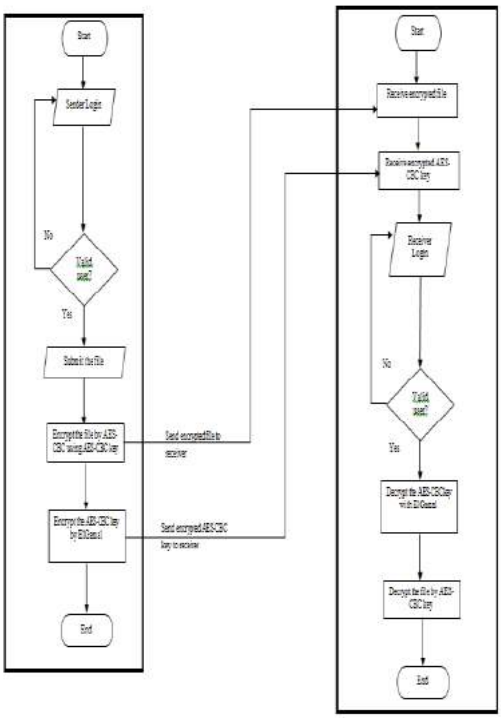



**Figure 3. The System Flow**

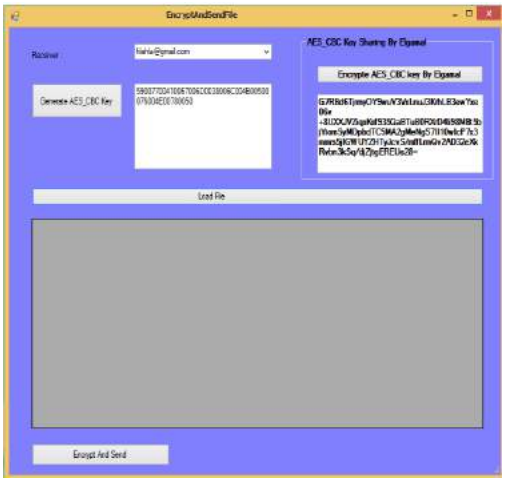



**Figure 4. Message Encryption By AES-CBC**

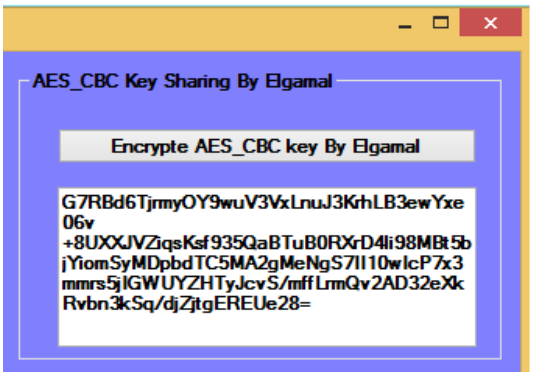



**Figure 5. AES-CBC Key Encryption by ELGamal**

## 5.1. Discussion and Analysis

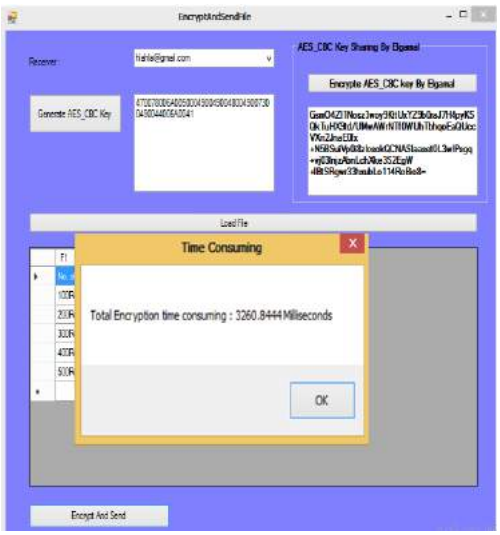



**Figure 6. Time Consuming Monitoring**

This section will discuss the secure data sharing in time consuming analysis point of view. This system will explore each execution time as shown in Figure 6.

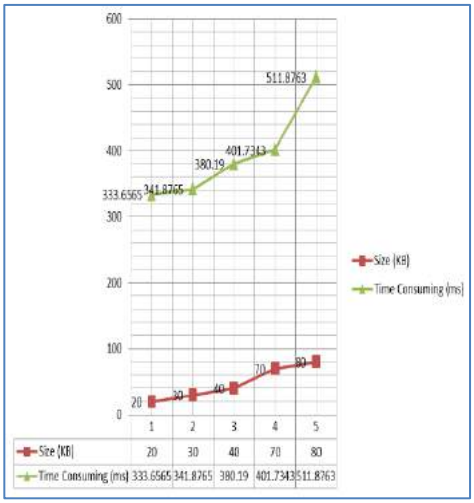



**Figure 7. a detailed time-consuming analysis of AES-CBC-ElGamal (Encryption)**

Different sizes of files are also processed and analyzed the various time values as shown in Figure 7. In Figure 8, five different sizes of files are tested and time consuming is shown in graph.

Based on Figure 7 and Figure 8, encryption and decryption time consuming are not quite different and times within the reasonable and acceptable to use the system for secure file sharing.

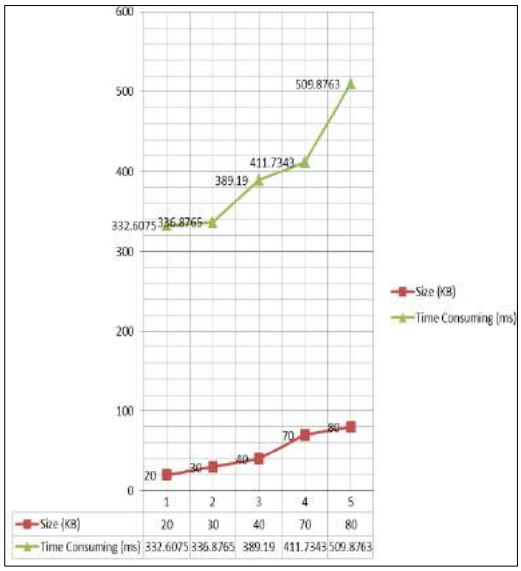



**Figure 8. a detailed time-consuming analysis of AES-CBC-ElGamal (Decryption)**

## 5. Conclusion

In the proposed system, a detailed time-consuming analysis of AES-CBC-ElGamal is presented in terms of encryption time, decryption time and throughput for .doc / .xlsx file encryption. The mode of operation in the proposed system is CBC which provides application of the block cipher on a stream of plaintext and make the algorithm more efficient. Key sharing process is provided by applying ElGamal. The proposed system can provide a service for securing file sharing for confidential data.

## References

[1] D. B. A. Bozhinovski Pachovski," MODES OF OPERATION OF THE AES ALGORITHM ",10th CIIT 2013.

[2] K. Muttaqin, J. Rahmadoni, "Analysis and Design of File Security System AES (Advanced Encryption Standard) Cryptography Based", JAET, Vol 1(2) 2020.

[3] N. An, "Using AES Algorithm Encryption and Decryption of Text File, Image and Audio in Openssl and Time Calculation for Execution", IOSR JCE, 2020.

[4] N. Dworkin. "Recommendation for Block Cipher Modes of Operation, Methods and Techniques". NIST Special Publication 800-38A Edition2001.

[5] S. Almuhammadi and I. Al-Hejri, "A Comparative Analysis of AES Common Modes of Operation", 2017 IEEE 30th CCECE.

[6] W. Stallings. Cryptography and Network Security: Principles and Practices. Fourth Edition. NJ, Prentice Hall, 2015

# Secure Educational Data Management using Lattice-Based Access Control

Yie Yie Nwe, Nilar Aye
*University of Computer Studies, Yangon*
*yieyie.nwe@ucsy.edu.mm, nilaraye@ucsy.edu.mm*

## Abstract

*Database security is the system, processes, and procedures that protect a database from unintended activity. The insurance of information stockpiling is a difficult and considerable errand thus the client information ought to be safeguarded against security dangers. In this framework, the client's information is safeguarded alongside the relationship of cross section-based security strategy. The proposed framework is produced for secure information access control in Training Degree School (EDC). This situation is expected to give right access control in view of client's jobs that are allocated by the undertaking's arrangement choice by utilizing Grid Put together Access Control Model with respect to information and characteristics of EDC's understudies. In the proposed framework, executive can get to all information and can make all exchange of the entire framework and the information control of the particular level. The users of the proposed system are Admin User (level-1), Department Head (level-2), Senior Teacher (level-3), Teaching Staff (level-4) and Student Affair (level-5). These levels are defined by system administrator or board of the organization. The assigned categories of the admin on the object access are 1) essential data submission, 2) Data Management (Limited by System Rules), and 3) user management (Subject management).*

*This system is implemented using C# programming language with Microsoft SQL server database engine.*

**Keywords**: EDC, Lattice-Based Access Control, database security.

## 1. Introduction

Due to the rapid development of Computer and Internet technology, an ever-increasing number of resources of an organization or an association are put away in computerized design in data set. Data sets are additionally broadly utilized in each individual's day to day routine of each and every association. These associations are danger to open the data set frameworks, the techniques to be thought about while getting a data set, and how to get a data set in various honesty levels of significant layers. The proposed framework will control the protected information access on understudies' instructive information of Training Degree School. This framework will be created as a protected information access control framework using Lattice-Based Access Control in Education Degree College. Lattice based access control provides Authorization and prevents authorization attack. In addition to providing access control, lattice-based access control is utilized to safeguard against unauthorized disclosure, modification, and availability. The data classification framework has defined these kinds of activities.

An access control policy with various rights for the organization's users is created in this system based on the roles of the organization. This system's primary function is to implement an access control policy that restricts access. In order to grant restricted access to the data, a lattice model has been constructed and an access control policy has been established based on it.

## 2. Related Work

Teacher Steve Demurjian Fall Jin Ma, portrayed Mandatory Access Control in Patient DB utilizing CORBA, Application predefines the security game plan (T, S, C, U) for resource, organization and procedure. The security access control levels are organized by climbing request (from lower to upper) concerning their entrance allowed. Security alludes to the assurance of information against unapproved divulgence, change, or obliteration [7].

The use of staggered security, portrayed by lattice-based responsiveness profiles, to ensure consistence with data access impediments between structures. This security approach obliges the complexities expected for prosperity data access and benefits from existing, exhibited instruments that are used for defend and public wellbeing applications [6].

Admittance to the information objects is constrained by access control approaches which are put away as strategy objects [1]. A climate where in different sort of exchanges are executed simultaneously; a portion of these might be exchanges refreshing strategy objects. Refreshing strategy objects while they are sent can prompt potential security issues. This proposed calculations that forestall such security issues, yet additionally guarantee serializable execution of exchanges. The calculations vary on the level of simultaneousness gave and the sorts of approaches each can refresh.

## 3. Motivations

It becomes difficult to guarantee the confidentiality and integrity of user data on an EDC system when security at the other end is compromised. Thus, to defeat the security issues, information should be done the approval cycle prior to putting away it. In this framework, an adaptable and powerful information security conspire is proposed to safeguard client information in light of access control strategy (Lattice Based).

## 4. Background Theory

There are two basic types of access control mechanisms used to protect information from unauthorized access in the database systems. Discretionary Access Control (DAC) and Mandatory Access Control (MAC). MAC strategy thinks about the awareness level at which the client is working to the responsiveness name of the item being gotten to and declines except if certain MAC checks are passed. MAC is mandatory since the marking of data happens naturally, and conventional clients can't change names except if a chairman approves them. Awareness marks are allotted to documents, gadgets, windows, hosts, organizations, and to other framework protests that client access. Chairman demonstrate the degree of trust or occupation obligation of anybody getting to the framework by doling out a freedom that sets the upper bound of a bunch of responsiveness names at which the client can work.

Manager likewise relegates a base responsiveness name that sets the lower bound. On the other hand, managers can design clients to work at a solitary mark. With compulsory control, just directors and not proprietors of assets might settle on choices that bear on or get from strategy. Just a director might change the class of an asset, and nobody might concede a right of access that is unequivocally illegal in the entrance control strategy.

MAC requires every one of the individuals who make, access, and keep up with data to adhere to guidelines set by directors. The limitations put on record control (perusing, composing, making, erasing) are those that are for the most part acknowledged while carrying out a MAC policy:

1. To read a file, the label of the process must dominate the label of the file.
2. To write or update a file, the label of the process must be dominated by the label of the file.

A cycle can make a document to the level of the mark. For instance, a client who is running an interaction at Mystery ought not be permitted to peruse a record with a mark of Top Secret. The entrance choice to understand protests and compose not entirely set in stone by an overall idea of identicalness and strength between the level of a process(subject) and the level of an item (record, catalog, and so on.). Characterizing strength is passed on to the adjusting execution, yet by and large a name "rules" another mark in the event that it is "equivalent or higher" in some characterized structure. For instance, in military terms, a name of Top-Secret rules a mark of Mystery.

### 4.1. Two methods of MAC

Two methods are commonly used for applying mandatory access control.

**Rule-based access controls:** This kind of control further characterizes explicit circumstances for admittance to a mentioned object. All MAC based frameworks execute a straightforward type of rule-based admittance ought to be conceded or denied

by coordinating with an item's responsiveness name and a subject's awareness mark.

**Lattice-based access control:** These can be utilized for complex access control choices including different items as well as subjects. A lattice model is a numerical construction that characterizes most prominent lower-bound and upper headed values for a couple of components, like a subject and an item.

### 4.2. Multi-Level Secure (MLS) DBMS

Mandatory Access Control (MAC) systems are suitable for some staggered secure applications. MLS is the utilization of a PC framework to handle data with various responsive qualities (i.e., at various security levels), license concurrent access by clients with various trusted status and has to-be aware, and keep clients from getting admittance to data for which they need approval. MLS permits simple admittance to less-delicate data by higher-cleared people, and it permits higher-cleared people to impart disinfected records to less-cleared people without any problem. A cleaned record is one that has been altered to eliminate data that the less-cleared individual isn't permitted to see.

### 4.3. Lattice Base Access Control

Lattice based access control gives the characterization of access strategy in light of the proprietor of the archive or information.

$$L = Ci * Y \quad (1)$$

where, Y is a set of additional constraints $1 \leq i \leq n$ and $C1 > C2 > C3 \ldots . > Cn$ is security values. The tasks in EDC were taken as a domain and a variety of data are available in this system. All the data defined in the system come under a data set called object O which is defined by equation (2).

$$O = Di \quad (2)$$

where $1 \leq i \leq n$

There are different roles in EDC's data access control system such as administrator, department head, senior teacher, teacher, and student affair, etc. Each role is defined as the subject S given in (3).

$$S = Ri \text{ where } 1 \leq i \leq n \quad (3)$$

The security esteem is given to archive and it takes the structure $v = Ci\ y$ where $1 \leq I \leq n$ and y contained in Y. In the event that v1 = C1 and v2 = C2, v1 "unrivaled than" v2. Based on the security esteem, the read or compose tasks are applied by various subjects. The lattice L is represented in the following diagram (Figure 1).

