FACE MASK DETECTION BY USING CONVOLUTIONAL NEURAL NETWORK

EI CHERRY LWIN

M.C.Sc DECEMBER 2022

FACE MASK DETECTION BY USING CONVOLUTIONAL NEURAL NETWORK

BY

EI CHERRY LWIN

A Dissertation Submitted in Partial Fullfillment of the Requirement for the Degree of

Master of Computer Science
(M.C.Sc)

University of Computer Studies, Yangon
DECEMBER 2022

ACKNOWLEDGEMENTS

To accomplish this thesis, many things are essential like my best effort, hard work, and important ideas from specific supporters while completing this thesis.

Firstly, I would like to express my special thanks to Dr. Mie Mie Khin, Rector of the University of Computer Studies, Yangon, for her kind permission and encouragement to submit this dissertation.

I would like to thank course coordinators, Dr. Si Si Mar Win and Dr. Tin Zar Thaw, Professors, Faculty of Computer Science, University of Computer Studies, Yangon, for their superior suggestions, administrative supports and encouragement during my academic study.

I would like to express my profound gratitude to my thesis supervisor, Dr. Myat Mon Kyaw, Associated Professor, Faculty of Information Science, University of Computer Studies, Yangon, for her close supervision, cooperation during my effort of study, understanding and generous encouragements during a long period of study during the progress of this work.

I also wish to express my deepest gratitude to Daw Aye Aye Khaing , Associated Professor, English Department, the University of Computer Studies,Yangon for her editing this theis from the language of point of view.

Moreover, I would like to extend my thanks to all my teachers who taught me throughout the master's course and my friends for their cooperation.

I especially thank to my parents, all of my colleagues, and friends for their encouragement and help during my thesis.

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 porposed system is constructed with MobilenetV2 model. If a person is wearing a mask, the system displays the face region with a green anchor box. If a person is not wearing a mask, the face region is displayed with a red anchor box. This system can be merged with other applications at airports, railway stations, workplaces, schools, and other public places for safety.The accuracy is 82 % for testing images in dataset.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

The 209th report of the world health organization (WHO) published on 16th August 2020 reported that coronavirus disease (COVID-19) caused by acute respiratory syndrome (SARS-CoV2) has globally infected more than 6 million people and caused over 379,941 deaths worldwide . According to Carissa F. Etienne, Director, Pan American Health Organization (PAHO), the key to control COVID-19 pandemic is to maintain social distancing, improving surveillance and strengthening health systems . Recently, a study on understanding measures to tackle COVID-19 pandemic carried by the researchers at the University of Edinburgh reveals that wearing a face mask or other covering over the nose and mouth cuts the risk of Coronavirus spread by avoiding forward distance travelled by a person's exhaled breath by more than 90% . Steffen et al also carried an exhaustive study to compute the community-wide impact of mask use in general public, a portion of which may be asymptomatically infectious in New York and Washington. The findings reveal that nearly universal adoption (80%) of even weak masks (20% effective) could prevent 17–45% of projected deaths over two months in New Work and reduces the peak daily death rate by 34–58% . Their results strongly recommend the use of the face masks in general public to curtail the spread of Coronavirus. Further, with the reopening of countries from COVID-19 lockdown, Government and Public health agencies are recommending face mask as essential measures to keep us safe when venturing into public. To mandate the use of facemask, it becomes essential to devise some techniques that enforce individuals to apply a mask before exposure to public places.

1.1 Overview of the Thesis

Coronavirus COVID-19 pandemic is continuously spreading until now everywhere on the earth, and causing a severe health crisis. As the COVID-19 (Coronavirus) pandemic continues to spread, most of the world's population has suffered as a result. COVID-19 is a respiratory disorder that results in severe cases of pneumonia in affected individuals. The disease is acquired via direct contact with an infected person, as well as through salivation beads, respiratory droplets, or nasal droplets released when the infected individual coughs, sneezes, or breathes out the virus into an airspace . Globally, thousands of individuals die from the COVID-19 virus daily. A Coronavirus (COVID-19) report by the World Health

Organization (WHO) reveals that, as of 22 November 2021, there were 258 million confirmed cases of COVID-19 cases and 5,148,221 deaths worldwide . Therefore, people should wear face masks and keep a social distance to avoid viral spread of disease. An effective and efficient computer vision strategy intends to develop a real-time application that monitors individuals publicly, whether they are wearing face masks or not.

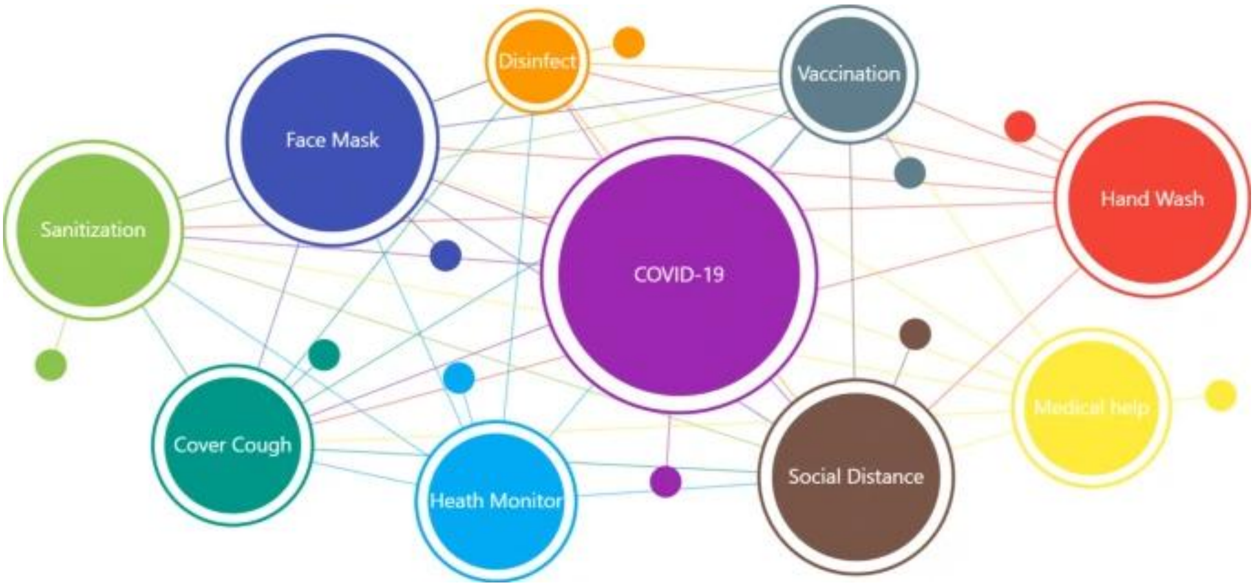


Figure 1.1: Precautions to avoid COVID-19

Face recognition (FR) systems are conventionally presented with primary facial features such as eyes, nose, and mouth that is non-occluded faces. However, a wide range of situations and circumstances impose that people wear masks in which faces are partially hidden or occluded. Such common situations include pandemics, laboratories, medical operations, or immoderate pollution. For instance, according to World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) , the best way to protect people from the COVID-19 virus and avoid spreading or being infected with the disease is wearing face masks and practicing social distancing. Accordingly, all countries in the world require that people wear a protective face mask in public places, which has driven a need to investigate and understand how such face recognition systems perform with masked faces.

However, implementing such safety guidelines earnestly challenges the existing security and authentication systems that rely on FR already put in place. Most of the recent algorithms have been proposed towards determining whether a face is occluded or not that is masked-face detection. Although saving people's lives is compelling, there is an urgent demand to authenticate persons wearing masks without the need to uncover them. For instance, premises access control and immigration points are among many locations where subjects make cooperative presentations to a camera, which raises a problem of face recognition because the occluded parts are necessary for face detection and recognition.

As a result of Covid-19 pandemic, Mask Detection has become a main safety measure that have more demand for somebody to put on facial masks, keep the social distance, and to wipe their hands with hand sanitizers, whereas a separate issue with social distance maintenance and cleanup square measure conveyed until currently, the thing of facial mask detection have not still exists dealt with enough. Identifying a facial maskthroughout this influenza pandemic could also be necessary significant protection and is the foremost important move forward in time once maintaining social distance is difficult to require care of recognizing a mask is crucial. So centers for illness management steered all pairs of years old and up to put on a mask publically different mutual distancing exists in some regions. The primary aim of this paper is to find the existence of a mask on human faces.

1.2 Motivation of the Thesis

COVID-19 is a highly contagious disease, and the WHO and other health agencies have recommmended that people use face mask tp prevent its transmission.The spread of covid is still evident in most places with new variants of the same virus. One thing that has become common while going out nowadays is the face mask. But still there are many people unaware of the situation and refusing / avoiding to wear a face mask.This system helps in identifying people who are not wearing a face mask.

1.3 Objectives of the Thesis

The objectives of the thesis are:

- To detect a person without a face mask for reducing the spread of COVID-19.
- To automate the process of identifying the people who are not wearing mask.
- To develop a surveillance system for COVID-19 precaution.
- To study mobilenetV2 and caffe face detector.
- To identify whether the person on image/video stream is wearing a face mask or not.
- To implement the trained model to detect masks in images.