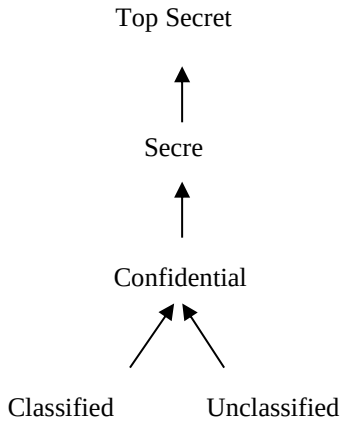



**Figure 1: Lattice Model for Proposed Method**

Each subject should make verification qualifications through the point of interaction. At the point when the subject is attempting to recover the substance, their qualifications would have been checked and in the event that the confirmation has been succeeded, just the verified clients might get to the information, alongside one seriously checking, called approval. In the recovery stage, prior to getting to the information, access control rules must be checked and either allowed to give the information access or deny the entrance. After the fruitful finishing of the lattice phase, then the user will be allowed to access the data.

### 4.4. Rules of the system users

**General rules of propose System**

Pi = < Subject (Si), Data Object (Oi), Operation (Ri)>

Ri = {read Y/N, write Y/N, Premium permission/ Null, Data Range or Data Area}

[Lower bound ➔ read Y/N: Upper bound ➔ write Y/N, Premium permission]

**Rules for Teacher**

$P_{ij}$ = < Subject (Si), Data Object (Oi), {r, 0, Related Data Object of academic years}>

$R_{ij}$ = {r, 0, (Data Object) Related Data Object of academic years}

**Rules for Senior Teachers**

$P_{ij}$ = < Subject (Si), Data Object (Oi), {r [Data Object of all academic years], w [Data Object of Academic year], Related Data Object of academic years}>

$R_{ij}$ = {r [Data Object of all academic years], w [Data Object of Academic year], Related Data Object of academic year}

**Rules for Head of Department**

Pij = < Subject (Si), Data Object (Oi), {r [Data Object of all academic years], w/update [Data Object of Academic year], Related Data Object of academic years}>

Rij = {r [Data Object of all academic years], w / Update [Data Object of Academic year], Related Data Object of academic year}

## 5. System Implementation

Through lattice, the proposed system ensures that the various levels of confidentiality are enforced at the point of data processing and prevents unauthorized disclosure. Lattice-based access control categorizes who has access to a document based on who owns it and what roles the EDC staff hold in order to determine who can view it. This accomplishes the security controls necessary for each classification as well.

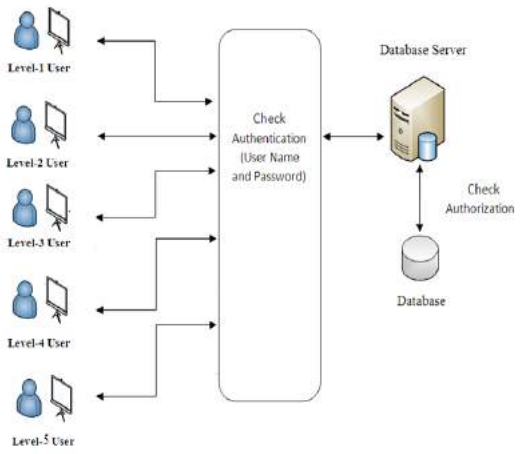



**Figure 2: System Overview**

In the proposed system, administrator can access all data and can make all transaction of the whole system and the data occupation of the respective level. The users of the proposed system are Admin User (level-1), Department Head (level-2), Senior Teacher (level-3), Teaching Staff (level-4) and Student Affair (level-5). These levels are defined by system administrator or board of the organization.

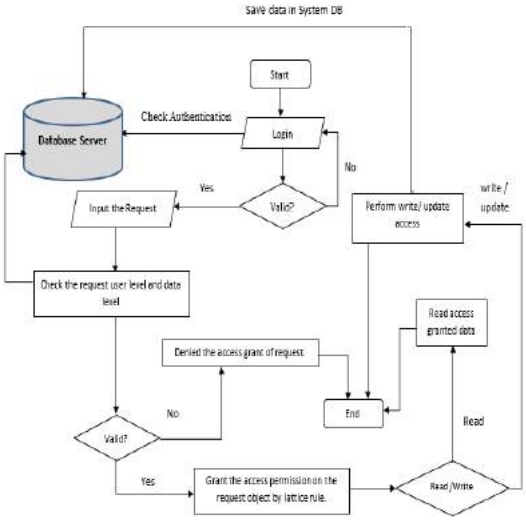



**Figure 2. The System Flow**

The assigned category of the admin on the object accesses are:

1) Essential data submission,
2) Data Management (Limited by System Rules), and
3) User management (Subject management).

**Authentication:** Authentication is the demonstration of laying out or affirming something (or somebody) as real, that will be that cases made by or about the subject are valid. The framework will allow the framework clients if their login name and secret word is right. And afterward the framework client can get to the information by their level characterized by chairman. This is called authorization.

**System Authorization:** Authorization is the capability of determining access privileges to assets, which is connected with data security and PC security overall and to get to control specifically. All the more officially, "to approve" is to characterize access strategy. For instance, HR staffs are regularly approved to get to

representative records, and this strategy is typically formalized as access control rules in a PC framework. During activity, the framework utilizes the entrance control rules to conclude whether access demands from (confirmed) buyers will be allowed or dismissed. Approval consent award or not of the proposed framework is controlled as displayed in figure 3 and figure 4.

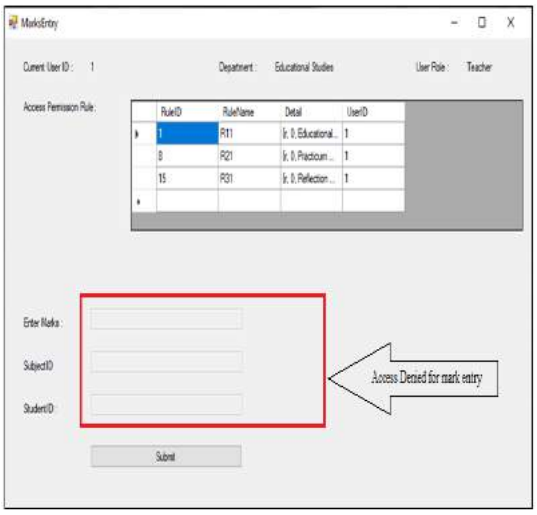



**Figure 3: Access Denied for Mark Entry**

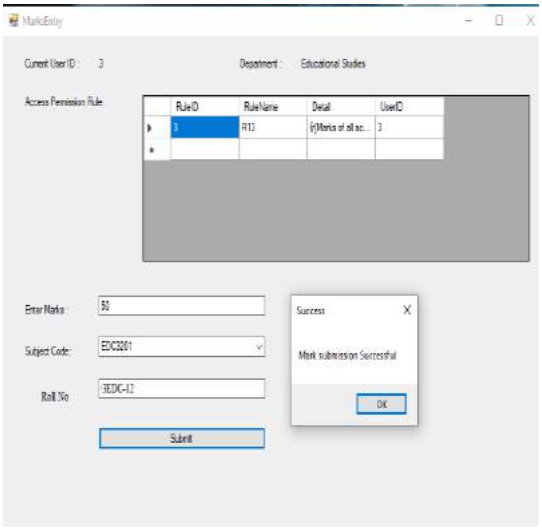



**Figure 4: Access Grant for Mark Entry**

## 6. Benefits of the System

The system can give privacy limitation. Besides, the chairman and proprietor of the framework can all the more really oversee and keep up with the significant data assets in a way steady with security strategies. At last, the delicate information in the data set can be saved safely. Conveyed admittance can be conceded.

## 7. Conclusion

Dynamic conditions present new test to get to control. In such circumstances, the substances change, the setups change and the functional modes change this demand ongoing update of access control arrangements. Cross section model gives security against unapproved divulgence and furthermore it offers assurance on modification of content, high accessibility through access control. These sorts of content assurance have been characterized by the information characterization structure. Additionally, the proposed model provides confidentiality. This access control of EDC system has multilevel to protect in relational database. These lattice rules will cooperate together to get secure strengthened. Authentication and authorization play a very important role in database security.

## References

[1] "A Lattice-Based Approach for Updating Access Control Policies in Real-Time", Tai Xin Indrakshi Ray Department of Computer Science Colorado State University, 2003.

[2] Database Server Security for Banking Information System, Zayar Aung, Htay Htay Thaung, University of Computer Studies, Yangon,2009.

[3] Katsumata, S. and Yamada, S. [2019]. Group signatures without NIZK: From lattices in the standard model, in Y. Ishai and V. Rijmen (eds), EUROCRYPT 2019, Part III, Vol. 11478 of LNCS, Springer, Heidelberg, pp. 312–344.

[4] Lattice Based Access Control for Protecting User Data in Cloud Environments with Hybrid Security, Saravanan, Dr. Umamakeswari, SASTRA University, India, 2021.

[5] Ling, S., Nguyen, K., Wang, H. and Xu, Y. [2017b]. Lattice-based group signatures: Achieving full dynamicity with ease, in D. Gollmann, A. Miyaji and H. Kikuchi (eds), ACNS 17, Vol. 10355 of LNCS, Springer, Heidelberg, pp. 293– 312

[6] Steven A. Demurjian Multi-Level Security in Healthcare Using a Lattice-Based Access Control Model, University of Connecticut, Storrs, USA, International Journal of Privacy and Health Information Management (IJPHIM),2019.

[7] Steve Demurjian Fall Jin Ma, "Implementation of Mandatory Access Control in Role-based Security System", Computer Science & Engineering the University of Connecticut, 2001.

# Vulnerability detection for HTTPS Spoofing and Email Hijacking attacks on web application using Boyer Moore String Matching Algorithm

Thazin Eaindra Bo, Zin Thu Thu Myint
*University of Computer Studies Yangon (UCSY)*
*thazineaindrabo@ucsy.edu.mm , zinthuthumyint@ucsy.edu.mm*

## Abstract

*Today, many businesses use the internet for promoting their work flow. Web application is one of the technologies using over internet to support this kind of e-businesses. A web application is software that can be accessed using any web browser such as Google, Chrome, Firefox and so on. Web applications have client side and server-side applications. When people access a web application from one browser, firstly a request sends to the web server and then server responds this request to web application and process required tasks. Web applications have become popular for people because they can be easily use, cost effective and other advantages. Maintaining web application security is the important fact for web application users because vulnerabilities can exist in web application. Web application vulnerabilities are weakness of E-business application and they can emerge from the variety of reasons, for example, application developer errors within code, application design weakness etc. Attackers can try by using these vulnerabilities to exploit system for getting privileges. In this paper, the proposed system can find the vulnerabilities that can accept two kinds of Man-In-The-Middle (MITM) attack which are HTTPS Spoofing and Email Hijacking attacks. MITM attack is well-known attack for cyber-security field. HTTPS Spoofing and Email Hijacking are kinds of MITM types of attack. These kinds of vulnerabilities are common risk for business. In this system, Boyer Moore string matching algorithm is used to search phishing URL by confirming with the pre-collected attacked URL datasets. This algorithm compare pattern and shift more than one position at a time so it saves time consuming. This proposed system is implemented by using Python programming language. Finally, the evaluation results show that how results can be accurate on false negative and false positive rate.*

**Keywords**: HTTPS Spoofing Attack, Email Hijacking Attack, Man-In-The-Middle Attack, Boyer Moore String Matching Algorithm, Web Application

## 1. Introduction

As developing technologies, people use many web applications for their respective fields. Sometimes, people need to load their credential data to web application for their specific use. Some attackers try to get unauthorized access and steal sensitive data and then can be used this data for malicious purposes such as identity theft, security violation and other illegitimate processes. Vulnerability is the drawback of web application that caused illegal accessing by attackers. So, this system can check vulnerabilities for information security before using web application. There are many types of vulnerabilities but Man-In-The-Middle attack is widespread and targeting connection between two people. MITM attack is one of the top 10 common types of Cyber-security. MITM attack has many kinds of attack such as HTTPS Spoofing, Email Hijacking, IP Spoofing, DNS Spoofing, Wi-Fi Eavesdropping and others.

Nowadays, most of the web application use HTTPS protocols for assisting the process of their works. HTTPS extension means Hypertext Transfer Protocol Secure. HTTPS protocol is communication protocol that encrypted http data over the internet for both client and server sides. In HTTPS spoofing, an attacker uses similar domain that look like targeted website of domain. Moreover, people widely use emails for their business field. In Email Hijacking, attackers try to access target's email account using target's

information. So, we propose a system for vulnerability detection for these two kinds of MITM attack using Boyer Moore String Matching algorithm. This algorithm is well-known algorithm for string matching process. There are many others string matching algorithm but some matching compares each character. So, they need many times for comparison with data. The proposed algorithm is reliable for string matching within system.

This paper is organized with the following parts. In section 2, this describes about the related works and section 3 explains background theory of system about vulnerability detection for HTTPS Spoofing and Email Hijacking of web application using Boyer Moore algorithm. Section 4 and 5 present the architecture of the system and evaluation result of the system's output.

## 2. Related Works

Firstly, the administrators propose to detect vulnerabilities for web application and to prevent from the attackers. This section explains the studies of vulnerability detection for web application.

In the first study, Argha Ghosh and A. Senthilrajan explained about an approach for detecting Man-In-The-Middle attack using DPI and DFI.[2] This paper uses a technique for detection this attack using Deep Packet Inspection and Deep Flow Inspection. DPI and DFI are technologies and DPI is used to manage and analyze for real-time network traffic with high-speed networks, DFI use to classify network flow feature for traffic of incoming network. Moreover, a co-ordinate module is used between DPI and DFI for maintaining data transmission and identifying incoming, outgoing traffic affected by MITM attack or not. In their experimental evaluation section, this system analyzes the performance of proposed approach using Wireshark.

In the next study, Robert A. Sowah, Kwadwo B. Ofori-Amanfo, Godfrey A. Mills, and Koudjo M.Koumad explained about detection and prevention of Man-in-the-Middle Spoofing attacks in MANETs using in Artificial Neural Networks (ANN).[6] A Mobile Ad-Hoc Network (MANET) means convenient wireless infrastructure and presents advantage for network settings. The ANN classification algorithm process attack detection, reconfiguration and isolation. It is measured on dataset with network-varied traffic and mobility patterns. This paper had 88.235% of final detection rate and process not only offered productive and cheap way to perform MITM attacks determining.

A. Zubaidah MohdSaleha, N. A. Rozalia, A. G. Bujaa, K. Abd. Jalila, F. H. M. Alia, T. F. Abdul Rahmana proposed a detection method for web application vulnerabilities using Boyer Moore String Matching Algorithm.[1] This paper showed that ability to detect vulnerabilities based on false negative and no false positive with minimum processing time. This paper detected four common vulnerabilities which are SQL Injection, Buffer Overflow, Cross Site Scripting and Cross Site Request Forgery. In this paper, the system only can detect web pages and these four vulnerabilities. Finally, this paper recommends that proposed method should be combined hybrid string matching for better detection.

U. Rahamathunnisa A.P, Vellore N. Manikandan, V. U. S. Kumaran, V. C. Niveditha analyzed preventing from phishing attack by implementing url pattern matching technique in web.[7] This paper showed that identification of phishing websites using uniform resource locators (URLs) matching webpages. This paper uses TF-IDF algorithm and measures appearance of particular term in whole document. This paper detected phishing attack by matching request URL with a blacklist and whitelist of database. This process runs browser backend and validates each and every request.

All of these papers are described in this section use variety of methods for vulnerability detection. Then, the system is continued to study method of vulnerability detection for Man-In-The-Middle attack.