1.4 Organization of the Thesis

This thesis is organized into five chapters.Chapter 1 includes introduction, overview of the thesis, motivation, objectives of the thesis. Chapter 2 describe deep learning and evaluation of system performance. Chapter 3 describes the architecture design and gives the algorithm used in the module design. Chapter 4 presents design and implementation of proposed system which includes system flow diagram, screen designs of the proposed system and experimental results. Finally, chapter 5 describes the conclusion, limitation and future extensions of the system.

CHAPTER 2

BACKGROUND THEORY

2.1 Theory of deep learning

Deep learning is a branch of machine learning in artificial intelligence that uses algorithms to give computers intelligence. These methods are motivated by the biological structure and function of the brain. Using deep learning, a computer model may learn to carry out categorization tasks directly from images, text, or sound. Deep learning models can achieve the highest levels of accuracy, occasionally even outperforming human ability. Models are trained using multi-layer neural network architectures and a large amount of labeled data. Deep learning is part of data science, which also includes statistics and predictive modeling. Deep learning makes it much quicker and simpler for data scientists to collect, analyze, and interpret vast amounts of data. Driverless automobiles must have deep learning in order to distinguish between a lamppost and a stop sign, which is a crucial component. This functionality enables voice control in consumer electronics including phones, tablets, TVs, and hands-free speakers.

2.2 Biological brains and artificial neural networks

Deep learning develops from artificial neural networks, which have been around since the 1940s. Neural networks are connected networks of processing units that resemble axons in a human brain. They are used to create artificial neurons. In a biological neuron, which can include millions of neurons, dendrites receive input signals from numerous adjacent neurons. The neuron's cell body, or soma, receives these modified signals and combines them before sending them to the axon. If the received input signal exceeds a predetermined threshold, the axon will send a signal to the nearby dendrites of other neurons[17].

With a few modifications for usability, artificial neuron units are based on biological neurons. Similar to dendrites, the input link to the neuron transmits attenuated or amplified input impulses from nearby neurons. The neuron receives the signals and passes them along. The neuron adds up the incoming signals and then determines what to output based on the total amount of information it has received. For instance, a binary threshold neuron generates an output value of 1 when the entire input exceeds a predetermined threshold; otherwise, the output is 0. Artificial neural networks employ a range of different neuronal types; the only distinction is the activation function that is applied to the complete input to form the neuron output.

2.3 The deep learning process

Since most deep learning techniques use neural network designs, deep learning models are frequently referred to as deep neural networks. Deep neural networks are neural networks with a significant number of hidden layers. Deep learning is the accumulation of many layers within a neural network over time, with performance increasing as the network's complexity increases. The data that is received by each layer of the network is processed in a different way, which then informs the layer above itTraditional neural networks only have 2-3 hidden layers, however deep neural networks can have up to 150. Massive amounts of labeled data and neural network topologies that derive features directly from the data rather than requiring laborious feature extraction are used to train deep learning models.The structure of a neural network is shown in Figure 2.1.

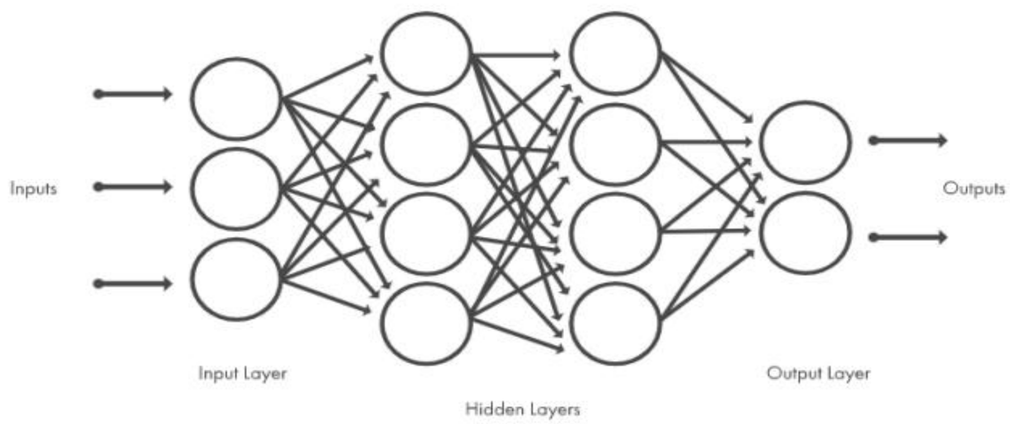


Figure 2.1 Layered neural network are made up of a series of interconnected nodes

2.4 Deep learning methods

There are numerous methods that can be utilized to build effective deep learning models. One of them is the rate of learning loss. When the weights are adjusted in response to the expected error, the learning rate hyperparameter controls how much the model alters. Prior to the learning process, it specifies the system or provides operational conditions. When learning rats are set too high, an unstable training process or the acquisition of a flawed set of weights may result. A long and unproductive training procedure could be the outcome of a slow learning rate. Another approach is transfer learning, which necessitates getting access to a network's internals and optimizing a model that has already been trained [18]. Users initially add new data to an existing network that already has previously unidentified classifications. With more specialized classification abilities, new tasks can be accomplished after the network has been upgraded. The benefit of this approach is that it requires a lot less data than previous approaches, cutting down computation time to minutes or hours.

As part of the training process, a developer must gather a massive data set and set up a network architecture capable of learning the features and model. This strategy is especially beneficial for newly developed applications and those that have a wide range of output formats. However, it is a less common strategy overall because it needs a large quantity of data and could take days or weeks to train. It has been demonstrated that the dropout strategy improves neuronal performance on supervised learning tasks in a variety of fields, such as speech recognition, document classification, and computational biology. This approach aims to solve the overfitting issue in networks with many parameters by randomly removing units and their connections from the neural network during training.

2.4.1 Workplace examples of deep learning

Applications for deep learning can be found in many different industries. Deep learning is being used by automotive professionals to automatically detect items like stop signs and traffic lights. Deep learning is also used to identify pedestrians, which reduces accidents. When determining whether troops are in safe or dangerous areas, it is employed in the aerospace and defense industries to discriminate between objects from satellites that locate points of interest. It is being used in the realm of medicine by cancer researchers to automatically detect malignant cells. A microscope that produces high-dimensional data that can be used to regularly train a deep learning system to recognize cancer cells has been improved by researchers. This device helps to improve worker safety near heavy machinery by automatically determining if persons or items are too close to it. Automatic hearing and voice translation are further applications of deep learning.

2.5 Related Work

Generally, most of the publication focus is on face construction and identity recognition when wearing face masks. In this research our focus is on recognizing the people who are not wearing face masks to help in decreasing the transmission and spreading of the COVID-19. Researchers and scientists have proved that wearing face masks help in minimizing the spreading rate of COVID-19.

Ge S., Li J., Ye Q., Luo Z built a model by using a dataset to find the unmasked and masked face. The dataset, which is called Masked Faces (MAFA), included 35,806 images of people wearing masks. The authors used a convolutional neural network to propose their model including three different modules, which are proposal, implementation, and authentication. Their work achieved 76.1% accuracy results [7].

Ejaz Md. S., Islam Md. R., Sifatullah M., Sarker A applied machine learning techniques to distinguish between people wearing face masks versus people not wearing face masks. They used Principal Component Analysis (PCA). The paper accomplished identifying people who are not wearing masks provides a better recognition rate in the Principal Component Analysis. Authors were able to find that extracting features from people wearing face masks is lesser than people who are not wearing face masks because of missing features for wearing mask which decrease the recognition rate. So, they concluded that traditional statistical algorithm Principal Component Analysis (PCA) is better for normal face recognition but not for masked face recognition. Finally, they also found that accuracy has much decreased after classifying people wearing a face mask, which gave an accuracy of 70% [4].

Mohamed Loey, Mohamed Hamed N. Taha, Gunasekaran Manogaran, and Nour Eldeen M. Khalifa created an approach to annotate and localize the medical face mask objects in real-life images. Their proposed model included two components: a feature extraction process using the ResNet-50 deep transfer learning model, and a medical face mask detection using YOLO v2. The authors were able to improve the detection performance by using the mean IoU. This helped in estimating the best number of anchor boxes. With that said, the authors were able to conclude that using Adam optimizer achieved the highest accuracy of 81% .

Xinbei Jiang, Tianhan Gao, Zichen Zhu and Yukang Zhao [15] proposed a system in Real Time Face Mask Detection using YOLOv3. The Properly Wearing Masked Face Detection Dataset (PWMFD) is used in the paper, which has 9205 images samples wearing masks. The relationships among channels are obtained by integrating the attention mechanism into Darknet53 using the SE block, so that the network can focus on the feature. In order to better describe the spatial difference between predicted and ground truth boxes and to improve the stability of bounding box regression, Glou loss is implemented. The extreme foreground-background class imbalance was solved using focal loss. The final results showed that SE-YOLOv3 is better than YOLOv3 and other state-of-the-art detectors on PWMFD. While comparing with YOLOv3, the proposed model achieved 8.6% higher mAP and detection speed.