## 3. Background Theory

In this section, the system discusses about web application vulnerability, HTTPS Spoofing attack, Email Hijacking attack and detailed information of Boyer Moore string matching algorithm.

### 3.1. Web Application Vulnerability

A web application's vulnerability is a weakness of the system in web-based application. These vulnerabilities can be found from testing such as invalid form inputs and web server, application design errors within the system. These vulnerabilities can harm the users of web application and make to exploit the security of web application. This paper implemented the system to find the vulnerabilities that can accept HTTPS Spoofing and Email Hijacking attacks.

### 3.2. HTTPS Spoofing Attack

HTTPS Spoofing attack is a kind of Man-In-Middle attack types. HTTPS supports the secure communications about user information. An attacker uses similar domain that look like the targeted website domain which he wants to connect website. Then, the attackers send the wrong URL to the intended users to get the privileges of the system access and performing the attack such as phishing. For example, an attacker can make fooling browser as trusted website actually this website is not. Users can believe and then accessing this website and steal personal information of users from user's sharing.

### 3.3. Email Hijacking Attack

Another form of Man-In-The-Middle attack types is Email Hijacking attack. In this attack, the attackers try to access to an email account of targeted person. And those attackers monitor communications processes of the targeted person and steal user's information for malicious purposes. For example, attackers can spoof email addresses of bank and send their instructions to targeted customers. Then, these customers believe to follow instructions of attacker rather than bank's instructions. As a result, these customers may lose money in attackers' hands.

### 3.4. Boyer Moore String Matching Algorithm

String or Pattern matching algorithm can check whether a specific sequence of characters, tokens and data are included in given data or not. Many Programmers use pattern matching techniques to determine the correctness of source files. This is used to find and replace a matching pattern or string in a text. Many applications that support search functionality uses pattern matching in various ways. There are many kinds of pattern or string-matching algorithms such as Boyer Moore, Naïve, Knuth Morris Pratt, Shift-Or, Karp-Rabin algorithms and others.

In this paper, system use Boyer Moore String Matching algorithm to search specious links of url and email in web application. Boyer Moore algorithm is well-known pattern matching algorithm and can be used for detecting and searching of specious features vulnerabilities in web application. In Boyer Moore algorithm, this has text and pattern for searching process. For searching, this compares text and pattern included character by character from right to left beginning with rightmost character. The preprocessing phase of this algorithm is O(m+n) time and space complexity. The complexity of the searching process is O(mn). When searching the worst case for a non-periodic pattern, it makes 3n text character comparisons. The best performance complexity is O(n / m) where m and n are lengths of pattern and text.

The following figure (Fig 1) is last-occurrence function part of Boyer Moore String Matching algorithm. This function pre-process pattern P and alphabet S to build last-occurrence function L by mapping S with index. It returns largest index i while P[i] = c, c means the last character of substring text and -1 for no index exists. Σ means collection of characters uniquely in substring text.

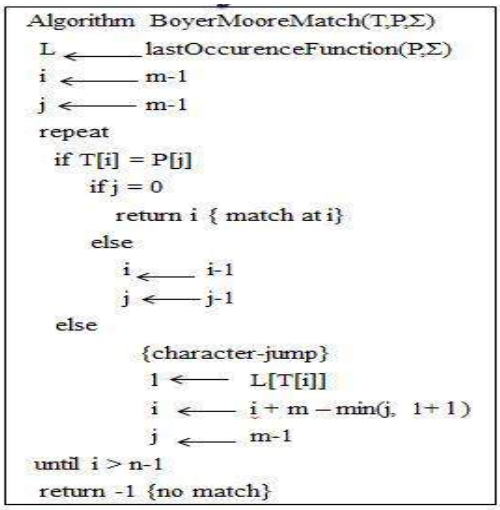



**Figure 1. Part of Boyer-Moore String Matching Algorithm**

The Boyer-Moore's string matching algorithm includes two principles, which are Right to Left

Scan and Bad character shift. The Bad character shift rule can shift more than one position at a time. In this algorithm, when a pattern is found in text, it returns the starting index of match pattern. Note that, length of text is always greater than length of pattern.

In this proposed system, the system uses Boyer Moore string matching algorithm to search specious features of URL in web application. Boyer Moore algorithm is better than other string-matching algorithm. Because it is time consuming for searching process. Its best performance complexity is O(n/m).

## 4. Overview of the Proposed System

In this section, the implemented vulnerability detection for HTTPS Spoofing and Email Hijacking attacks of web application using Boyer Moore String Matching algorithm will be explained. This system applies Boyer-Moore (BM) String Matching Algorithm to find attack signatures in URL and Email. The following figures are overview of the proposed system.

### 4.1. Proposed System for HTTPS Spoofing

In this paper, the system process four steps. The detailed system architecture shows in Figure 2.

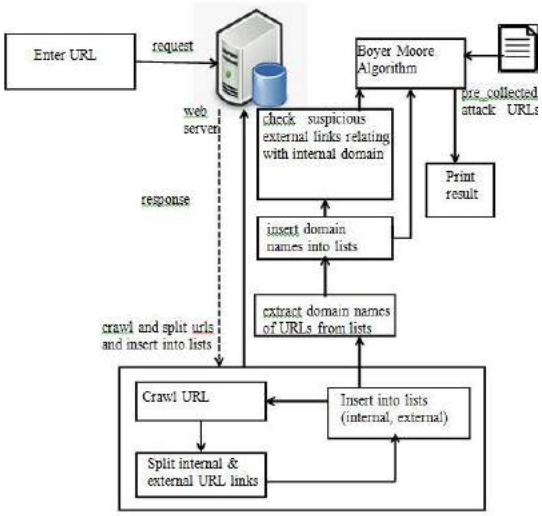



**Figure 2. Architecture of HTTPS Spoofing Attack**

*In Step 1*: Enter URL and request to the server. Firstly, to use this web application, users enter their urls to check this vulnerability.

*In Step 2*: When server responses URL, then system crawl the related URL of corresponding web page. After that, the system split these URLs into internal and external URL links and insert into lists. Then, the domain names are also extracted and insert into corresponding lists.

*In Step 3*: Getting domain names check with pre_collected attack urls to find vulnerabilities by using Boyer Moore algorithm.

*In Step 4*: And then, if not exit vulnerability, external URL links check again according to the relating with internal domain by using algorithm.

By using above the steps, this system can decide whether the vulnerability of HTTPS spoofing attack exists or not.

### 4.2. Proposed System for Email Hijacking Attack

For Email Hijacking attack, the system does the following four stages. In Figure 3 shows the detailed system architecture.

*In Step 1*: Enter owner and sender's gmail address. The system check sender's mail address with pre_collected blacklists email address.

*In Step 2*: Read the specified email in user laptop that are downloaded from Mail Server using post office protocol (POP). Then, login into owner's gmail inbox and select sender's mail. And email's detailed information and links are extracted from this email.

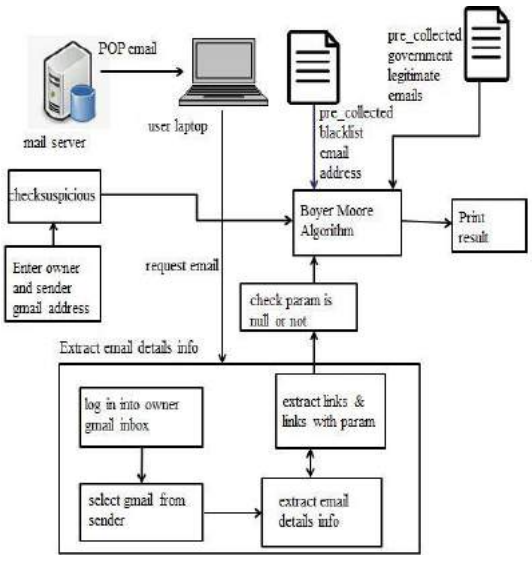



**Figure 3. Architecture of Email Hijacking Attack**

*In Step 3*: Then, extracted email links make checking with pre_collected attack links to find vulnerability by using Boyer Moore algorithm.

*In Step 4*: And then, extracted external links check again including parameter or not. If these links exit parameter, system checks sender's email address with pre_collected legitimate email lists. When sender's email includes in this list, this email cannot be assumed as Email Hijacking attack. If not, this email can be assumed as including attack.

## 5. Experiment and Result Analysis

The proposed method was evaluated by terms of false positive, false negative, true positive and true negative rates. The accuracy of this system is indicated by comparing with the attacked datasets. This proposed system was also measured on the correctness (accuracy) of vulnerability detection. The accuracy index can be calculated using the following confusion matrix in Table 1.

**Table 1. Specification of True Negative, True Positive, False Negative and False Positive Rate**

| | Normal | Attack |
|---|---|---|
| Normal | True Positive (TP) | False Positive (FN) |
| Attack | False Negative (FP) | True Negative (TN) |

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN}$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP+FN}$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN}$$

$$\text{True Negative Rate (TNR)} = \frac{TN}{TN+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

**Figure 4. Equations for accuracy measure**

In Figure 4, the proposed system shows that TN is the system can accurately detect the vulnerability when the website is actually vulnerable and TP means the system can actually detect that there is no vulnerability when the website is not vulnerable. FP is the system can detect that there has no vulnerability while the website has vulnerability and FN means the system can detect that there has vulnerability but actually, there is no vulnerability in the website.

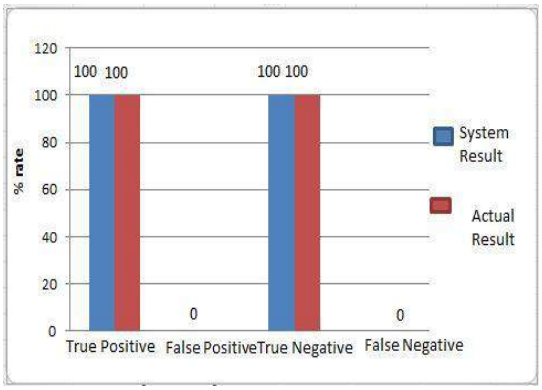



**Figure 5. Comparison results of proposed system and attacked dataset**

In this Figure 5 of evaluating result, firstly, the system collected 100 URLs and Emails and then these links were examined by their corresponding type of vulnerabilities using Boyer Moore String Matching Algorithm. In this checking, the system used attacked URLs datasets to confirm the web application has vulnerability or not. In this proposed system, this found 80 links of true positive, 20 links of true negative and zero links in both false positive and false negative. The proposed system's result was equal the actual dataset's result. The accuracy result of the proposed system was calculated by the following equation and show in Table 2.

Accuracy=
((TPR+TNR)/(TPR+FPR+TNR+FNR)) * 100

**Table 2. Accuracy of the proposed system**

| Measure | Value (%) |
|---|---|
| True Positive Rate | 100 |
| False Positive Rate | 0 |
| True Negative Rate | 100 |
| False Negative Rate | 0 |
| Accuracy | 100 |

In Table 2 express that accuracy of the proposed system is 100% and tested on 100 URLs and Emails links. The processing time of the proposed system is not very long and saves the time for vulnerability detection for URLs and Emails. This proposed system is intended to help

many people to save for using URLs and Emails according to do their business and other fields.

## 6. Conclusion

In this system, Boyer Moore Algorithm focus on the detection of web application vulnerability in two kinds of MITM attacks, which are HTTPS Spoofing and Email Hijacking attack. The system uses the input as Uniform Resource Locator (URL) and email for detecting web application vulnerability. The users can detect the type of vulnerability, which is either HTTPS Spoofing or Email Hijacking. Moreover, the Boyer Moore String Matching Algorithm is the effective algorithm for vulnerability detection because it compares character and then shifts the whole pattern. So, it saves time. HTTPS Spoofing and Email Hijacking attacks are dangerous to everyone in the society including organizations, person and hence these attacks need to be detected accurately. This provides a great help for ordinary man in protecting their important information by detecting vulnerabilities of these attacks.

In this study, the system does not consider other kinds of Man-In-The-Middle attack such as DNS Spoofing, IP Spoofing and others. In future work, the system will be aimed to detect other vulnerabilities in web application with low rate of false positive and false negative.

## References

[1] Ain Zubaidah Mohd Saleha, Nur Amizah Rozalia, Alya Geogiana Bujaa, Kamarularifin Abd. Jalila, Fakariah Hani Mohd Alia, Teh Fradilla Abdul Fahmana, "A Method for Web Application Vulnerabilities Detection by Using Boyer-Moore String Matching Algorithm", Information Systems International Conference (ISICO 2015), Universiti Teknologi MARA, Shah Alam 40000, Selangor, 2015.

[2] Argha Ghosh(&) and A. Senthilrajan, "An Approach for Detecting Man-In-The-Middle Attack Using DPI and DFI", A. P. Pandian et al. (Eds.): ICCBI 2019, LNDECT 49, Department of Computational Logistics, Alagappa University, Karaikudi, India, pp. 563–574, 2020

[3] Email blacklist domain name data are collected https://github.com/disposable-email-domains /disposable-email-domains/, visit on February 2022.

[4] Email legitimate data are collected from 'Government Directory-Myanmar National Portal'website https://myanmar.gov.mm/governmentdirectory/, visit on March 2022.

[5] OWASP Top Ten, [online] available: https://owasp.org/www-project-top-ten/, visit on February 2022.

[6] Robert A. Sowah , Kwadwo B. Ofori-Amanfo, Godfrey A. Mills, and Koudjo M. Koumad, "Detection and Prevention of Man-in-the-Middle Spoofing Attacks in MANETs Using Predictive Techniques in Artificial Neural Networks (ANN)", Journal of Computer Networks and Communications Volume 2019, Article ID 4683982, Department of Computer Engineering, University of Ghana,2019

[7] U. Rahamathunnisa A.P (Sr), SITE, Vellore N. Manikandan A.P (SG), SITE, Vellore U. Senthil Kumaran Associate Professor, Vellore C. Niveditha Student, MCA, "Preventing from phishing attack by implementing url pattern matching technique in web", Article in International Journal of Civil Engineering and Technology, VIT University, Vellore, September 2017

[8] URL attack data are stored from https://github.com/mitchellkrogza/Phishing.Database/, visit on February 2022.

[9] What is Spoofing and How to Prevent a Spoofing Attack, [online] available: https://www.pandasecurity.com/en/media/center/panda-security/what-ispoofing/, visit on March 2022.

[10] Xin Wanga,一, Runpu Wua, Jinxin Maa, Gang Longa, Jedeng Hana, "Research on Vulnerability Detection Technology for WEB MailSystem ",8th International Congress of Information and Communication Technology (ICICT-2018),China Information Technology Security Evaluation Center,China,2018

[11] Zhiping Jiang, Kun Zhao, Rui Li, Jizhong Zhao & Junzhao Du, "identity spoofing attack detection and prevention for a wireless edge network", Journal of Cloud Computing volume 9, Article number: 5 (2020).

# Detection of SQL Injection Attacks in Online Learning System Using Rabin-Karp Pattern Matching Algorithm

San San Wai, Yi Mon Thet
*University of Computer Studies, Yangon, Myanmar*
*sansanwai1@ucsy.edu.mm, yimonthet@ucsy.edu.mm*

## Abstract

*Nowadays, SQL injection is one of the most threatening web application attacks used against SQL database servers and web applications such as online learning, online banking, and online shopping, etc. Due to the pandemic of COVID-19, many online learning platforms are appearing because online learning is an important role in universities, colleges, institutions and schools for continuous learning from anywhere and anytime. Attackers can use SQL Injection to get unauthorized access and perform unauthorized data modification. To overcome this problem from attacking with SQLi in web applications, this paper proposes the detection of SQLIAs in online learning system by using pattern matching approach - Rabin-Karp Pattern Matching Algorithm. The proposed system uses SQL injection dataset with 1224 injection patterns from Kaggle and the experimental results show that the detection of SQLIA types and attackers' information (such as MAC address, IP address, etc.) and evaluate the performance in SQLI detection in terms of Accuracy (ACC).*

**Keywords**: SQL injection, information security, attack detection, Rabin-Karp Pattern matching Algorithm

## 1. Introduction

A variety of web applications are available for day-to-day activities such as online learning, online banking, online shopping, etc. Along with the prosperity of web applications, inevitably they are becoming the main targets of malicious attackers. According to the Web Application Attack Statistics: 2021 [11], a web application on average 500 - 700 attacks per day. As one of the most serious threats to web applications, SQL Injection Attacks (SQLIAs) are widely used by attackers to obtain unauthorized access to sensitive information. Therefore, this proposed system aims to detect SQLIAs for online learning web application by using Rabin-Karp pattern matching algorithm.

The remainder of this paper is structured as follows. Section 1 describes the introduction to web application attacks especially in SQLIA, and the related works with the proposed system are presented in Section 2. Section 3 explains the background theory that is applied in the system and the most common types of SQLIAs. and Section 4 presents the proposed system of the design. Section 5 represents the implementation, experimental results and evaluations and finally Section 6 ends with conclusion and future work.

## 2. Related Work

The different researches and approaches have been presented and published many techniques for detection and prevention of SQL Injection Attack (SQLIA). In web- based security problems, SQLIA has the top most priority.

The researchers in [1] proposed a novel technique for SQLi attacks detection and prevention using Bitap string matching algorithm. This system begins by precomputing a set of bitmasks containing one bit for each element of the pattern.

A novel technique for SQL injection prevention in cross-site scripting attacks using Knuth-Morris-Pratt string match algorithm was also proposed in [2]. In this work, the filter function was used to pass every input string. This function will block the user, reset the HTTP request, and display a corresponding warning message if at least one function returns True.

The approach in [3] based on comparing, at run time, the parse tree of the SQL statement before inclusion of user input with that resulting after inclusion of input. Code conversion to each and

every user input is more time consuming as well as the database size will also increase.

In the study [4], a hybrid technique for SQL Injection Attacks detection and prevention was proposed that detects and prevents all types of SQLIAs in different system categories regardless of the system development language or the database engine. However, it takes lot of time delay when the database recovery operation is performing after the SQLIA is detected.

The researchers in [5] proposed a scheme SQL Injection Attack detection and prevention using Aho–Corasick pattern matching algorithm. The proposed scheme was evaluated by using sample of well-known attack patterns.

## 3. Background Theory

SQL injection, also known as SQLI, is a common type of attack that use malicious SQL code for backend database manipulation to access information that was not intended to be displayed. SQL injection attacks can be used to target any application that uses a SQL database. SQL injection is a technique in which attacker injects an input query in order to the query and illegally gain the access of the database.

### 3.1. Web Application Security

Web applications allow various types of users to access the obtainable services. The permanent availability of web applications will increase the opportunity for everyone who is looking to exploit and damage these applications for illegal purposes.

### 3.2. Vulnerabilities in Web Application

There are several types of web application vulnerability; each one has special properties, such as the vulnerability style, the detection and prevention techniques. Figure 1 shows the statistics of OWASP (open web application project) top ten vulnerabilities that is used in the hacking of web application in 2022 [13]. The statistics have been conducted according to the number of exploiting the same vulnerability. OWASP top ten 2022 SQL Injection vulnerability is as follow:

***Injection***: This type occurs when the attacker injects the application command or queries by untrusted data. SQL Injection attacks are one of the major attacks targeting web applications as reported by Open Web Application Security Project (OWASP) [11]. This can leave the system vulnerable and can result in severe loss of data [6]. A SQLI attack is one of the deadliest attacks because it compromises authentication, integrity, authorization and confidentiality [7].

The SQLi compromises the main security services: confidentiality, authentication, authorization and integrity [8]. It can be considered as the most dangerous attacks of the injection category. SQLI attack consists in injecting (inserting) malicious SQL commands into input forms or queries to get access to a database or manipulate its data (e.g. send the database contents to the attacker, modify or delete the database content, etc.) [9], [10].

### 3.3. Types of SQL Injection Attacks (SQLIAs)

The most common types of SQL injection attacks are Boolean-based, Union-based, Like-based, Batch Query, Comment-based, and Time-based SQL injection attacks.

**Table 1. Most Common Types of SQLIAs**

| No. | Attack Types | Sample Injection Code |
|---|---|---|
| 1. | Boolean-Based SQLi | anything' or 'x' = 'x |
| | | 123 or 1 = 1; -- |
| 2. | Like-Based SQLi | username LIKE a% |
| | | or uname like %s |
| 3. | Union-Based SQLi | 'UNION select * from users |
| | | union select * from uid; |
| 4. | Batch Query | drop table users; |
| | | drop table temp;-- |
| 5. | Comment-based SQLi | -- |
| | | "_" |
| 6. | Time-based SQLi | 'sleep 50' |
| | | 1 waitfor delay '0:0:10'-- |

#### 3.3.1. Rabin-Karp Pattern Matching Algorithm

Unlike Naïve string-matching algorithm, Rabin–Karp algorithm or Karp–Rabin algorithm does not travel through every character in initial phase. It filters the characters that do not match the hash values and performs comparison. In this

proposed system, Rabin-Karp pattern matching algorithm is used to detect SQL injection attacks in online learning system. The step-by-step procedure how the Rabin-Karp algorithm works is described as follows: (1) Takes a sequence of characters, (2) Checks for possibility of the presence of the required string and (3) If the possibility is found then, character matching is performed.

### 3.3.2. Procedure of Rabin-Karp

The procedure Rabin-Karp and nested procedure calls are described as follows. Firstly, the procedure will calculate the hash value of pattern and text string by calling procedure Calculate-Hash(). Then, the two hash values are compared. If the two hash values, the characters in pattern and text string will be performed the string matching. When the characters are exactly equal, the pattern can be regarded as found. If not, the pattern can be regarded as not found.

```
Procedure Rabin-Karp (Text, Pattern, Prime):
m := Pattern.Length
HashValue := Calculate-Hash(Pattern, Prime, m)        //Compute Hash Value of Pattern
CurrValue := Calculate-Hash(Text, Prime, m)           //Compute Hash Value of Text with size m
for i from 1 to Text.length – m
    if HashValue == CurrValue and String-Match(Text, Pattern, i) is true        //Found
 Return i
   end if
 CurrValue := Recalculate-Hash(String, i+1, Prime, CurrValue)          //For Next Text Window
   end for
Return -1          //Not Found
```

**Figure 2. Procedure of Rabin-Karp Algorithm**

```
Procedure Calculate-Hash(String, Prime, x):
hash := 0
for m from 1 to x                          // Here x denotes the length to be considered
    hash := hash + (Value of String[m])    // to find the hash value
  end for
Return hash
```

**Figure 3. Procedure of Hash Value Calculation**

```
Procedure String-Match(Text, Pattern, m):
for i from m to Pattern. Length + m - 1
   if Text[i] is not equal to Pattern[i]
      Return false
   end if
end for
      Return true
```

**Figure 4. Procedure of String Matching**

```
Procedure Recalculate-Hash(String, Curr, Prime, Hash):
   Hash := Hash - Value of String[Curr-1]   //here Curr denotes First Letter of Previous String
   Hash := Hash % Prime
   m := Pattern.length
  New := Curr + m - 1
   Hash := Hash + (Value of String[New])
Return Hash
```

**Figure 5. Procedure of Hash Value Recalculation**

## 4. Design of the Proposed System

There are two main phases in the proposed system: data collection phase and detection phase. When the attacker enters the SQL injection patterns into the system, the proposed system will compare these patterns with the injection patterns from the database by using Rabin-Karp pattern matching algorithm. In this section, the overview system design of the proposed system and the system flow diagram are described in Figure 6 and Figure 7.

The overview system design of the proposed system is described as follows. SQL Injection Attacks (SQLIAs) occurs when an attacker can user input query and the attacker can inject queries on the URL. Compare user input query (substring) with of retrieved pattern from existing SQL database by using Rabin-Karp pattern matching algorithm. Check whether the pattern matches or not? Pattern is matched and SQL injection attack is detected. Pattern is not matched with any SQL injection pattern; and the system will allow the user to login as an authenticated user.

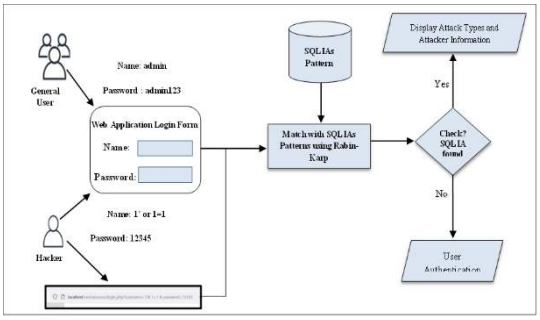



**Figure 6. Overview System Design of the Proposed System**

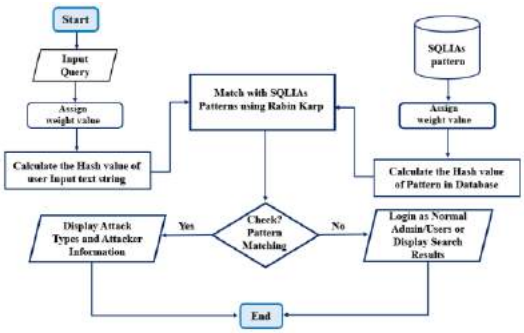



**Figure 7. System Flow Diagram**

The step-by-step procedure of the proposed system is described as follows. As shown in the pattern matching algorithm, the attacker will input the injection query to login into the system or modify the credentials of users in database. The system will detect by comparing the attacker input query (substring) with all of retrieved patterns with Rabin-Karp pattern matching algorithm. The system will check whether the pattern matches or not. If the patterns are equal, SQL injection attack is detected and SQL injection attack type and attacker's information are displayed by the system. If the input pattern is not equal with any SQL injection pattern, the system will allow the user to login as an authenticated user or display the query's results.

**Algorithm** - Pattern Matching
**Input** - User Input SQL query.
**Output** - Pattern found or not found.
**Step 1** - User input query is made as substring according to pattern size.
- Assign weight value.
- Calculate the hash value.

**Step 2** - Compare user input query (substring) with all of retrieved patterns from SQL injection pattern database until end of string query by using Rabin Karp pattern matching algorithm.
**Step 3** - Check whether the pattern matches or not?
- If it is equal to one of retrieved patterns, SQL injection attack is detected and SQLI attack type is displayed by the system.
- If it is not equal with any pattern, then the system will allow login as authenticated user or display query's results.

**Figure 8. Pattern Matching Algorithm**

## 4.2. Data Collection

In this proposed system, the SQL injection patterns are needed to collect to build SQL injection patterns database. Therefore, the SQL injection dataset in csv file is downloaded from Kaggle, the world's largest data science community with powerful tools and resources to help us achieve data science goals. The total number of SQL Injection Queries is 1128 samples in this dataset will be used to build SQL injection patterns database.

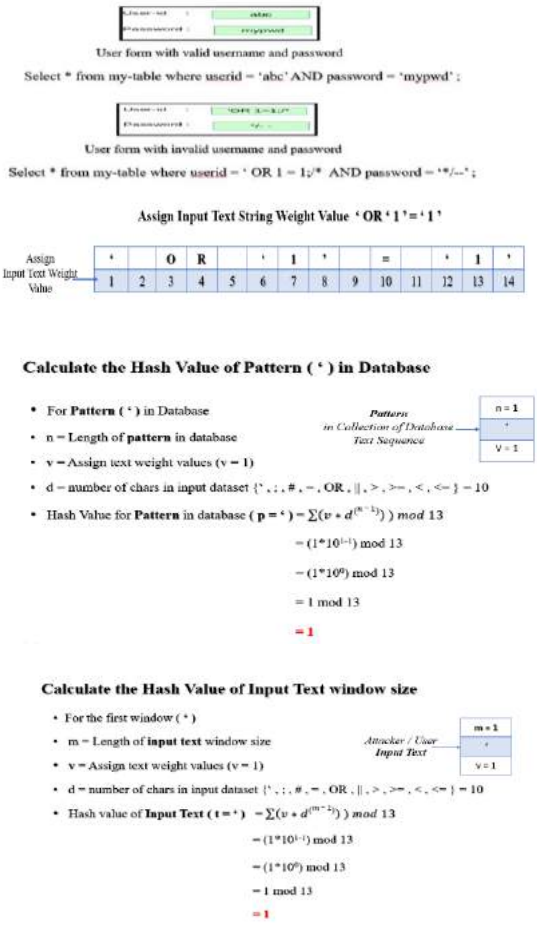
## 4.3. SQL Injection Detection

The proposed system will detect SQL injection attacks into different attack types: Boolean-based, Union-based, Like-based, Batch Query, Comment-based and Time-based. This system uses Rabin-Karp pattern matching algorithm to SQL injection attack. The procedures and calculation steps of the proposed algorithm will be described in the following subsections.

### 4.3.1. Using Boolean-Based SQLIAs

Boolean-Based SQL injection strings have a single quote (') followed by logical operator OR and a true statement such as '1' = '1';#, 'a' <> 'b' ;#, '2 + 3' < = '10' ;#.

**Example**



Because the hash value of the pattern in database is equal to the hash value of the input text, character-matching is performed. Both the characters match, SQL Injection Attack found.

## 5. Implementation of the Proposed System

The purpose of this section is to present the implementation of the proposed system. The proposed system uses SQL injection dataset from Kaggle to detect the SQL injection attack. (https://www.kaggle.com/datasets/syedsaqlainhussain/sql-injection-dataset). The downloaded dataset file is CSV (Comma Separated Value) file format which can be opened with as Excel file. Kaggle is the world's largest data science community with powerful tools and resources to help us achieve data science goals. The total number of SQL injection patterns is 1224 inject patterns that are included in this dataset. Figure 9 shows the SQL injection detection result by the proposed system.


**Figure 9. SQL Injection Detection**

### 5.1. Experimental Results

In this section, the proposed system evaluates the performance in SQL injection detection in terms of Accuracy (ACC) with the following formula. In this formula, TP is the True Positive rate, FP is False Positive rate, FN is False Negative rate and TN is True Negative rate.

| | | Actual Class | |
|---|---|---|---|
| | | *Attacked* | *Normal* |
| Predicted Class | *Attacked* | TP (True Positives) | FP (False Positives) |
| | *Normal* | FN (False Negatives) | TN (True Negatives) |

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}$$

The following figures show the performance evaluations and experimental results according to different types of injection patterns, the total number of SQL injection queries. As the figures show, the results showed that the more injection patterns with different types, the more accuracy percentages are found.

Total numbers of SQL injection patterns: 100 inject patterns (Boolean-based SQLi) in this dataset.