Sunil Singh, Umang Ahuja, Munish Kumar, Krishna Kumar and Monika Sachdeva [16] proposed a system in Face Mask Detection using YOLOv3 and faster R-CNN models. This paper, draws bounding boxes on people on the screen in red or green color whether they are wearing a mask or not and keeps the ratio of people wearing masks on a daily basis.

CHAPTER 3

THE PROPOSED SYSTEM METHODOLOGY

3.1 Convolutional neural network (CNN)

Artificial neural networks are being utilized more frequently to process unstructured data, including speech, text, pictures, and audio. Convolutional neural networks (CNNs) are effective when given information that is this illogical. Convolutional neural networks may recognize important components within data when the data has a topology. In terms of architecture, CNNs are influenced by multi-layer perceptron systems. By establishing local connection restrictions between neurons in adjacent layers, CNN uses local spatial correlation. Convolutional neural networks' core function is to process data using the convolution technique. When any signal is convolution with another signal, a third signal is produced that may provide more information about the signal than the original signal [19].

3.2 CNN architectures

Convolutional and pooling (or subsampling) layers are organized into modules in all CNN architectures, which come in a range of sizes and shapes. These modules are followed by one or more fully connected layers, like a normal feedforward neural networkModules are typically piled on top of one another to build a deep model. Direct image delivery to the network is followed by multiple iterations of convolution and pooling processing. The results of these processes are then fed into one or more fully connected layers. The output layer receives the inputs from the layers above it, executes the calculations using its neurons, and then computes the output. Even though that this is the fundamental architecture that has been utilized the most in the literature, a number of design changes have been suggested recently with the aim of improving picture classification accuracy or reducing computing costs. The relationship from input layer to output layer is shown in Figure 3.1[20].

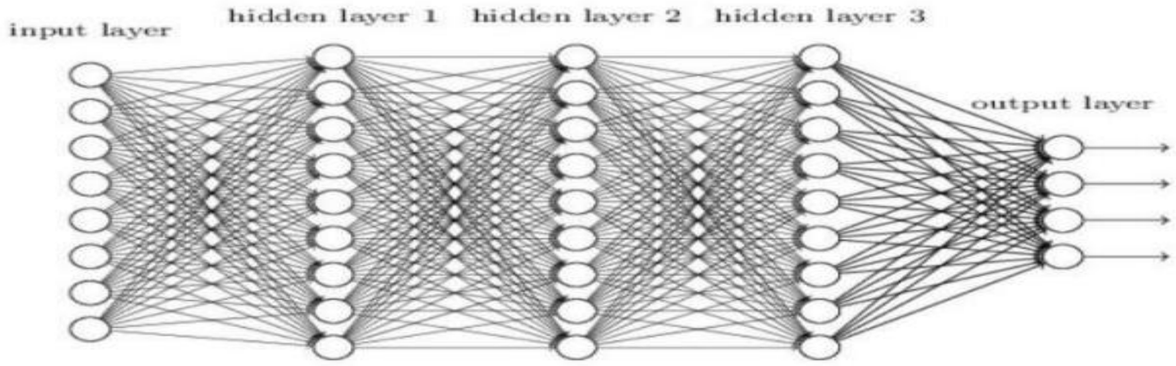


Figure 3.1. CNN architecture

3.3 Workflow of CNN

The higher performance of convolutional neural networks with picture, speech, or audio signals inputs distinguishes them from other neural networks. The three most common types of layers are fully connected, pooling, and convolutional layers. The convolutional layer is the top layer of a convolutional network. However, the fully connected layer is the final layer and can be followed by additional convolutional layers or pooling layers. The CNN becomes more complicated with each layer, detecting larger areas of the image. Early layers emphasize fundamental components like colors and boundaries. The target object is eventually identified as the visual data progresses through the CNN layers and starts to discern larger pieces or features of the item.

3.4 Design of Convolutional Neural Network

The basic architecture of convolutional neural networks is the same as other neural networks, with input, hidden and output layers. The only difference is that a convolutional neural network is a multilayer feedforward neural network. This means that it is made up of many hidden layers, which is the origin of the word "convolutional" in its name. Among these layers, there are four main layers, namely convolutional, Non Linearity, pooling, and fully connected layers. Convolutional layers are usually used as the first layer of a convolutional network, after which more convolutional layers or additional pooling layers can be added as needed, but the fully connected layer must be the last layer. A Non Linearity layer can be added after the convolution to increase the activation function as a way to add nonlinearity [21].

3.4.1 Convolutional layer

The convolutional layer is the most important component of a CNN because it is where most of the computation takes place. It requires input data, a filter, and a feature map, among other things. As an example, the input layer receives a color image made up of a 3D matrix of pixels. Accordingly, the input will have three dimensions: height, width, and depth, which correspond to a picture's rgb color space.The receptive fields of the image will be traversed by a feature detector, also known as a kernel or filter, which will look for the presence of the feature. The technical word for this process is convolution.A section of the image is represented by the two-dimensional (2-D) weighted array that serves as the feature detector. The 3x3 matrix that typically makes up the filter, whose size might vary, also influences how large the receptive field will be. The method is then repeated until the kernel has swept through the entire image. Next, the filter is applied to a section of the image, and a dot product between the input pixels and the filter then shifts by a stride.A series of dot products from the input and filter culminate in a feature map, activation map, or convolved feature.

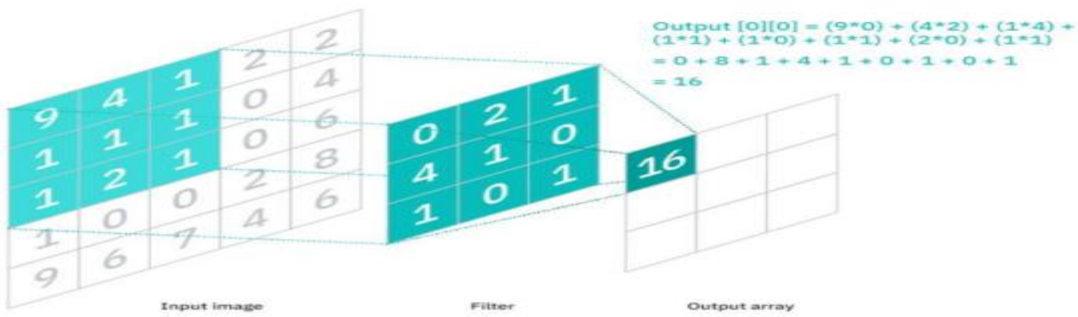


Figure 3.2. Visual representation of a convolutional layers

Each output value in the feature map does not have to relate to each pixel value in the input image. It simply must be connected to the receptive field, which is where the filter is applied. Convolutional and pooling layers are often referred to as “partially connected” layers because the output array does not have to map directly to each input value. The feature detector’s weights remain constant as it advances over the image, a technique known as parameter sharing. Through the process of backpropagation and gradient descent, some parameters, such as weight values, adjust during training. Before the neural network can be trained, three hyperparameters that determine the output volume size must be established. Number of filters, stride, and zero padding are among them. The depth of the output is influenced by the number of filters used. The kernel’s stride is the number of pixels it moves over the input matrix. Although stride values of two or more are uncommon, a larger stride produces a smaller output. When the filters do not fit the input image, zero-padding is utilized. All members outside of the input matrix are set to zero, resulting in a larger or similarly sized output.

3.4.2 Pooling layer

Downsampling, often known as pooling layers, is a dimensionality reduction technique that reduces the number of factors in an input. The pooling process, like the convolutional layer, sweeps a filter across the entire input, but this filter has no weights. Instead, the kernel populates the output array by applying an aggregation function to the values in the receptive field. Pooling is divided into two categories: Max pooling and Average pooling. Max Pooling is a convolutional procedure in which the filter or Kernel takes the pixel with the highest value from the input and sends it to the output array. This approach is used more commonly as compared to average pooling. Average pooling, on the other hand, is a pooling technique that employs the average value for patches of a feature map to build a downsampled feature map.

3.4.3 Fully connected layer

The fully connected layer's name is self-explanatory. In a neural network, fully connected layers are those where all the inputs from one layer are connected to each activation unit of the next layer. In partially connected layers, the pixel values of the input image are not directly connected to the output layer, as previously stated. Each node in the output layer is connected directly to a node in the previous layer in the fully connected layer. This layer performs categorization based on the features extracted by the preceding layers and the filters applied to them. While convolutional and pooling layers typically utilize relu functions to classify inputs, fully connected layers typically use a softmax activation function to produce a probability ranging from 0 to 1.

3.5 Network Architecture

MobileNetV2 is a convolutional neural network architecture that seeks to perform well on mobile devices. MobilenetV2 is the second version of the Mobilenet models. It significantly has a lower number of parameters in the deep neural network. This results in more lightweight deep neural networks. Being lightweight, it is best suited for embedded systems and mobile devices. This makes it even more efficient and powerful. The MobileNetV2 models are faster due to the reduced model size and complexity. MobilenetV2 is a pre-trained model for image classification and a refined version of MobilenetV1. Pre-trained models are deep neural networks that are trained using a large images dataset. As a whole, the architecture of MobileNetV2 contains the initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers..