Total Numbers of SQLi queries = 50
Normal attempt to Login or Search = 23
No of correctly predicted malicious requests, TP = 22
No of incorrectly predicted malicious requests, FP = 5
No of incorrectly predicted normal requests, FN = 3
No of correctly predicted normal requests, TN = 20

$$ACC = \frac{22 + 20}{22 + 5 + 3 + 20} = \frac{42}{50} = 0.84 = 84\,\%$$

**Figure 10. Performance Evaluation**

Total numbers of SQL injection patterns: 100 inject patterns (Boolean-based SQLi, Like-based SQLi, Union-based SQLi, Comment-based SQLi) in this dataset.
Total Numbers of SQLi queries = 75
SQL Injection query = 58
Normal attempt to Login or Search = 17
No of correctly predicted malicious requests, TP = 50
No of incorrectly predicted malicious requests, FP = 8
No of incorrectly predicted normal requests, FN = 0
No of correctly predicted normal requests, TN **= 17**

$$ACC = \frac{50 + 17}{50 + 8 + 0 + 17} = \frac{67}{75} = 0.89 = 89\,\%$$

**Figure 11. Performance Evaluation**

Total numbers of SQL injection patterns: 1224 inject patterns (Boolean-based SQLi, Like-based SQLi, Union-based SQLi, Comment-based SQLi, Batch Query, and Time-based SQLi) in this dataset.
Total Numbers of SQLi queries = 100
SQL Injection query = 63
Normal attempt to Login or Search = 37
No of correctly predicted malicious requests, TP = 56
No of incorrectly predicted malicious requests, FP = 7
No of incorrectly predicted normal requests, FN = 3
No of correctly predicted normal requests, TN = 34

$$ACC = \frac{56 + 34}{56 + 7 + 3 + 34} = \frac{90}{100} = 0.9 = 90\%$$

**Figure 12. Performance Evaluation**

In this section, the efficiency and performance of enhanced Rabin-Karp Pattern Matching algorithm is calculated using the measures

accuracy. In this database, a malicious query which consists of attacks like Boolean-based SQL, Like-based SQL, Union-based SQL, Comment-based SQL, Batch Query, and Time-based SQL are injected into the database. The efficiency of the proposed technique is to identify and detect the database from SQLIA is presented in the above chart and it shows the accuracy (ACC). Calculating 100 (Boolean-based).

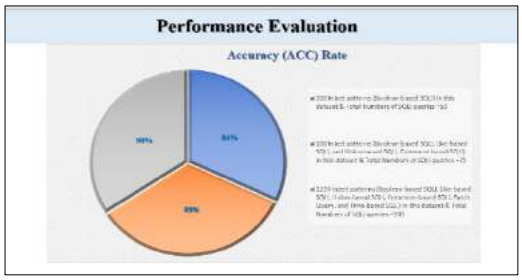



**Figure 13. Performance Evaluation**

SQLi patterns in the dataset and the total numbers of 50 SQLi queries, an accuracy value is 84 %. Calculating 100 (Boolean-based, Like-based, Union-based, and Comment-based) SQL injection patterns in the dataset and the total numbers of 75 SQLi queries, an accuracy value is 89 %. Calculating 1224 (Boolean-based, Like-based, Union-based, and Comment-based, Batch Query and Time-based) SQL injection patterns in the dataset and the total numbers of 100 SQLi queries, an accuracy value is 90%.

## 6. Conclusion

SQLIA is a common type of attack over the web application. SQLIA occurs when an attacker is able to insert a series of malicious SQL statements into a query through manipulating user input data and URL for execution the back-end database. In this proposed system, Rabin-Karp Pattern Matching Algorithm is used to detect SQL injection attacks on the Online Learning System and defining SQLI attack types (such as Boolean-based, Like-based and Union-based, etc.). The performance of the proposed system will be measured by correctly classifying user input queries entering the online learning system as normal queries or malicious queries.

### 6.1. Limitation and Further Extension

In this paper, the proposed system will not detect unknown SQL injection attacks that are not included in the dataset. This system detects only the different types of attacks: Boolean-based, Union-based, Like-based, Batch Query, Comment-based and Time-based SQL injection attacks. As further extension, the system needs to develop to detect other types of SQL injection attacks and build more sufficient SQLi patterns dataset.

## References

[1] N. Karthikeyan, R. Vivekanandan, M. Sakthivel, N. Dinesh, “A Novel Technique to Detect and Prevent SQL Injection Attacks using Bitap String Matching Algorithm”, Volume 27, Issue 4, 2021

[2] Abikoye, “A Novel Technique to Prevent SQL Injection and Cross-Site Scription Attacks using Knuth-Morris-Pratt String Match Algorithm”, EURASIP Journal on Information Security (2020)

[3] A. John, A. Agarwal, M. Bhardwaj, “An Adaptive Algorithm to Prevent SQL Injection”, American Journal of Networks and Communications, 2015

[4] J. O. Atoum and A. J. Qaralleh, “A Hybrid Technique for SQL Injection Attacks Detection and Prevention”, ( IJDMS ) 2014

[5] S. Kharche, J. Patil, K. Gohad, B. Ambetkar, “Implementation of Pattern Matching Algorithm to Prevent SQL Injection Attack”, IJSRST, 2018

[6] M. Hirani, A. Falor, H. Vedant, “A Deep Learning Approach for Detection of SQL Injection Attacks using Convolutional Neural Networks”, 2020

[7] I. Jemal, O. Cheikhrouhou, H. Hamam, A. Mahfoudhi, “SQL Injection Attack Detection and Prevention Techniques Using Machine Learning”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 15, Number 6 (2020) pp. 569-580

[8] G. Deepa, P. S. Thilagam, F. A. Khan, A. Praseed, A. R. Pais, and N. Palsetia, “Black-box detection of xquery injection and parameter tampering vulnerabilities in web applications,” International Journal of Information Security, 2018.

[9] Y. Fang, J. Peng, L. Liu, and C. Huang, “Wovsqli:Detection of sql injection behaviors using word vector and lstm,” in Proceedings of the 2nd International Conference on

Cryptography, Security and Privacy. ACM, 2018, pp. 170–174.

[10] Q. Li, F. Wang, J. Wang, and W. Li, "Lstm-based sql injection detection method for intelligent transportation system," IEEE Transactions on Vehicular Technology, 2019.

[11]OWASP, "Owasp top ten project," https://www.owasp.org/index.php/Category: OWASP Top Ten Project, 2019, accessed on April 2019

[12] A. Tajpour, S. Ibrahim, M. Masrom, "SQL Injection Detection and Prevention Techniques", International Journal of Advancements in Computing Technology · August 2011

[13]https://www.sonarqube.org/features/security/owasp

# SQL Injection Pattern Recognition Based on Naïve Bayes Model

Hsu Wai Tun, Khaing Khaing Wai
*University of Computer Studies, Yangon*
*hsuwaitun3377@gmail.com, khaingkhaingwai@ucsy.edu.mm*

## Abstract

*In recent years, sharing information through the Internet across various platforms and web applications has grown increasingly widespread. Users' critical information is stored in databases by the web-based applications that receive it. Due to their availability over the Internet, these apps and the databases that are connected to them are vulnerable to numerous cybersecurity incidents. Therefore, cyber-security is critical for securing users' critical data and information in this technology era. The attacker can steal critical and confidential information by using various threats. The threats include attacks such as Cross Side Scripting (CSS), Denial of Service Attacks (DoS), and Structured Query Language (SQL) Injection attacks. One of the 10 most popular risks and weaknesses to web applications with backend databases is SQL injection. It utilizes malicious SQL queries to modify internal data and retrieve information from the back-end database that was not intended to be displayed. Since there are countless cyberattacks every day and have really been needing on developing a more secure system that can predict them and prevent them from happening. In this thesis, the proposed system can also detect SQL Injection Attacks successfully by applying a machine learning algorithm based on Naïve Bayes Method. The proposed model was trained and evaluated with 21,523 instances of a dataset which comprises SQL Injection and no Injection. The user interface is created for a test case that anticipates either a malicious or benign question from the user. Finally, this system displayed the result of detecting the query that is SQL Injection or not, and evaluated how accurate the results are based on having false negative and false positive rates.*

**Keywords**: SQL Injection Attack, Naïve Bayes, Feature Extraction, Regular Expression, Gaussian Naïve Bayes, Confusion matrix

## 1. Introduction

The majority of the programs we use on a daily basis are web-based programs. Web applications accept users' crucial information and store this information in a database [7]. Attackers try to get easy access to the underlying database of web applications, making them susceptible. They use to attack various threats such as Cross Side Scripting (CSS), Denial of Service Attacks (DoS), and Structured Query Language (SQL) Injection attacks [4]. Among these, one of the most common cyber-attacks is SQL Injection Attacks. SQL injection is a popular attack method that allows access to data that was not intended to be revealed by altering the backend database utilizing malicious SQL code. Attackers may be able to access the databases that underlie Web applications using SQL Injection Attacks because it gives them the opportunity to leak, edit, or even destroy information that is retained in these databases

One of the most common and harmful types of hacker assault is SQL Injection Assault. Prevention of SQL injector attacks is a crucial and complex topic to learn in information system security. A SQL injection attack (SQLIA) in Web applications supported by a database is the main security risk. Through this vulnerability, attackers can quickly access the application's underlying database and any potentially sensitive data that may be present there. By carefully crafting input, a hacker can access database content that would otherwise be impossible. Typically, this is accomplished by altering the SQL statements employed by online programs. Due to the safety of web applications, researchers have extensively researched SQLIA detection and prevention and developed a variety of strategies.

However various techniques have been developed to counteract such attacks, and fraudsters continue to find ways to circumvent the different protections put in place to prevent SQL Injection attacks. Currently, there has been a

significant amount of debate concerning the use of machine learning algorithms to detect and prevent certain cyber security risks. The effectiveness of using supervised and unsupervised learning techniques to identify security threats cannot be challenged, but the computing power and processing time needed to run such complicated algorithms continue to be a major concern for the community working on cyber security, which is constantly evolving [6]. For the purpose of recognizing SQL Injection attacks, a significant amount of study has been done utilizing various machine learning methods. This system uses the Naïve Bayes algorithm to detect SQL Injection Attacks.

## 2. Related Work

Sangeeta, S Nagasundari, and PrasadB Honnavali [1] introduced SQL Injection Attack Detection using ResNet. In this system, the ResNet model is constructed using data collected from a variety of devices and the internet. The ResNet method is used to train on processed samples. The user interface is created for an experiment that predicts either a malicious or innocent question from the user. If a malicious or legitimate input request is made, the trained ResNet model can detect with accuracy which one it is. This system shows how ResNet can successfully identify several types of SQLIA. Muhammad Amirulluqman Azman, Mohd Fadzli Marhusin, and Rossilawati Sulaiman introduced [2] Machine Learning-Based Technique to Detect SQL Injection Attacks. In this system, machine learning is employed, and the SQL Injection detector is trained using training data before being checked against testing data. The access log is extracted and divided into a test set and a training set. The detector learns from the training set and develops a Knowledge Base (KB). Finally, the detector checks the test set depending on KB. The outcome of the detection demonstrates that the proposed technique achieves excellent accuracy in differentiating between malicious and legitimate web requests. M.KarthiKeyan [3] introduced An Efficient Technique for Preventing SQL Injection Attack Using Pattern Matching Algorithm. In this system, they suggested employing the Aho-Corasick pattern matching algorithm as a detection and prevention method to prevent SQL Injection Attacks (SQLIA). This system is evaluated by using a sample of well-known attack patterns. The proposed scheme has the following two modules, 1) Static Phase and 2) Dynamic Phase. In the Static Pattern List, they keep a list of known Anomaly Patterns. In the Dynamic Phase, an alarm will occur, and a new anomaly pattern will be established whenever a new anomaly of any kind unexpectedly emerges. The Static Pattern List will be updated with the new anomalous pattern. The initial evaluation suggests that the proposed approach is effective against the SQL Injection Attack by taking into consideration a sample of common attack patterns.

## 3. SQL Injection Attack

SQL injection sometimes referred to as SQLI, is a popular attack method that utilizes malicious SQL code to manipulate the backend database and access data that was not intended to be exposed [4].

Given the right conditions, a hacker may use a SQL Injection vulnerability in order to bypass an internet application's security and authorization measures and retrieve the entire database's contents.

SQL Injection is a server-side code injection technique that uses a predefined SQL statement to take advantage of a web application's vulnerability to attack the system, inserting the query into the URL or input fields. This query is sent by the web application to the database, which then processes it and sends data back to the web application. By using SQL Injection, the attackers successfully get access to the database in this manner. Therefore, attackers get access to sensitive data, the opportunity to modify database information, to run database administrator commands, and the ability to recover system files.

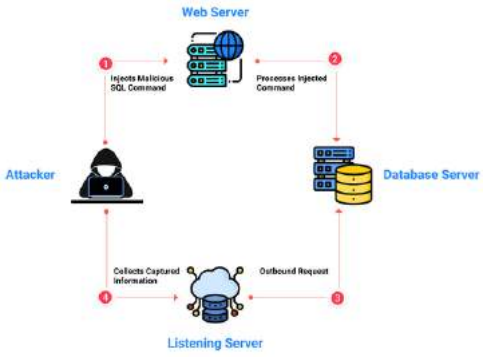



**Figure 1. SQL Injection Attack**

The following are the impact of SQL Injection when it enters the application.

- **Steal credentials**— attackers can impersonate users and use relevant privileges after obtaining credentials using SQLI.
- **Access databases**— attackers can access the private information on database servers.
- **Alter data**— attackers have access to the accessed database and can edit or add new data.
- **Delete data**— attackers are able to delete entire tables or erase database records.
- **Lateral movement**— attackers who have operating system privileges can log into database servers and utilize these rights to break into other sensitive systems.

### 3.1. Sample of SQLI Attacks

**1. Using SQLI to Authenticate as an Administrator**

This sample shows how a hacker can bypass authentication in an application and achieve administrative rights by using SQL injection. Assume a basic username-and-password database table-based authentication solution. The variables user and pass are obtained via a user's POST request, which is added to the following SQL statement:

```
SQL = "SELECT id FROM users
WHERE username='" + user + "' AND
password='" + pass + "'"
```

The attacker uses this SQL statement by concatenating a string like this instead of the passed variable:

```
password' OR 5=5
```

**As a result:**

```
SELECT id FROM users WHERE
username='user' AND password='pass'
OR 5=5'
```

In this statement, 5=5 is always true. So, the first ID from the users' table, which is typically the administrator, will be returned by the WHERE clause. This indicates that the attacker has administrator rights and can access the application without logging in.

**2. Using SQLI to Access Sensitive Data**

In this sample, the code gets the current username before checking for items with a certain item name whose owner is the current user.

```
string userName = ctx.getAuthenticatedUserName();
string query = "SELECT * FROM items WHERE owner = '"
        + userName + "' AND itemname = '"
        + ItemName.Text + "'";
```

The code generates the following query when the username and item name are combined:

```
SELECT * FROM items WHERE owner
=AND  itemname = ;
```

If the attacker inserts the following string for the item name:

```
Widget' OR 5=5
```

SQL statement becomes:

```
SELECT * FROM items WHERE owner
= 'John'
AND itemname = 'Widget' OR 5=5';
```

This is similar to

```
SELECT * FROM items;
```

As a result, the query will return all of the table's data, giving the attacker unauthorized access to confidential information.