MobileNetV2 is an architecture of bottleneck depthseparable convolution building of basic blocks with residuals . It has two types of blocks as shown in Figure.3.3.Both blocks consist of three layers.The first layer is 1x1 convolutions with “ReLU6” and a onestride residual block. The second layer contains depth-wise convolution, a residual block with stride 2 and is used for shrinking. The last layer contains a 1x1 “convolution” with no non-linearity. The methodology of object detection make 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 serve as a simple feature extractor for object detection tasks to take input images and produce feature maps for each image.

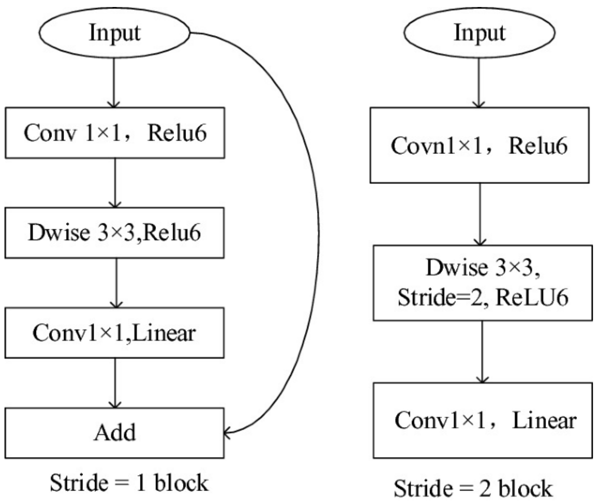


Figure 3.3: MobileNetV2 blocks

The pretrained techniques are usually used to extract feature maps with high-quality classification problems. This part of the model is called the base model. The base model which is used the “image net” weights. ImageNet is an image database.It has been trained on hundreds of thousands of images and is extremely useful for image categorization.During training,the evaluated “bounding boxes” are compared to the “ground truth boxes”. During backpropagation, the trainable parameters are changed as needed. The MobileNet design consists of two parts. They are a base model and a classifier.

CHAPTER 4

IMPLEMENTATION AND DESIGN

4.1 Overview of the Proposed System

The proposed system is a face mask detection system based on Convolutional Neural Network (MobilenetV2) method. The method detects whether a person is wearing a face mask or not using a facial detection system. Provide an image of a few people wearing a mask and not wearing a mask as input dataset and the segmented image of the same is obtained as output. Then the model is implemented by using a webcam where the video is read by frame and resized as necessary. Then, the preprocessing function is called to get the result of people wearing a mask and not wearing a mask along with the accuracy in percentage.

The proposed system comprises of two phases and they are as follows:

For training phase,

1. Data Visualization
2. Image Preprocessing
3. Data Augmentation
4. Model Training

For testing phase,

1. Get the testing images.
2. Image Segmentation using Mask R-CNN

4.2 The Proposed System Design

This system intends to implement Convolutional Neural Network Model. The aim of this model is to classify face mask wearing or not for COVID-19 precaution. The detailed designs of the proposed system are described in Figure 4.1.The proposed system uses MobilenetV2, a Convolution Neural Network. The proposed system consists of two modules: building mask detector model (training phase) and classifing face mask wearing or not (testing phase).

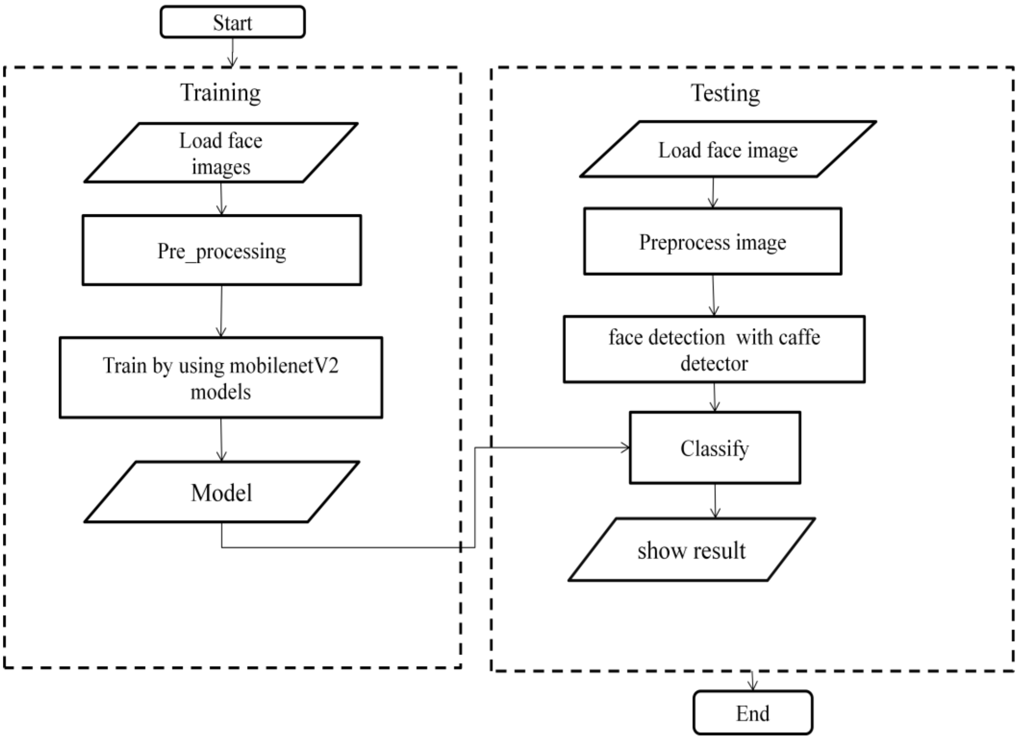


Figure 4.1: Flow Diagram of the Proposed System

The proposed system is divided into two parts.They are training part and testing part. In training part, the system uses 3832 RGB image in two folders as (with mask and without mask). Firstly, thse input image is loaded face image in the dataset. These image are difference size. So, preprocessing step is necessary. In preprocessing, there are three step.They are resizing, one-hot encoding and data augmentation. After the preprocessing, images are trained by MobilenetV2 model. Finally, the output is evaluted classifer model.

To evaluate the system performance, analysis results are emphasized on training time and system accuracy results. The performance value is accomplished by employing precision, recall and f-measure.

In testing part, the input image is loaded face image in the dataset, google and real world captured images. The input image is unstructured image sizes. Therefore, preprocessing step is necessary to resize image. Then, the input image is detected face detection with caffe detector model. Ather the face detection step, the input image is detected face mask classifier whether to know wearing face mask or not. Finally, the result is if a person is wearing a mask, the face region is displayed with a green anchor box and if a person is not wearing a mask, the face region is displayed with a red anchor box.

4.2.1 Dataset Description

The image dataset from Kaggle is applied in this system. It has two classes. The first one is the image with people who wear masks. The second one is the 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. Figure 4.2 shows images in which people wear masks. Figure 4.3 shows images in which people do not wear masks. This dataset was used to pre-train the model. Figure 4.4 shows the total number of different classes. Likewise, figures 4.5 shows a distribution of with_mask class image sizes and figure 4.6 shows a distribution of without_mask class image sizes.

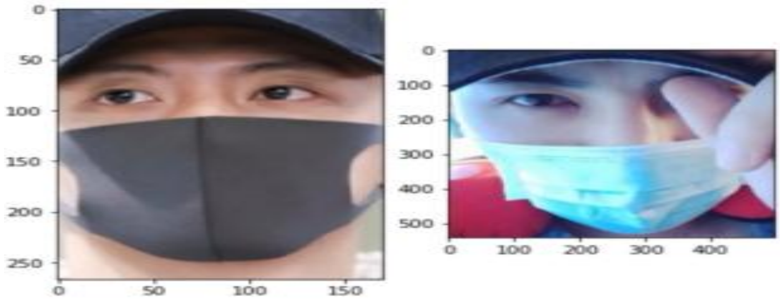


Figure 4.2: Images in which people wear masks

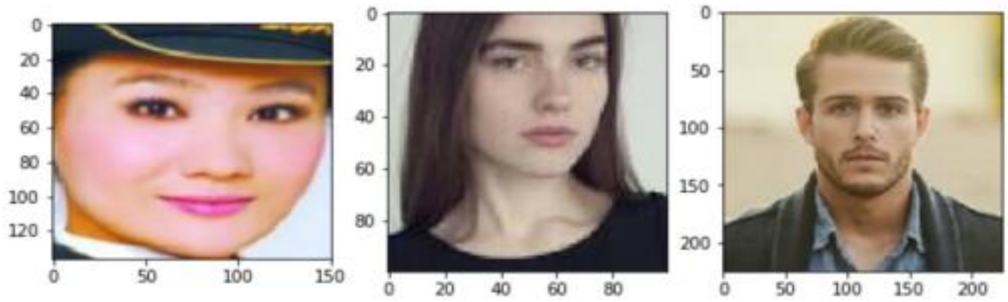


Figure 4.3: Images in which people do not wear masks

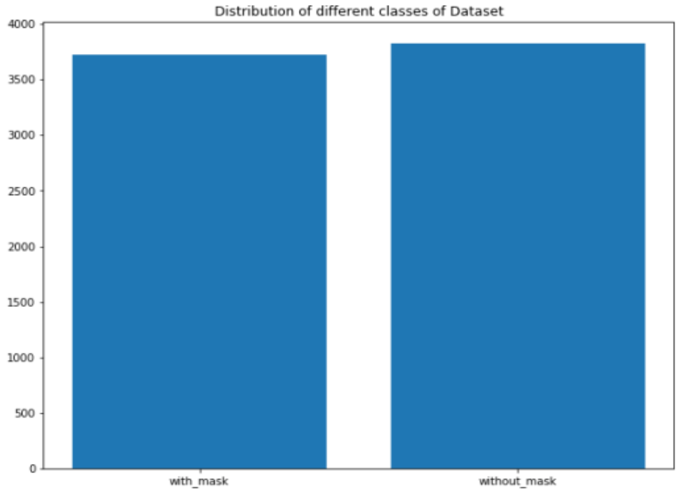


Figure 4.4: Total number of different classes of dataset

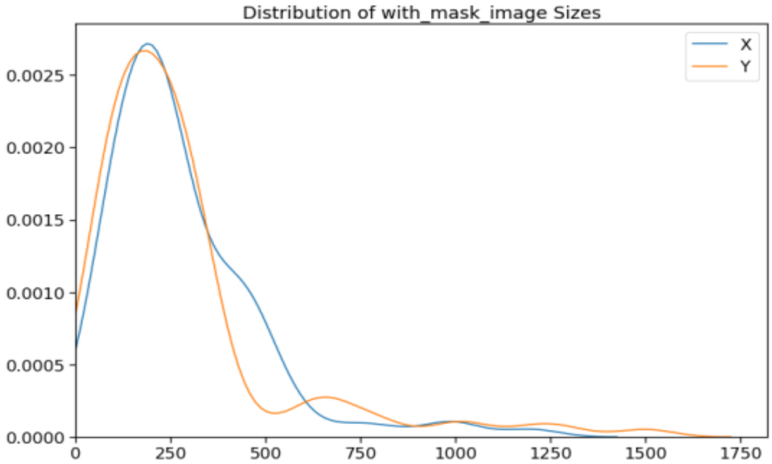


Figure 4.5: Distribution of with_mask class image sizes

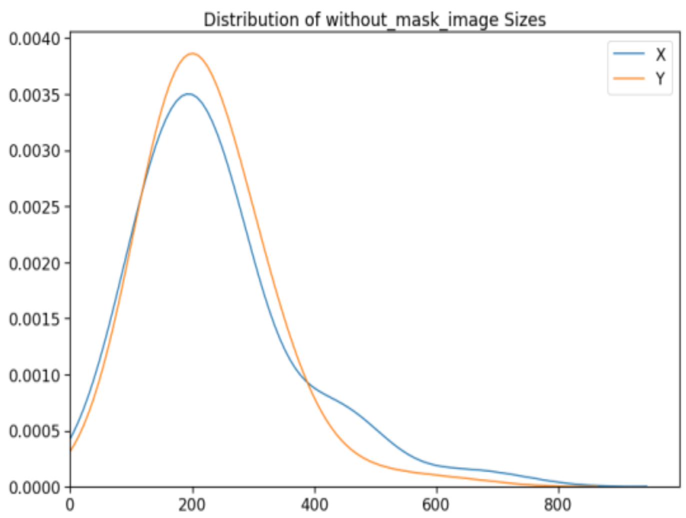


Figure 4.6: Distribution of without_mask class image sizes

4.2.2 Data Visualization

Data visualization is the process of transforming abstract data to meaningful representations using knowledge communication and insight discovery through encodings. It is helpful to study a particular pattern in the dataset. The total number of images in the dataset is virtualized in both categories_ 'with mask' and 'without mask'.

4.2.3 Image Preprocessing

Data pre-processing is a very significant step that helps in improving the quality of data to help the extraction of important understandings from the data. Data preprocessing involves conversion of data from a given format to much more user friendly, desired and meaningful format. It can be in any form like tables, images, videos, graphs, etc. These organized information fit in with an information model or composition and captures relationship between different entities. The images are resized into 224x224.Image resizing is n importand step in preprocessing. To perform One-Hot Encoding, the image is firstly converted into a NumPy array. One-Hot Encoding is used for encoding categorical data into vectors 0s and 1s.Images with masks are encoded to 0 and without masks are encoded to 1. After this process the train_test_split function can split, some of the dataset for training and some for testing.

Figure 4.7: 224x224 resize image of preprocessing

4.2.4 Data Augmentation

A face mask detection system does not take input data, converts it randomly and returns both input and transformed data. The image data generator in keras uses the input image and transforms randomly into transformed data. A collection of techniques introducing random jitters and perturbations and creating a new training sample from existing one is called data augmentation. The model's generalizability is improved by using data augmentation. In data augmentations, images are applied with a few techniques to generate a larger dataset from the given one and thus for getting a better model for the problem. The augmentation techniques are rescale, rotation_range , width_shift_range , height_shift_range, shear_range , zoom_range and horizontal_flip. Random rotation augmentation will randomly rotate the images from 0 to 360 degrees in clock wise direction. In this module, firstly the image is rotated with random rotation augmentation.The zoom augmentation method is used to zooming the image. This method randomly zooms the image either by zooming in or it adds some pixels around the image to enlarge the image. After rotation, the image is zoomed.Reversing the entire rows and columns of an image pixels in horizontally is called horizontal flip augmentation. Finally, the image is flipped horizontally.

4.3 Model Training

The training process is also known as backpropagation. The training step is very important for the model to find the weights that best accurately represent the input data to match its correct output class. Thus, these weights are constantly updated and moved towards their optimal output class. In this study, a Face Mask Detection Dataset was used to train the MobilenetV2 model. During the training process, training data was split into smaller sizes of batches of 32. Batch size includes splitting the whole dataset into a chain of the reduced amounts of data fed into the model one at a time. Splitting the training dataset into batches helps in training the model faster and controlling the gradient error accuracy. Likewise, the learning rate of 0.001 has been applied to set the size of a step in the direction of the minimizing the loss function. The optimization algorithm, forward pass, loss function, backward pass, and weights updates are followed to train the model from the labelled data.

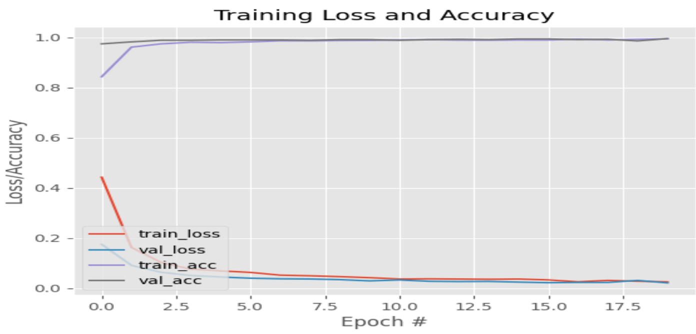
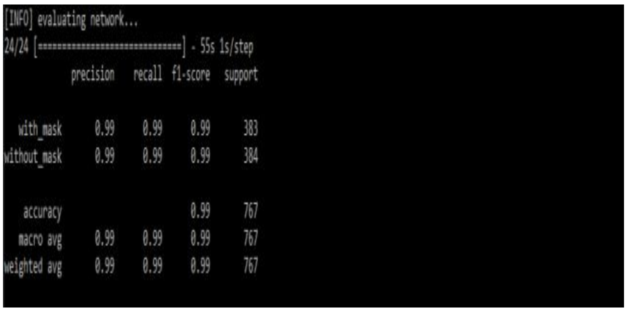


Figure 4.8 :Training Loss and Accuracy


[INFO] evaluating network...
24/24 [==============================] - 55s 1s/step

	precision	recall	f1-score	support
with_mask	0.99	0.99	0.99	383
without_mask	0.99	0.99	0.99	384
accuracy			0.99	767
macro avg	0.99	0.99	0.99	767
weighted avg	0.99	0.99	0.99	767

Figure 4.9 : Classification Report

4.3.1 Adam Optimizer

During the training process , Adam optimizer has been applied to the CNN model. Adam is an adaptive learning rate optimization algorithm that has been proposed especially for training deep neural networks. Adam optimizer works by calculating each learning rate according to different parameters within the model. Thus, using Adam optimization algorithms was very beneficial for this methodology because it uses valuations of the first and second gradient moments to acclimate the learning rate for each of the neural network weights.

4.3.2 Mean Squared loss (MSE)

In this approach, to achieve the best accuracy results, the MSE loss function was used. MSE stands for Mean Square Error, which is commonly used and is the sum of squared distances between target variable and predicted values. Equation shows how the MSE loss function is being calculated, where n is the number of data points, ytrue is the actual value for data point i and ypredicted is the value returned by the model.

4.3.3 Running Epochs

An epoch refers to each cycle that a model takes through the full training dataset. For example, feeding the neural network with training data for more than one epoch, the result should become better in terms of predicting the given unseen data, which is the test data. In this approach, 30 epochs were applied to achieve the best accuracy results of predicting the test data. Table 4.1: shows the epochs on the loss percentage.

Table 4.1 Epoch Number and Loss Percentage Results

Epoch Number	Percentage of Loss
1	0.44
5	0.06
10	0.04
15	0.03
20	0.02

4.4 Testing

In this very last step of testing, sample testing images were passed through the convolutional neural network, and the predicted and actual true classes were compared. 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.