**3. Injecting Malicious Statements into Form Field**

This is a simple user-input-based SQL injection attack. The attacker utilizes a form that asks for the user's first and last names.

The SQL statement that receives the form inputs has the following format:

```
SELECT id, firstname, lastname FROM
authors
```

When the attacker inserts a fraudulent expression to the first name, the SQL statement becomes:

```
SELECT id, firstname, lastname FROM
authors WHERE firstname = 'malicious'ex' and
lastname ='newman'
```

Due to the single punctuation, the database recognizes invalid syntax and attempts to execute the malicious statement.

## 4. Methodology of the Proposed System

Machine Learning is powerful for classification and regression. It has typically 4 techniques such as supervised learning, unsupervised learning, reinforcement, and semi-supervised learning. Supervised learning uses labeled datasets, in order to train algorithms that effectively identify data or predict outcomes. Among various supervised learning methods, the proposed system detects SQL Injection or not by using the Naïve Bayes method because it is simplest to understand when the technique is explained using binary or categorical input values.

### 4.1. Feature Extraction

Feature extraction is the process of breaking down the input data into a set of features so that the task at hand can be carried out using this condensed representation rather than the original, full-size input [5]. It is a method for reducing important features from big input data collection. Any extraction technique is used to extract each feature from datasets. A machine learning model can benefit from feature extraction when being trained. Any extraction technique is used to extract every phrase from datasets. In this paper, Regular Expression is used.

**Regular Expression**

Regular Expression is a set of symbols that defines a pattern of text that must match in order to make a filter more specific or general. RegEx sometimes referred to as regular expression, is a generalized expression that is used to match patterns with different character sequences [8]. A string formatted sequence of characters is defined by regular expressions. It uses patterns to match strings. The following are samples of Regular Expression.

**Table 1. Sample of Regular Expression**

| Pattern | Description |
|---|---|
| -- | Single line comment |
| /* */ | Multiline comments |
| ' | Single quotation |
| " | Double quotation |
| [](),; | Punctuations |
| && \|\| != == | Logical operator |
| + - * / | Arithmetic operators |

### 4.2. Naïve Bayes

The Naive Bayes classification algorithm is appropriate for both binary and multiclass classification. Compared to numerical input variables, naive Bayes performs better in cases of categorical input variables. It is helpful for analyzing data and making predictions based on past situations. The Naive Bayes algorithm is a probabilistic classifier. It is based on probability models that make extensive assumptions about independence. The term "Naive Bayes classifiers" refers to a set of classification methods built on the Bayes theorem.

**Bayes Theorem**

The Bayes Theorem determines the likelihood of an event occurring given the likelihood of an earlier event occurring. The mathematical formulation of Bayes' theorem is given by the equation:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

where, X = data sample
H = hypothesis
P(H|X) = posterior probability
P(H) = prior probability
P(X) = probability that sample data is observed
P(X|H) = likelihood

### 4.3. Gaussian Naïve Bayes

When there are several random variables, the states of the normal distribution which are averages of the random variables converge to the normal distribution and are normally distributed. Gaussian Naïve Bayes assumes that every class has a Gaussian distribution. All the continuous variables associated with each feature are distributed according to Gaussian Distribution,

$$\mathbf{f(x)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\overline{x})^2}{2\sigma^2}}$$

**Algorithm of the Proposed System**

Input:

T=Training dataset,

F= $(f_1, f_2, f_3, \ldots\ldots, f_n)$ //value of the predictor variable in testing dataset,

Output:

A class of testing dataset.

Step:

1. Read the training dataset T;
2. Calculate the mean and standard deviation of the predictor variables in each class;
3. Repeat
   Calculate the probability of $f_i$ using the gaussian distribution equation in each class;
   Until the probability of all predictor variables $(f_1, f_2, f_3, \ldots\ldots, f_n)$ has been calculated;
4. Calculate the posterior probability for each class;
5. Get the greatest likelihood;

The given dynamic query Test is categorized as malicious or non-malicious based on this result.

## 5. Description of the Dataset

The experimental dataset for the SQL Injection Attack was obtained from the Kaggle website. The sample dataset has two columns of data and 34,048 different values. The first column represents a query that has to be identified as either a normal statement or a SQL Injection Attack, and the second column provides a numeric value that helps identify the type of statement it is. In this case, the sentence has been represented by the value 1 as a SQL Injection query and by the value 0 as a regular statement.

**Table 2. SQL Injection Dataset**

| Dataset | Training | Testing |
|---|---|---|
| 100% | 70% | 30% |
| SQL query | 21523 | 9225 |

## 6. Implementation of the System

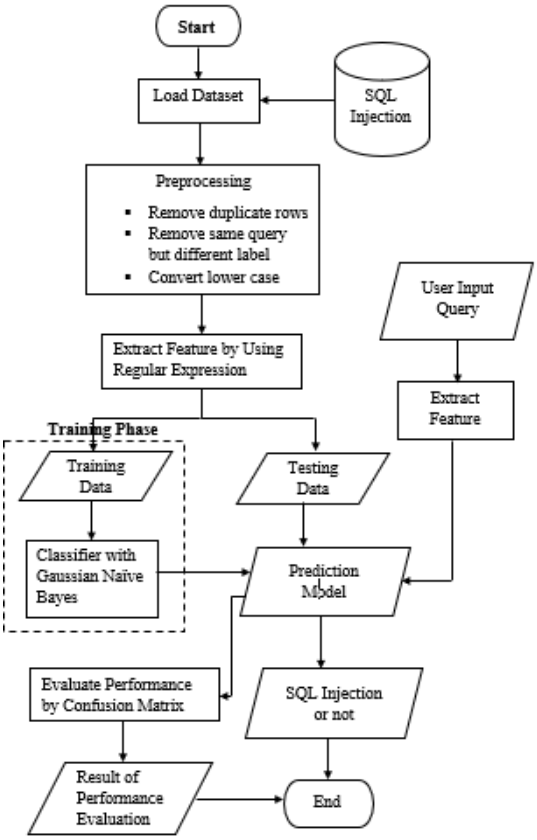



**Figure 2. Flow Chart of the Proposed System**

The flow diagram of the proposed system is shown in figure 2. Firstly, this system loads the dataset and then checks duplicate rows in the dataset. If it has, remove duplicate rows. Additionally, this system checks the same query but a different label in the dataset. If it has, drop rows that happen in this case. Finally, the system converts the lowercase of all the datasets. And then this system extracts the feature by using a regular expression from the dataset and saves it in the database. After that, this dataset splits into 70% training dataset and 30% testing dataset. Furthermore, this system builds a model with the Naïve Bayes method by using a training dataset and then checks model accuracy with the testing dataset by using a confusion matrix. Finally, this

system displays the results that are SQL Injection or not.

## 7. Performance Evaluation

The machine learning method includes performance evaluation as a crucial step. It is crucial to evaluate how machine learning models generalize on test data in order to confidently trust their predictions. A confusion matrix is used to check the performance evaluation of the proposed system.

### 7.1. Confusion Matrix

A confusion Matrix is a diagram that illustrates the differences between actual and predicted values. It measures how well a machine learning classification model is performing.

**Table 3. Confusion Matrix**

| Predict / Actual | SQLInjection(Positive) | SQL(Negative) |
|---|---|---|
| SQLInjection | TP | FN |
| SQL | FP | TN |

**Table 4. Types of results are produced by prediction**

| Prediction Result | Explanation |
|---|---|
| TP (True Positive) | predicted SQL Injection and are actually SQL Injection |
| FP (False Positive) | predicted no SQL Injection and are actually SQL Injection |
| TN (True Negative) | predicted no SQL Injection and are actually no SQL Injection. |
| FN (False Negative) | predicted SQL Injection and are actually no SQL Injection |

This paper focuses on SQL Injection detection system because SQL Injection attack is the serious threat that describes in OWASP to exploit the Web App. The following figure 3. shows the performance evaluation of the system. Results are evaluated by four evaluation methods (accuracy, precision, recall and F-measure) called confusion matrix. F-measure is the mean of recall and precision, that is F = 2PR/ (P+R).

**Table 5. Evaluation Result of the Proposed System**

| | Precision | Recall | F1-score |
|---|---|---|---|
| Naïve Bayes Algorithm | 86% | 97% | 91% |


**Figure 3: Evaluation of the Proposed System**

## 8. Conclusion

It is crucial to identify and shield SQLI vulnerabilities from an attack since SQLIA is one of the most significant web vulnerabilities. In this proposed system, a Nave Bayes-based machine learning algorithm is used to recognize SQL injection attacks. This method uses a classifier to identify fraudulent queries. The Naïve Bayes Classifier is one of the most simple and effective classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions. This method uses a classifier to identify fraudulent queries. The proposed classifier classifies the test set with 88% accuracy. The suggested approach can be improved to detect various types of SQL injection attacks by appropriately extracting features.

## References

[1] Sangeeta, S Nagasundari and PrasadB Honnavali, “SQL Injection Attack Detection using ResNet,” IEEE – 45670, Int Conf. 10th International Conference on Computing, Communication and Networking Technologies,2019

[2] Muhammad Amirulluqman Azman, Mohd Fadzli Marhusin and Rossilawati Sulaiman, “Machine Learning-Based Technique to Detect SQL Injection Attack,” Journal-of-Computer-Science-1549-3636, 2021.

[3] M.KarthiKeyan, “An Efficient Technique For Preventing SQL Injection Attack Using Pattern Matching Algorithm,” IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), 2013.

[4] Musaab Hasan, Zayed Balbahaith, and Mohammed Tarique, “Detection of SQL Injection Attacks: A Machine Learning Approach,” International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2019.

[5] Anamika Joshiand and Geetha V, " SQL Injection Detection using Machine Learning," International Conference on Control, Instrumentation, Communication and Computational Technologies, 2014.

[6] Ryohei Komiya, Incheon Paik, Masayuki Hisada, “Classification of Malicious Web Code by Machine Learning,” Awareness Science and Technology (iCAST), 2011 3rd International Conference on, vol., no., pp.406,411, 27-30 Sept. 2011.

[7] G.T.Buehrer, RW.Weide, and P.AG.Sivilotti, "Using Parse Tree Validation to Prevent SQL Injection Attacks," International Workshop on Software Engineering and Middleware (SEM), 2005. [8] Sonali Mishra “SQL Injection Detection Using Machine Learning,” 2019

[9] Z.Su and G.Wassermann " SQL Injection Detection Using Machine Learning," The 33rd Annual Symposium on Principles of Programming Language (POPL 2006), Jan 2006.

[10] Sql Injection Dataset https://www.kaggle.com/datasets/syedsaqlainhussain/sql-injection-dataset

# Data Deduplication for Myanmar Language Storage using Secure Hash Algorithm

Thae Nu Aye, Tin Thein Thwel
*University of Computer Studies, Taungoo, University of Computer Studies, Yangon*
*thaenuaye@ucsy.edu.mm, tintheinthwel@ucsy.edu.mm*

## Abstract

*There is a vast amount of duplicated or redundant data in storage systems. The existing data deduplication attempted to reduce the storage spaces in file-level, sub-file-level data storage in terms of byte-level. There is also a need to reduce content level data deduplication, especially in Myanmar language contents. This study aims to deduplicate the data for sentences written in Burmese. The system accepts Myanmar sentences as input and uses Text Splitter to segment the input file into chunks according to the whitespace. Input a separated chunk into the ChunkID generator, which is using the Secure Hash Algorithm (SHA1). The system will search for duplicate phrases, and then we will work on reducing those duplicate phrases.*

## 1. Introduction

The amount of storage needed by servers grows along with the amount of internet usage, raising the possibility of duplicate data, especially in situations where persistent storage must be kept for a long time. As a result, the storage environment presents an intriguing research topic for data deduplication. Deduplication is a method for reducing storage space by removing redundant data copies kept in various places. This method decreases the amount of data stored on each individual device while also decreasing storage cost.

There is currently a significant amount of redundant and redundant data in storage devices. Data duplication happens both within and between versions of the same file. These vast quantities of duplication of data require more storage space and power, which significantly reduces storage utilization. Data deduplication can get rid of duplicated sections or whole files as well as multiple copies of the same data. Unique data chunks are found, and by replacing redundant chunks with new ones using the right strategy, duplicate data is removed. Users of the Burmese language typically use space however they see fit. In order to reduce data, we will try to suggest Myanmar Language Deduplication in this paper using String Tokenizer and the Data Deduplication method. We aim to remove redundant data in order to reduce storage space.

- We identify the same data by using the advantages of cryptographic technology.
- Additionally, we strive to improve data processing speed and efficiency by eliminating unnecessary data.
- We aim to obtain an efficient mechanism for searching identical data in the existing storage for Myanmar.

## 2. Related Work

M. Lillibridge et al. described the problem statement as "chunk-lookup disk bottleneck/full chunk indexing encountered in in-line deduplication" and tried to solve it using sampling, sparse indexing, and chunk locality. However, because it is assumed that if two segments share one chunk, it is likely that it will be shared with other chunks, only a limited number of segments are deduplicated, and they can't do fully deduplication because they sometimes store duplicate chunks.

Ohnmar Htun, the author of “Phonetic similarities of Myanmar Internationalized Domain Names (IDNs)” explained about phonetic similarity for Burmese language and SDF the method. This paper proposed to retrieve phonetically similar Myanmar IDNs, IPA (International Phonetic Alphabet)-Soundex functions were used for matching character values based on their phonetic similarities of Burmese. IDNs are domain names represented by

characters other than the traditional ASCII characters (a through z). Such domain names could contain letters or characters. The normalized similarity method is capable of measuring similarity not only in a single language, but also in a cross-language comparison. This paper used Stepped Distance Function (SDF), to calculate the phonetic and orthographic similarity of two letters.

Walter Santos solved the problem of identification of replicas in databases with a Parallel deduplication algorithm using the filter-stream model and only considered databases, not sub-file levels of the file, which have their respective secure hash codes.

# 3. Proposed System

## 3.1. Data Deduplication

Data deduplication is the method of storing and/or sending only unique data after examining a set of data or byte stream at the sub-file level. The approach taken to assess, identify, track, and prevent duplication is the fundamental basis for understanding the term "sub-file level". The deduplication procedure entails updating tracking data, storing and/or sending new and unique data and ignoring any duplicate data.

Data is compressed by being encoded in order to use less storage space. While loss data compression methods permanently discard some of the original data, lossless data compression techniques enable exact reconstruction of the original data from the compressed data. Data deduplication is a process that gets rid of extra copies of data and drastically reduces the amount of storage space needed. Deduplication can be implemented as a background process to remove duplicates after the data has been written to disk or as an inline process to remove duplicates as the data is being written into the storage system. By removing duplicate data blocks and storing only unique data blocks, deduplication works. The deduplication of the Myanmar language was emphasized in this proposal. Data deduplication's mechanism of operation: Deduplication divides an incoming data stream into distinct data segments, then uniquely identifies the segments and compares them to previously saved data. The segment is saved on disk if it is distinct. However, a reference is made to the incoming data segment, and it isn't stored if the segment is a duplicate of what has already been stored.