1. I(x) ←Input Images
2. Cl← Convolutional Layer
3. S (b) ←Size of Box
4. F (m) ←Feature Map
5. D← dimension of boxes
6. F(C) ←Fully connected layer
7. I(c) ←Change in Intensity of pixel
8. L(r) ←Learning Rate, B(s) ←Batch size, E (p) ←Epochs
9. A ←Threshold value
10. Co ← Confidence
11. T← No. of truth box 18
12. B← Number of default boxes
13. E B 2t×4 Truth boxes set T
14. Class[l] ←Class labels set
15. L (M) ←Load Model
16. N← Total no. of class labels
17. Obj ←Final Object

Pseudocode for Face Mask Detection

```
1. Initialize the MobileNet V2 model
2. Read the input Frame
3. While true
4. Initialize the L(r), B(s), E (p)
5. Resize the I(x), height, and the width
```

6. Load the base Model
7. F (m) ← {Minimum Cl +Maximum I(c)}
8. End for 9. Calculate the blob
10.For each Co, lens, S (b) do
11.If lens > 0 12.Width(W) = Xmin*Xmax
13.Height(H) = Ymin*Ymax
14.else
15.Resize the box with possible dimension
16.else if
17.else for
18.Initialize the all objects
19.X = cen [x] – w ÷2
20.Y = cen[y] – h ÷2
21.Assign x, y, w, h, Co, lens values
22.for i in indexes
23. for each i the truth box having class label class [1] do
24. Class_id = max(scores)
25. if confidence > 0
26. Confidences.append(float(confidence))
27. end if
28. end for
29.end for
30.for I in class_ids
31. return label, confidence
32.end for
33.end for

4.4.1 Image Segmentation using Mask R-CNN

A CNN for image segmentation and instance segmentation which is developed on top of faster R-CNN used to locate objects and boundaries is the Mask R-CNN. The precise detection of all objects in an image and segmenting each instance of those is called Instance segmentation or Instance recognition. It comes out as the result of object detection, object localization and object classification. In this type of segmentation, a clear difference between each object classified as similar instances can be observed.

In the instance segmentation process, each person is separated as a single entity. It is also called foreground segmentation as it works on the subjects of the image rather than the background. R-CNN can get 2 outputs for each object, a class label and a bounding box offset while Mask R-CNN can get 3 outputs where there is the object mask in addition to the class label and bounding box offset. The additional mask output is different from the other two outputs where this requires extraction of a much finer spatial layout of an object. Mask R-CNN is basically faster R-CNN with addition of an output for object mask along with existing outputs like class label and bounding box.

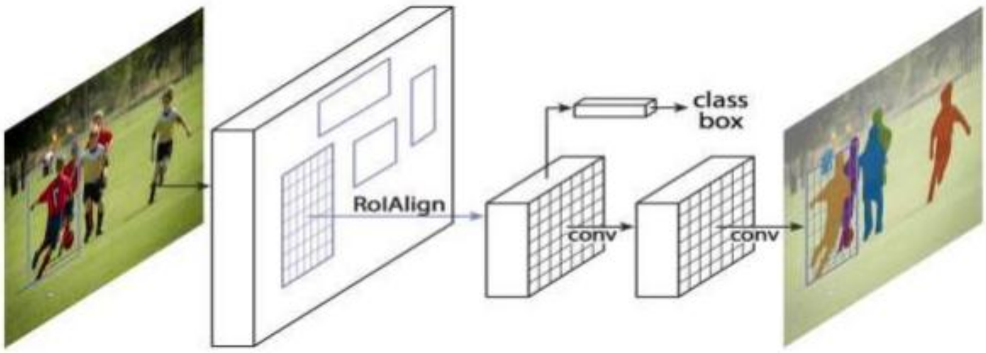


Figure 4.10 : Mask R-CNN

4.5 Evaluation of System Performance

If there is no system performance evaluation, it is impossible to know how well the system is performing. The evalution method can examine true positive, true negative ,false positive and false negative.The confusion matrix can be used to analyze the potential of a classifier. A confusion matrix generates actual values and predicted values after the classification process. The confusion matrix table as following.

Table 4.2: Confusion Matrix

Actural class	Prediction class	
	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

TRUE POSITIVE, TP In the case of mask detection in an image, a TP value also refers to a detection value, where there is a facemask detected on the model, and the facemask exists on an image. FALSE POSITIVE, FP This is the case when a false detection occurs by the model. That is, the model is detecting a mask, but there is no mask. TRUE NEGATIVE, TN A true negative value is more likely to describe the non-object regions and detect them as a non-object region. This value is barely used in the detection of masks as it does not affect the performance. FALSE NEGATIVE, FN A false negative refers to when the model misses a facemask on an image to detect.

Accuracy is one of the most widely used evaluation metrics for recognition and classification problems. It represents the ratio between the correct number of predictions and the total number of samples, which can be defined as follows:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \tag{4.1}$$

Precision is a positive prediction number that shows how correct the system is. The following is a mathematical equation to consider:

$$\text{Precision} = TP/(FP + TP) \tag{4.2}$$

Recall is also referred to as sensitivity, and it indicates how many confident instances the model properly identifies. A mathematical expression is as follows:

$$\text{Recall} = TP/(FN+TP) \tag{4.3}$$

In above formulas, True positive values refer to images which were labelled true and after prediction by model gave true result. Likewise, for True negative refers to images which were labelled true but after prediction resulted in false result. False positive refers to images which were labelled false and after prediction resulted in false hence false positives. False negative refers to images which were labelled false and after prediction resulted in true hence false negatives. The accuracy was a good measure to start with, because the classes were balanced. Precision gave the measure of positive predicted values. Recall gave the ability to a classifier to find all positive samples and f1 score gave the measure of test accuracy. These evaluation metrics were chosen because of their ability to give best results in balanced dataset.

Total images = 201 images

Table 4.3: Actural Class and Predicted Class

Actual Class	Predicted Class	
	Positive	Negative
Positive	TP=71	FN=8
Negative	FP=29	TN=93

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

4.6 Implementation

The proposed system implemented a model trained 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.

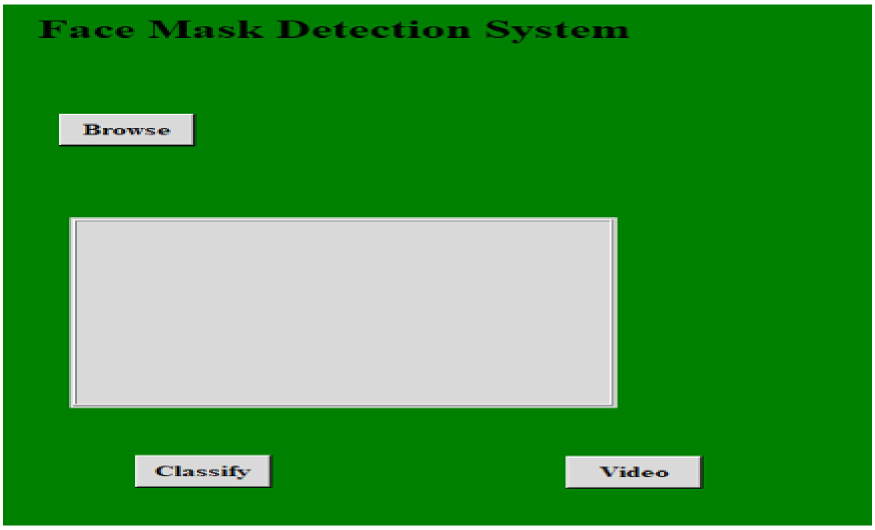


Figure 4.11: Home page of the porposed system

This is main page of the proposed system. All images in the testdataset is JPG and PNG formats. Dimension of the images is between 600 width and 500 height. These image take various sources such as testing image from dataset,google and real world captured images. The user have to browse an image that need to be classified as 'wearing mask or not' as shown in figure 4.12.