## 3.2. File Chunking

The chunking algorithm divides a file into discrete units known as "chunks" in the context of data deduplication. It can assist in reducing the amount of data sent over the network or removing duplicate copies of repeating data from storage. Each chunk has a header with some parameters in it (e.g., type, ChunkID, size, etc.). Chunking allows for the comparison of a large file to be split up into smaller pieces, making it easier and faster to identify duplicate chunks, which can increase the efficiency of data deduplication. Additionally, since the chunking procedure can be easily undone, the original data's integrity is unaffected.

## 3.3. Proposed System

### 3.3.1. System Architecture Overview

Figure 1 illustrates the proposed system's system's architectural layout. The following are the system's principal parts: The text splitter, ChunkID generator, duplicate finder, metadata, and storage are the system's major components. Txt and docx are two possible file types for input in the Burmese language. File Chunker splits input files into chunks of variable length. The File clunker's output chunks are used by the ChunkID Generator to generate ChunkID by applying the secure hash function SHA1, which is renowned for its resistance to hash collisions. Duplicate Finder uses the chunk ID to determine whether or not that chunk ID is already present in the metadata. The metadata maintains the hash code of the already stored file information, including ChunkID. Filename, chunk serial number, Chunk_ID, and Number_of_Chunk are all part of the file information. The storage space contains the contents of the data chunk. After deduplication the data or files, the system uses Reconstructor to restore the files or folders to their original state. It can reconstruct the desired file from the original file. When the reconstructed file has been deleted, the Reconstructor can reconstruct it again from the metadata and chunks. This means that the system can also recover a file that the user has accidentally deleted.

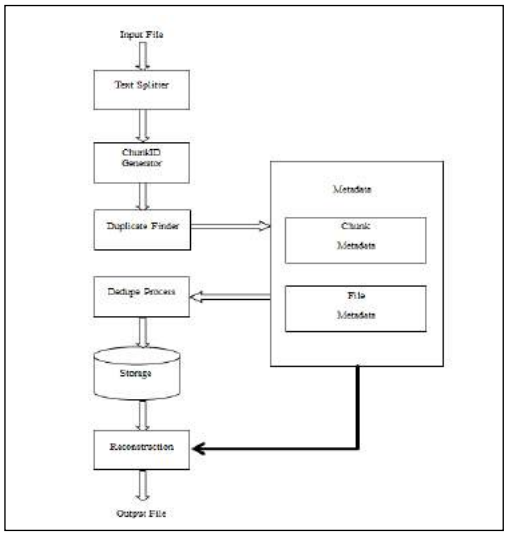



**Figure 1. Proposed system architecture**

### 3.3.2. Chunking (Text Splitter)

A file is split into segments or chunks by the process of "chunking". Python's split method can be used to break down text and make it easier for computers to understand. Almost every word written in Myanmar language documents is already formatted and segmented with spaces. This process, which makes text analysis easier, is called tokenization. In this system, a simple string Tokenizer from Natural Language Processing (NLP) is used.

### 3.3.3. Hashing (SHA1)

To generate the Chunk_ID, the Chunk_ID generator uses SHA1 which produces 160 bits signature for each chunk.

### 3.3.4. Deduplication (Duplicate Finder)

Its finds the duplicate Chunk_ID in the ChunkIDMetadata. The data deduplication procedure and algorithm are described below.

S1: Input File.
S2: Separate the file into Chunks by using Text Splitter.
S3: Each Chunk is identified by Hash value.
S4: Check whether duplicate chunks exist or not.
S5: If ChunkID has already existed, append it to the existing metadata.
S6: Else, insert the ChunkID into the metadata and store the chunk data in memory.

**Figure 2. Data Deduplication Procedure**

```
Duplicate Finder (ChunkID, ChunkIDMetadata, Chunk)
Begin
    found ⟵ searchInchunklib (ChunkID)
    if found in Chunklib then
        get index of that chunkID's content
        updateFileMetadata (chunkIdIndex, ChunkIDMetadata)
    else
        updateChunklib (ChunkID)
        updateFileMetadata (chunkIdIndex, ChunkIDMetadata)
        store (chunk)
End
```

**Figure 3. Data Deduplication Algorithm**

## 4. Experimental Result

Figure 4 shows the space reduction ratio relevant in most situations, which reflects all of the complementary capacity optimization technologies actually used. The chunkID's in our implementation are generated using the SHA1 hash algorithm. Internally, the data is saved as files. A data deduplication ratio over a particular time period is the number of bytes input to a process divided by its output. This ratio provides an indication of how well the deduplication process has worked and thus gives us a valuable insight into how much disk space we are able to save. With the use of the SHA1 hash algorithm to generate chunkID's, we can achieve an efficient data deduplication ratio over any given period. By taking advantage of the fact that data is typically composed of multiple copies and that each file contains similar information, we can use deduplication technology to detect and store only one copy of each block of data.

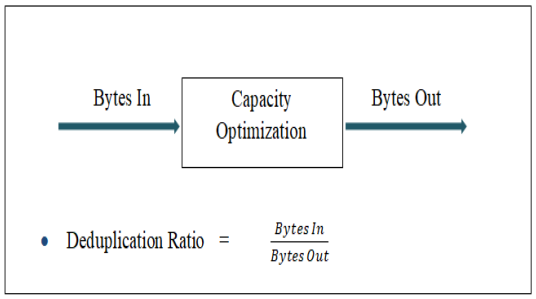



**Figure 4. Space Reduction Ratio**

Space Reduction Ratio (%) = 1-(1/Deduplication Ratio)*100 (1)

**For 100% duplicate file (.docx, .txt):** Tested results for the 100% deduplicate .docx and .txt files are shown in figure 5 and 6 below. File1

contains text content that has never been saved, and File2 and File3 contains the same content as File1. After duplicating the File1 for the first time, the storage of the ChunkLib increased and the storage of metadata increased. This means that no matter how many times you duplicate them, the storage size of ChunkLib and its metadata will not increase. The size of files 1, 2 and 3 is the same as the contents of File1, so you don't need to worry about increasing file sizes.

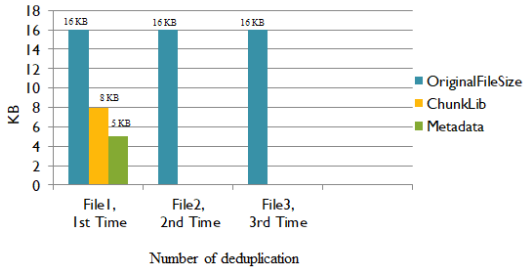



**Figure 5. Deduplication for 100% duplicate file (.docx)**

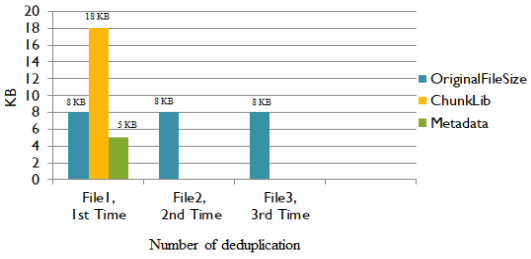



**Figure 6. Deduplication for 100% duplicate file (.txt)**

If we save the same files for the first time, ChunkLib and metadata increase.

**For 50% duplicate file (.docx, .txt):** Tested results for the 50% deduplicate .docx and .txt files are shown in figure below.

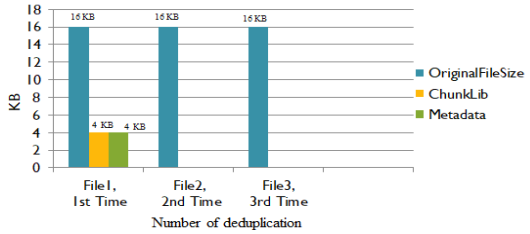



**Figure 7. Deduplication for 50% duplicate file (.docx)**

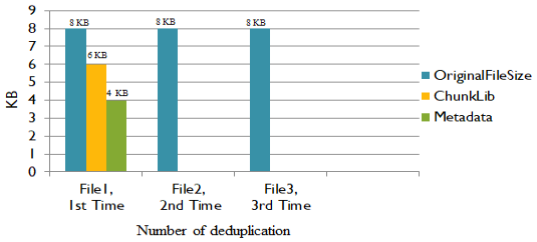



**Figure 8. Deduplication for 50% duplicate file (.txt)**

If we save files that are half the same data as the previously stored file, the storage space will increase in only half of that file with metadata in Figure 7 and 8 shows that the file tested here is half the size and contains the same content as previously saved and duplicated files, so the weight savings are significant.

**For 25% duplicate file (.docx, .txt):** Tested results for the 25% deduplicate .docx and .txt files are shown in figure below. Since the content of this file is a quarter of the previously duplicated file, it can be seen that the chunks in the file have been decreased in size to one-fourth of their original size in this image. Less than half of the total has been increased by ChunkLib.

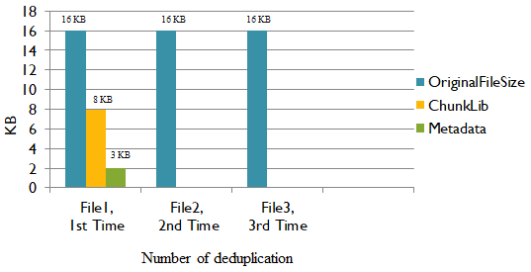



**Figure 9. Deduplication for 25% duplicate file (.docx)**

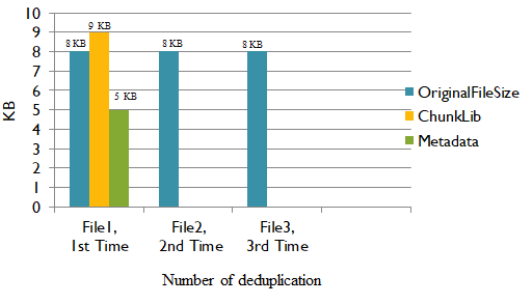



**Figure 10. Deduplication for 25% duplicate file (.txt)**

Moreover, if the files containing only one quarter duplicated data, ¾ storage space will increase.

Then, if we try to store the same contents as the previously stored file, the storage space will not increase, but only for metadata.

## 5. Conclusion

The proposed system can be used effectively by using data deduplication to store files written in Burmese. It can reduce the storage space as its potential purpose. Eliminating redundant data can significantly shrink storage requirements and improve bandwidth efficiency. Due to the advantages of proposed ChunkID based Segmentation mechanism in this work, even the reconstructed file is deleted by the user; it can restore that file whenever it is needed as it retains the concerned metadata in permanent matter. Because of using effective Hash Algorithm for creating Chunk_ID to check duplicate data, it can be more effective in finding duplicate data in the existing storage.

## References

[1] A Technical White Paper: “Data DeDuplication Background”.

[2] Dirk Meister, “Advanced Data Deduplication Techniques and their application”, Dissertation Johannes Gutenberg University Mainz, March 2013.

[3] Matthew Brise, Quantum Gideon Senderow, “Data Deduplication methods for Achieving Data Efficiency”, Journal of SNIA Education Committee.

[4] Qinlu He, Zhanhuai Li, Xiao Zhang, “Data Deduplication Techniques”, Department of Computer Science North-western Poly technical University Xi'an, P.R.

[5] Tom Sas – Hewlett-Packard, “Understanding Data Deduplication”, and SNIA Program in Storage Networking Industry Association, 2010.

[6] Tin Thein Thwel, Ni Lar Thein, “Data Deduplication using B+ Tree Indexing”, Research in University of Computer Studies, Yangon, 2010.

.

# Detecting Web Application's Broken Authentication by Using Combinatorial Algorithm

Ohnmar Thet, Zin Thu Thu Myint
*University of Computer Studies, Yangon*
*ohnmarthet@ucsy.edu.mm, zinthuthumyint@ucsy.edu.mm*

## Abstract

*To confirm the identity of the web application's users, the authentication to check the uniqueness of the user and session management functions to process the request securely between web server and the client user are used. The attacker can enter the system via the passwords, keys or sessions tokens if the system implementation is not designed securely. Only the authenticated user can access the private data based on the authenticated session of the system. This system is especially aimed to conduct a web application security using simple brute force and dictionary attack in broken authentication with combinatorial algorithm. Although there is no best way guard the user's computer security, but there is always try to develop. This proposed system guesses the password of the vulnerable web site using Dictionary based attack and Brute force attack. The experimental result is based on length of password, time consuming and type of methods. The 6 characters length of password and only numbers (for example - 456789), the possible combinations for brute force is 10^ 6 =1,000,000 (for 10 digits from 0 to 10). C# programming language with Microsoft SQL Server Database Engine are used to implement this system.*

**Keywords:** Authentication, brute force, diction attack, combinatorial algorithm

## 1. Introduction

In the act of using web applications, the sensitive private information that people give are stored on that web applications. On the other hand, some unethical attackers exploit the web application to gain illegitimate access and illegal use of personal information and make other things such as identity theft, privacy violation, and other cyber-attacks. These illegal facts allow the attackers to do whatever they want through the weaknesses of the web application. The weak point of the web application called vulnerability is caused by unawareness of the developers who cannot be controlled verification of the user inputs, appropriate validation approaches, and so on. Because of those reasons, detection of vulnerability is needed more.

This system will check whether the system provided a secure login function or not and to know how secure the authentication is. And it will give a message on whether or not the authentication is secured. In this system, the first step is to guess the password of the login page from the proposed web application. This system will guess the password in two ways: the dictionary based and brute-force attacks. In a dictionary attack, the system will use a wordlist of common passwords/phrases, add user-specific words (social engineered information), add "pass phrases" to the wordlists, generate variants for each word using several common "complexification" patterns. In dictionary attack, conduct the password of login page with the dictionary word in word file that is pre-created and pass the system via the success guess password. For brute force attack, the system will use target language letter frequency combinations, use a wordlist (letters/digits/symbols) as the building blocks, try to crack a passphrase. In brute-force attack, generating the pattern using combination algorithm and guess the possible password to pass through the login page of web page. The outcome of the security testing will be displayed as a graph that were noticed during the detection [3].

## 2. Related Work

As the first study, it detected and defined against brute force attacks with Power Rules. Signal Sciences facilitate DevOps and security

teams to encounter and safeguard against brute force attacks with Power Rules. The objective is to detect and defend against brute force attacks with Power Rules There are two findings in this paper. The first one is to tried to detect brute force attack by defining threads hold on failed login requests from IP addresses and the second is that if it is reached the threshold specified in the Power Rule, it defined as brute-force attack [10].

The second study showed that Brute force attacks are used to login to network services with pairs of username and passwords. S.Saito, K. Maruhashi, M. Takenaka and S. Torii described about the Detection and Prevention of Brute Force Attacks with Discipline IPs from IDS Logs For network service administrators. This paper highlights about the Remote Desktop Protocol (RDP) to prevent brute-force attack by analyzing the IDS logs and regularity of login trials This paper also presented that the network administrators have to set the limits of login attempts and block the traffic of brute force attacks at the entry point with an intrusion prevention system (IPS). The aim is to implement IDS and IPS based on integrated log file for multiple sites of distributed area.

As third study, L. Bošnjak, J. Sreš and B. Brumen tried to to demonstrate how easy it is to crack most of the user-created passwords using simple and predictable patterns and to perform an analysis of the cracked and uncracked passwords and measured their strength. [7] The system in this paper tried to hash the password in order to get difficulty in guessing the user passwords. [9] The results have shown that even a single low to mid-range modern GPU can crack over 95% of passwords in just few days, while a more dedicated system can crack all but the strongest 0.5% of them.