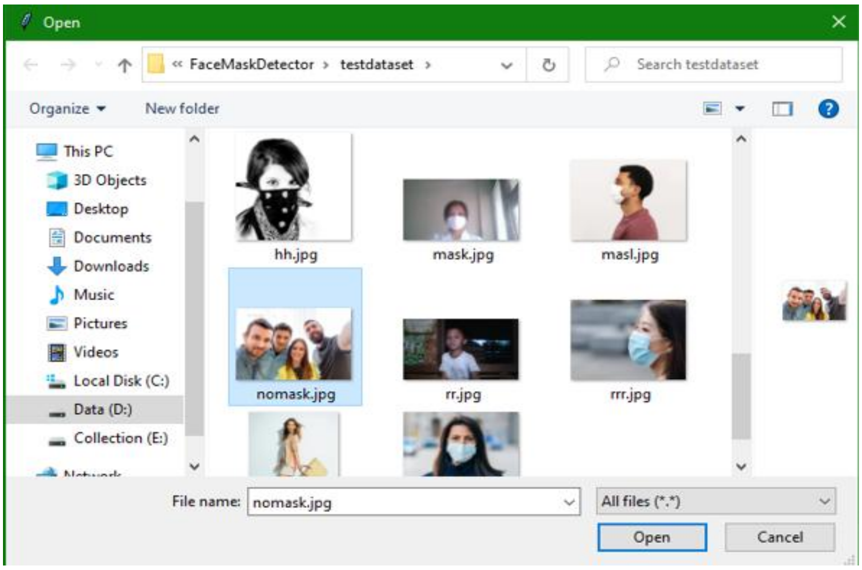


Figure 4.12: Browsing an image

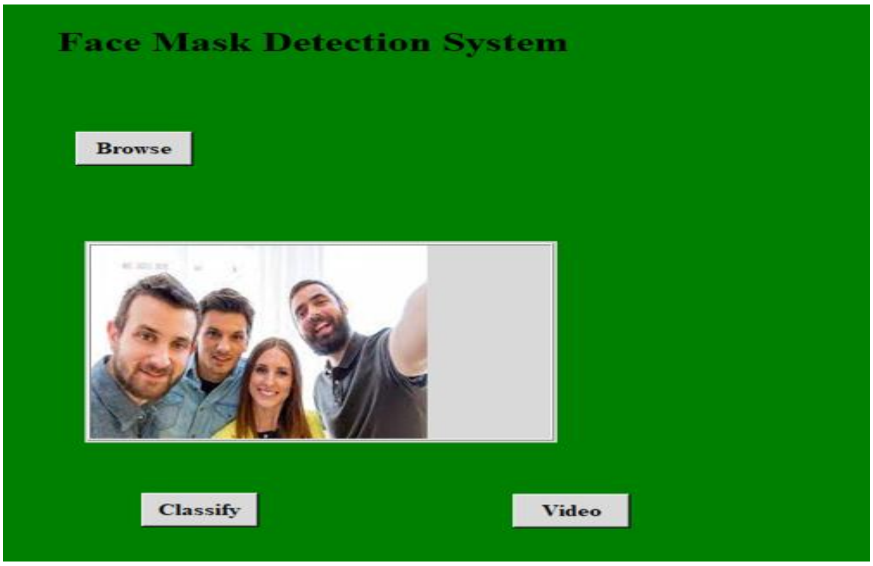


Figure: 4.13 Image without mask loaded into the system

After the input image is put into the system, it need to be preprocessed and classified with the trained model as shown in figure 4.13.

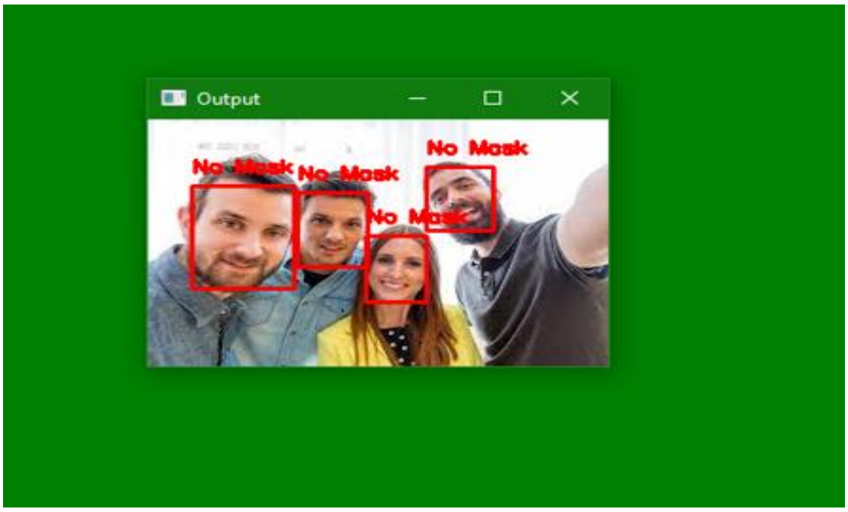


Figure 4.14: Result for detection of without mask

If the users clicks the 'classify' button, the result is shown over the image with anchor box and message (mask or no mask). The four people in an input image do not wear masks so the proposed system classify as 'no mask' and shows with red anchor box as shown in figure 4.14.

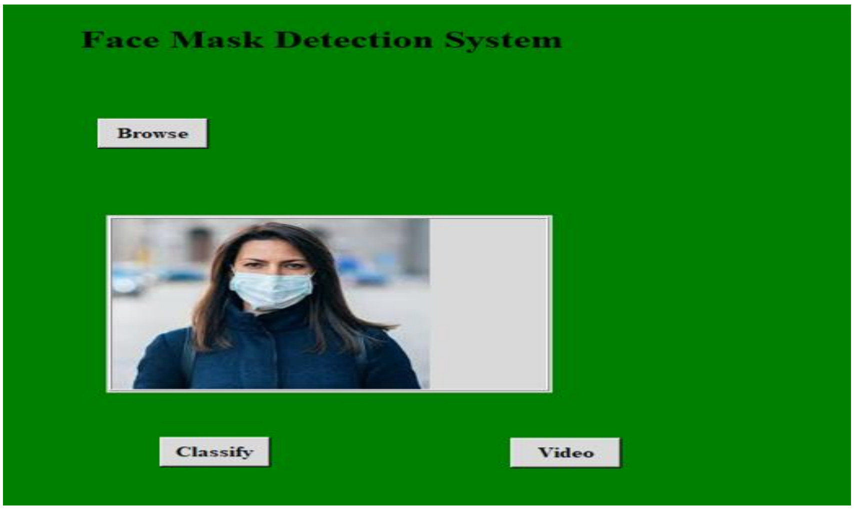


Figure 4.15: Image with mask loaded into the system

If the users clicks the 'classify' button, the result is shown over the image with green anchor box and message (mask). The person in an input image wears mask so the proposed system classify as 'mask' as shown in figure 4.16.

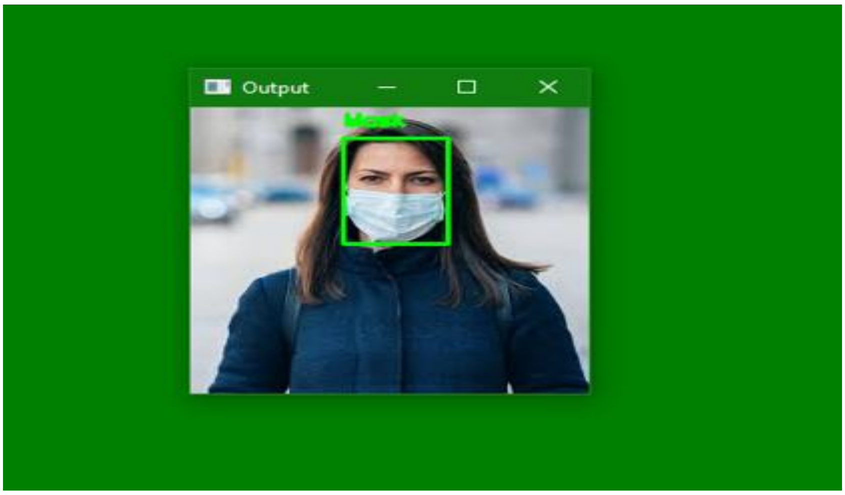


Figure 4.16: Result for detection of with mask

Figure 4.17: Image without mask loaded into the system

After the input image is put into the system as shown in figure 4.17, it need to be preprocessed and classified with the trained model. If the users clicks the 'classify' button, the result is shown over the image with anchor box and message (mask or no mask). The person in an input image do not wear mask. So the proposed system classifies as 'no mask' and shows with red anchor box as shown in figure 4.17. The image shows the girl' whole body and has only small face region but the system detects correctly that she does not wear mask as shown in figure 4.18.

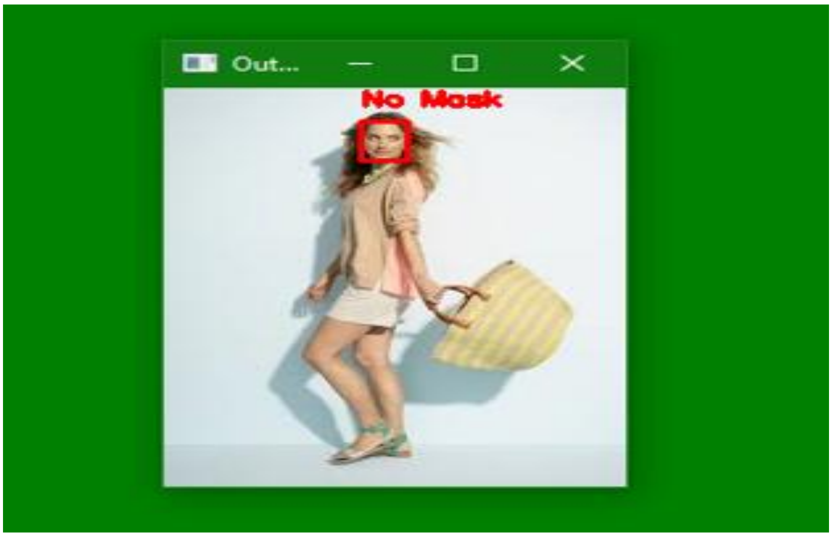


Figure 4.18: Result for detection of without mask

The proposed system can detect a person is wearing mask or not from a live video stream. If the input is video stream, the system detects the face region of a moving person in a video and detects that person is wearing mask or not as the continuous detection. If the user clicks the 'video' button, the input live video stream for 0.2 second is browsed into the system. The system can detect a person in the video is wearing mask or not. If a person is wearing mask during the frames of video stream, the system detects the face region and shows the result that he/she is wearing mask. If he/she takes off the mask in that video stream, the system detects the face region and shows the result that he/she is not wearing mask immediately. The system can detects the changes immediately and continuously as 'wearing mask or not' during the entire live video stream as shown in figure 4.19 (a) and (b).