## 3. Background Theory

There is a non-profit foundation that works to improve the security of the web application is known as The Open Web Application Security Project (OWASP). OWASP Top 10 Attacks are: [8]

1) Injection
2) Broken Authentication and Session Management
3) Cross-Site Scripting (XSS)
4) Insecure Direct Object Reference
5) Security Misconfiguration
6) Sensitive Data Exposure
7) Missing Function Level Access Control
8) Cross-Site Request Forgery (CSRF)
9) Using Components with Know Vulnerabilities
10) Un-validated Redirect and Forward

Nowadays, many people use the internet for more than one purposes. A web application is composed of a web server and web browser in other terms client-side and server-side. When people access a web application from any one browser, firstly they send a request to the web server and then this web server responds this request to the web application server and processing continued tasks Today, web applications are popular for people because these have many advantages: easily use and cost effective for users.

Maintaining web application security is the important case for users because web application may have vulnerabilities. Web application vulnerabilities are weakness of this web application and can find many kinds of reasons. For example, application developer errors within coding, application design weakness and so on. So, attackers can be tried by using these vulnerabilities to exploit system for getting privileges and personal information. This system will study the password attacking using brute force attack and dictionary-based attack.

### 3.1. Brute Force Attack

A brute force attack tries to access the login info, encryption keys or find a hidden web page by using trial-and-error approach. Hackers try to attempt the guess correct, they generate all possible combinations of the characters. [7] Brute force attacks means using extreme forceful attempts to try and ‘force’ their way into the private account(s). This is an old attack method, but it's still effective and popular for hackers. Because of depending on the length and complexity of the password, cracking can take anywhere from a few seconds to many years [6].

Brute force attackers have to put in a bit of effort to make these schemes improve.

Here's how hackers benefit from brute force attacks:

- Earning from ads or collecting activity data
- Stealing personal data and valuables
- Spreading malware to cause disruptions
- Hijacking the system for malicious activity
- Ruining a website's reputation

Each brute force attack can use different methods to crack the sensitive data. The following attacks are the most popular brute force methods:

- Simple Brute Force Attacks
- Dictionary Attacks
- Hybrid Brute Force Attacks
- Reverse Brute Force Attacks
- Credential Stuffing

## 3.2. Dictionary Based Attack

A dictionary attack can crack a password-protected computer system, network or other IT resource by consistently trying every word in a dictionary as a password. To find the key necessary to decrypt an encrypted message or document, a dictionary based attack can also be used. Dictionary based attacks can crack passwords easily because most of the computer system users and businesses organization keep on using system default keywords as passwords. The weak point of dictionary attack is that it cannot crack the password included multiple words and some passwords that consist of letters (both uppercase and lowercase) and number in random character combinations. The brute-force attack is effective in some strong password systems when possible combination of all characters and spaces is generated up to a certain maximum length. Strong passwords cannot be easily guessed, and they are not included in the pre-created password library. it takes a certain amount of time to produce brute force results Because a dictionary attack's guess attempts are limited to a pre-created wordlist, it is quite hard to crack non-predictable passwords [7].

### Work of Dictionary Based Attack

Preselected lists of words and phrases (such as my white walking shoes, my favorite hair color) are used to guess passwords in dictionary attack. It works when the users use from a basic list of passwords, such as "password," "admin1" and "12345678."

The lists of predictable patterns (words) in dictionary attack can be different by the region. For example, hackers trying to execute a dictionary attack on a New York-based group of targets, the test phrases might look like "kicks2020" or "newyorkknicks111." Attackers mix the words related to company names, types of interest, name of cities live in, name of department, and other specific contents when construction the dictionary library files.

The lists that are used in dictionary-based attack cannot expand compared with brute-force attacks, but they can develop into quite large lists of words. Processing and testing all these passwords are difficult to do manually. Therefore, some extra technology is normally required to boost up the process. Sometimes, attackers use free related software from the free site, such as password dictionaries or other brute-force attack tools. [11]

There are online and offline dictionary attacks and these attacks depend on whether the account, network or device. In an online attack, the attacker must be aware of the number of attempts when trying to guess the correct password. The online dictionary attack may detect by the system administrator, user account manager, system user or intrusion detection system and can limit the password attempt. If none of these are happen, the system automatically locks the attacker out.

Dictionary-based attacks can be successful when passwords are shorter prioritized list. Sophisticated attackers can also disable the intrusion detection features or number of login password attempts.

There are few limitations for hacker in offline dictionary attacks when it comes to the number of passwords they can predict. However, accessing to the password storage file is needed to try an offline dictionary attack. Only then can a dictionary attack be executed in an offline setting of the system.

## 3.3. Brute-force attack vs. dictionary attack

The number of password permutations that are attempted make the main difference between brute force and dictionary attack.

**Brute-force attacks**

Because of the systematic approach to generate all possible combination of ASCII characters, brute-force attack will need a large amount of time to complete. [1]

To make a combination of five-digit lock is a most familiar non-tech example of the difference. An attacker can use the brute force attack to try every possible permutation for the lock of five-digit. A lock of five-digit with values from zero to nine has the permutations of 100,000 possible [3].

**Dictionary attacks**

Dictionary attack will use a list of passwords and try numbers of attempts to break into the system. Instead of trying to input every possible permutation, an attacker uses a dictionary approach that would attempts all the permutations in its pre-generated library [5].

Sequential digit passcodes, like "123456," and static passcodes, like "000000," would be tested in dictionary attack. If the six-digit permutation is especially unique, the dictionary attack likely would not guess that six digits. Dictionary attacks can predict that what percentage of the users or accounts they try to attack will be vulnerable and will have an easily detectable six-digit passcode.

### 3.4 Combination Algorithm for Brute Force Attack

```
function combination:
pass in:inputArray, combinationArray, start, end, index, r
        if index is equal to r:
        for each element in combinationArray:
        print each element
        return
        for i = start:
        if i <=end and end – i+1 > r - index:
        combinationArray[index] = inputArray[i]
        call combination function again with updated parameter
```

**Figure 3.1 The combinatorial algorithm**

The Combination is a kind of arrangement of some given objects. In mathematical terms, Combination is a set of choices/selection of items from a unique set of items/objects. For example, you're given a bag with 5 different colors and asked to generate a pattern with any 3 colors. You can also pick any 3 colors out of 4, then arrange them in different orders [2].

Let's assume the colors are RGYBI (R= Red, G= Green, Y= Yellow, B= Blue, I= Indigo). So, the possible pattern can be RGB, RGY, etc [1].

**Formula of combination:**

$$\mathbf{C}^{\mathbf{n}}_{\mathbf{r}} = \binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (1)$$

- This has 5 colors meaning n = 5, and at any given time, we need to pick any 3. So, r = 3.

  After calculating, we get,

$$_5C_3 = \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{120}{6*2} = 10$$

- total of **10** color combinations [2]

## 4. Overview of the Proposed System

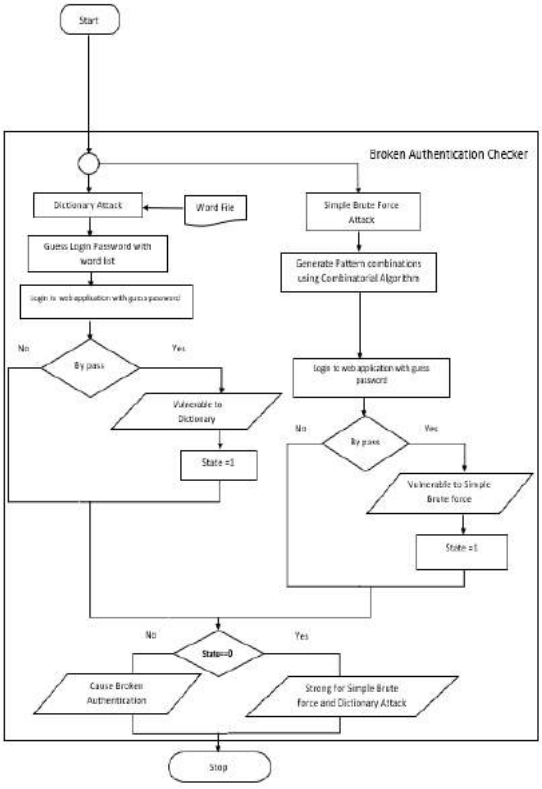



**Figure 4.1 System Flow**

In this system, the password guessing process will guess the vulnerable website's password by two ways: dictionary based and brute force attack. Set the state to 0 and start checking with dictionary and brute force attack. In dictionary attack, the proposed attacking system will guess the password using the words in downloaded word file. Bypass that website with the guess password, it is set the state to 1 and if it is not passed, it is set the state to 0. And at the same time, generate the pattern combination using combination algorithm in brute force. Bypass that website with the guess password, it is set the state to 1. If it is not passed, it is set the state to 0. After bypassing with one of two methods, return the message "Strong for simple brute force and dictionary attack" depending on the state. If not return the message "Cause broken authentication".

### 4.1. Dictionary for Attack

The dictionary for the proposed attacking system is downloaded from "Crack Station's Password Cracking Dictionary Site" [4]. The passwords are obtained from some of the websites that are affected by the data breaches. Sample contents of downloaded dictionary are shown in following Figure 4.2.

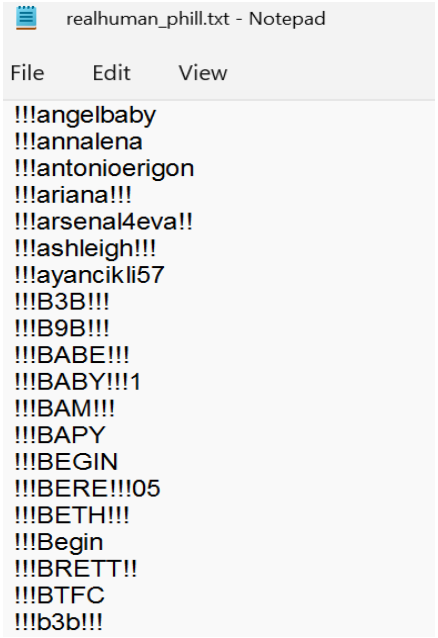



**Figure 4.2. Sample Password Contents of Dictionary Based Attack**

## 5. Experimental Results

This system guesses the password of the vulnerable web site using Dictionary based attack and Brute force attack. The experimental result is based on length of password, time consuming and type of methods. The 6 characters length of password and only numbers (for example - 456789), the possible combinations for brute force is 10^ 6 =1,000,000 (10 is for 10 digits from 0 to 10 and power 6 is for length of the password). The experiment is shown in following Table 5.1 and Figure 5.1.

**Table 5.1 Cracking Time for Brute Force and Dictionary Attacks**

| Attacks / Password Length | Brute Force Attack(ms) ( combination time+ estimate cracking time) | Dictionary Based Attack(ms) |
|---|---|---|
| Password Length 4 | 8784.378529 | 1944.3195 |
| Password Length 5 | 85212.78191 | 924.1135 |
| Password Length 6 | 3449000 | 1729.2016 |
| Password Length 7 | 32,423,879.70963 | 805.0287 |
| Password Length 8 | 3,047,844.69 | 2223.8646 |

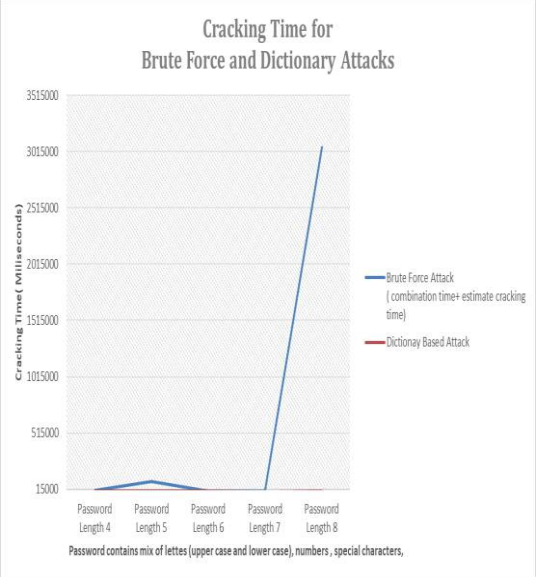



**Figure 5.1. Line Graph of Cracking Time for Brute Force and Dictionary Attacks**

In figure 5.1, testing is made for eight times in each method on different password length. In the experiment of dictionary-based method is obvious to guess the password in minimal time. Brute force attack cannot guess the password during the system defined limited time interval

(900,000ms = 15 minutes). In real situation, brute force attack can guess all types of passwords whether complex or simple. But the time consuming may be in minimal time to decade, century and so on.[5] But in the thesis testing, time interval is needed to limit. So, the proposed system defined the time interval as 15 minutes (900,000ms). According the experiment, the dictionary-based method is fast to guess the password which is contained in the dictionary. But it can't effort the complex passwords which are not contain in its dictionary. So, both methods have pros and cons by their nature.

## 6. Conclusion

The proposed system provided the secure authentication that is strong for brute force and dictionary attack. This system implements the Brute Force Attack by using combination algorithm, it is easy to implement and can try out all the possible combinations of ASCII character. It collects the dictionary words from the file given in crackstation.net to implement the Dictionary Attack. It uses these kinds of attacks as *white attack* to test the security of web application upon Broken Authentication. So, this system can detect the vulnerability on web application for Broken Authentication by forcing only Brute force and Dictionary Attack.

## References


[1] A. Christy Sathyabama University, Chennai, Tamilnadu, India, D. Saravanan Sathyabama University, Chennai, Tamilnadu, India, "An Analysis of Markov Password Against Brute Force Attack for Effective Web Applications", January 2014. Pg-5824, 5825

[2] ALBERT NJENHUIS and HERBERT S. WILF, "Combinatorial Algorithms For Computers and Calculators, Second Edition", pg-45-61, 1978.

[3] B. R. Heap, "[4[s by interchanges", The Computer Journal, Volume 6, Issue 3, November 1963, Pages 293–298 01 November 1963.

[4] Crack Word list text file from "https://crackstation.net/crackstation-wordlist-password-cracking-dictionary.htm"

[5] David C. Feldmeier and Philip R. Karn Bellcore, "UNIX Password Security- Ten Years Later", 445 South Street Morristown, NJ 07960, pg-2,3,4.

[6] Konark Truptiben Dave, "Brute-force Attack "Seeking but Distressing", Department of Computer Engg. & Information Tech. C.U.Shah Technical Institute of Diploma Studies, Surendranagar-363001, Gujarat, India, International Journal of Innovations in Engineering and Technology (IJIET), pg-75.

[7] L. Bošnjak, J. Sreš and B. Brumen, "Brute-force and dictionary attack on hashed real-world passwords", May 2018, pg-4,5.

[8] OWASP Top Ten, [online] available: https://owasp.org/www-project-top-ten/, visit on February 2022.

[9] S. Vaithyasubramanian Sathyabama University, Chennai, Tamilnadu, India Satomi Saito Koji Maruhashi1 Masahiko Takenaka1 Satoru Torii1"TOPASE: Detection and Prevention of Brute Force Attacks with Displined IPs from IDs Logs, November, 2015,pg 217,218.

[10] Signal Sciences, "Brute Force Attack Protection". pg-1,2.

[11] Tobias Lundberg, "Comparison of Automated Password Guessing Strategies", Master of Science Thesis in Electrical Engineering Department of Electrical Engineering, Linköping University, 2019, pg-23,24.