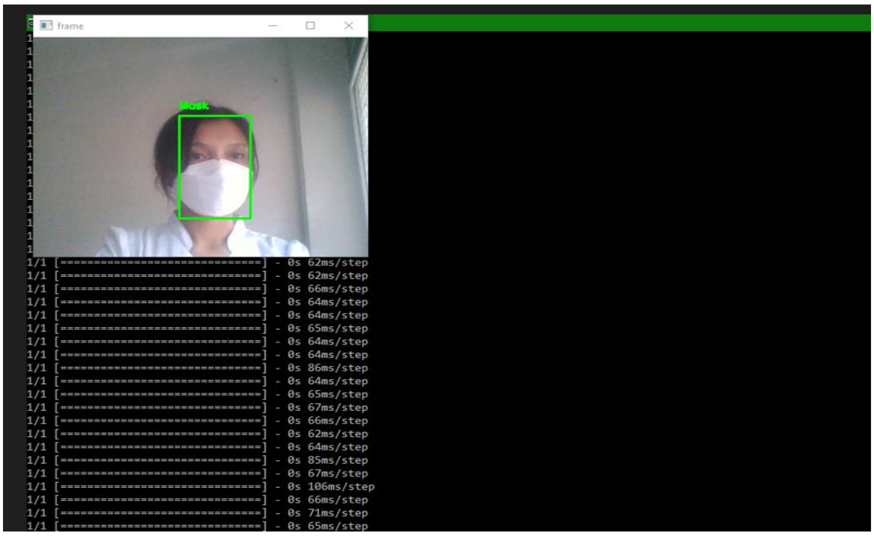


Figure 4.19 (a) Live video stream

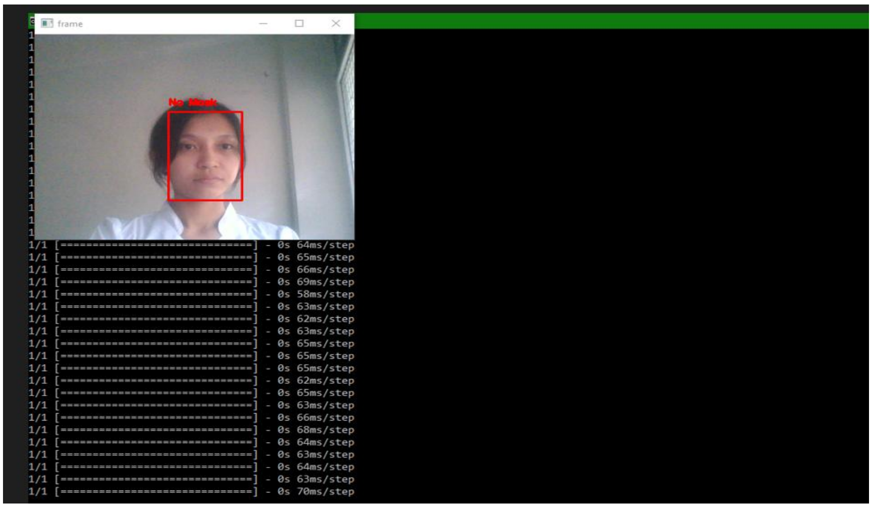


Figure 4.19 (b) Live video screen

CHAPTER 5
CONCLUSION

This thesis intends to detect whether a person is wearing a face mask or not. The face mask is an important shield for protection against Covid-19. The various components of this system are investigated and their contributions to the overall performance of the system are analyzed. In this chapter, the main contents of thesis are concluded. The limitation of the system and future work are suggested in this chapter.

The spread of Covid-19 is increasing every day in every corner of the world. This needs to be controlled to get back to our normal lives. While the specialists take care of the vaccine part, can help them by following the guidelines provided by WHO to remove/control the spread of this virus. COVID-19, the prevailing virus outbreak, has made us recognize the benefits of face masks. The use of face masks is essential when using public transportation. 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 used for training model and 201 images are used for testing.The accuracy is 82 % for testing images in dataset.This system can be applied in many crowded areas like metro stations, markets, schools, railway stations and so on to monitor the crowd and to ensure that everyone is wearing mask.

5.1 Advantages and Limitation of the System

Different types, color and shapes of masks are accetped and can be tested correctly with the proposed system.This system has provided good accuracy for training and testing. Input images and video stream from various sources such as real time live stream, real time captured images, google, testing images from dataset are correctly classified with the proposed system. As a drawback, the system cannot recognize precisely all face regions if the input video stream is live stream.

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 can be applied.

AUTHOR'S PUBLICATIONS

[1] Ei Cherry Lwin, Myat Mon Kyaw "Face Mask Detection By Using Convilutional Neural Network" The National Journal of Parallel and Soft Computing, (NJPSC 2022), 2022.

REFERENCES

[1] WHO EMRO | About COVID-19 | COVID-19 | Health topics. [Online]. Available: http://www.emro.who.int/healthtopics/corona-virus/about-covid-19.html, accessed on: Jul. 26, 2020.

[2] World Health Organization et al. Coronavirus disease 2019 (covid-19): situation report, 96. 2020. - Google Search. (n.d.). https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200816-covid-19-sitrep-209.pdf?sfvrsn=5dde1ca2_2.

[3] W.H.O., "Coronavirus disease 2019 (covid-19): situation report, 205". 2020

[4] Ejaz Md. S., Islam Md. R., Sifatullah M., Sarker, "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://www.researchgate.net/publication/338071004_Implementation_of_Principal_Component_Analysis_on_Masked_and_Non-masked_Face_Recognition.

[5] Ejaz, M.S., Islam, M.R., Sifatullah, M., Sarker, A.: Implementation of principal component analysis on masked and non-masked face recognition. In: 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), pp. 1–5 (2019). IEEE

[6] Lin, K., Zhao, H., Lv, J., Zhan, J., Liu, X., Chen, R., Li, C., Huang, Z.: Face detection and segmentation with generalized intersection over union based on mask r-cnn. In: International Conference on Brain Inspired Cognitive Systems, pp. 106–116 (2019). Springer

[7] Shimin Ge1 , Jia Li2 , Qiting Ye1,3 , Zhao Luo, "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.

[8] Mohamed Loey, Mohamed Hamed N. Taha, Gunasekaran Manogaran, and Nour Eldeen M. Khalifa, "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/ .

[9] Sunil Singh, Umang Ahuja, Munish Kumar, Krishna Kumar and Monika Sachdeva., Face Mask Detection using YOLOv3 and faster R-CNN, International Journal Of Engineering and Technology, 12, pp.81-104, 2021.

[10] Xinbei Jiang, Tianhan Gao, Zichen Zhu and Yukang Zhao., Real Time Face Mask Detection Method Based on YOLOv3, Electronics, 7, pp.130-147, 2021.

[11] M. Burugupalli, "IMAGE CLASSIFICATION USING TRANSFER LEARNING AND CONVOLUTIONNEURALNETWORKS,"July2020.[Online].Available:https://library.ndsu.edu/ir/bitstream/handle/10365/31517/Image%20Classification%20Using%20Transfer%20Learning%20and%20Convolution%20Neural%20Networks.pdf?seque nce=1.

[12] M. Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR), 2018

[13] 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.

[14] O.Gurav,"FaceMaskDetectionDataset"[Online].Available:https://www.kaggle.com/datasets/omkargurav /face-mask-dataset.

[15] Xinbei Jiang, Tianhan Gao, Zichen Zhu and Yukang Zhao., Real Time Face Mask Detection Method Based on YOLOv3, Electronics, 7, pp.130-147, 2021.

[16] Sunil Singh, Umang Ahuja, Munish Kumar, Krishna Kumar and Monika Sachdeva., Face Mask Detection using YOLOv3 and faster R-CNN, International Journal Of Engineering and Technology, 12, pp.81-104, 2021.

[17] Santanu Pattanayak, Pro Deep Learning with TensorFlow, https://link.springer.com/book/1.1007 /978-1-4842-3096-1.

[18] Li Deng, Dong Yu, Deep Learning: Methods and Applications, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DeepLearningBook_RefsByLastFirstNames.pdf.

[19] Sanjibani Sudha Pattanayak, 1Sateesh Kumar Pradhan and 2Ramesh Chandra Malik, Performance Evaluation of Deep Learning Networks on Printed Odia Characters, Journal of Computer Science,2020.

[20] Aurélien Géron,Hands-On Machine Learning with Scikit-Learn and TensorFlow, https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.2017.

[21] Sha, Runsheng, Pneumonia Classification Model by Convolutional Neural Network, DEGREE PROGRAMME IN ARTIFICIAL INTELLIGENCE 2022,pp.16-20,2022