Skin Cancer Diagnosis using Support Vector Machine based on Gray Level Co-occurrence Matrix

MYINT MYINT WAI

M.C.Sc DECEMBER 2022

Skin Cancer Diagnosis using Support Vector Machine based on Gray Level Co-occurrence Matrix

BY

MYINT MYINT WAI

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

Master of Computer Science
(M.C.Sc)

University of Computer Studies, Yangon
DECEMBER 2022

ACKNOWLEDGEMENTS

To start with, I would like to acknowledge particular thanks to the Union Minister, the Ministry of Science and Technology for the full facilities and supports during the M.C.Sc. course at the University of Computer Studies, Yangon.

First and foremost, I would like to express my deepest gratitude and my thanks to **Dr. Mie Mie Khin,** Rector, University of Computer Studies, Yangon, for her kind permission to submit this thesis.

Deep appreciation is extended to course coordinators, **Dr. Si Si Mar Win and Dr. Tin Zar Thaw,** Professors, Faculty of Computer Sciences, 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. Win Lelt Lelt Phyu,** Professor, Faculty of Computer Sciences, 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 Hnin Yee Aung**, Lecturer, Department of English, University of Computer Studies, Yangon, for her editing this thesis from the language point of view.

I am also so grateful to **all my teachers** who taught me computer science throughout the master's degree course, and **my friends** for their corporation, and extremely grateful to **my beloved parents** for their continuous support, patience, understanding, consideration and love that enable me to complete this thesis.

Finally, thanks to all who helped with necessary assistance for the completion of this thesis.

ABSTRACT

As today's world is more developed than ever, the way to diagnose the diseases is better and more precise. Cancer has been identified as one of the leading causes of death. In this study, features from dermoscopy pictures were extracted using the Local Binary Pattern (LBP) and Gray Level Co-Occurrence Matrix (GLCM) methods. Melanoma, the most dangerous type of skin cancer, or benign skin tumors were then classified using Support Vector Machine (non-cancerous). Five features, contrast, dissimilarity, homogeneity, energy and correlation were extracted by GLCM. Radial Basis Function (RBF) Kernel of SVM was trained with the features and, then tested the images and classified whether the cancer or not. Performance was evaluated with the confusion matrix by testing accuracy, specificity, sensitivity and precision. Using image processing software, this study proposes a technique for locating skin cancer. In order to identify whether skin cancer is present, the system first gets an image of a skin lesion as input and analyzes it using image processing methods. The contrast, homogeneity, dissimilarity, energy, and correlation analysis are all checked by the Lesion Image analysis tools during the image segmentation and feature stages. The image is categorized as either a benign or malignant cancer lesion using the derived feature parameters.

Keywords—Texture feature extraction, Gray level co-occurrence matrix (GLCM), Support vector machine (SVM), Texture features.

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

The most terrifying malignancies are those that affect the skin. It is important to advise skin cancer patients to receive treatment as soon as feasible. The greatest part of the entire body, the body, can become terribly sick if untreated if it spreads to the complete body. If the disease is not known and left untreated, it can easily spread to the whole body, and it can be very difficult to treat, and it can affect the life expectancy. Although it has been treated, it can be found that the hard leaves remain on the body.

This approach may be used to diagnose the ailment accurately and without difficulty, and it will give you the right answer. Using this system, if a diagnosis is found, doctors can be treated quickly. The body's other tissues could also suffer injury. By accurately diagnosing the disease and assisting clinical decision-making, early classification of skin lesions may improve the likelihood of curing cancer before it spreads. Skin Cancer diagnosis is a technique used in hospital and cancer treatment centers. The purpose system describes the methods used to ascertain the presence or absence of skin cancer.

The biggest advantage of Skin Cancer Detection Using SVM is that, thanks to its improved magnification, it can avoid the wasteful excision of perfectly healthy moles and skin lesions. There have been several attempts to introduce traditional medicine around the world, particularly in less technologically developed nations, but these efforts have run into difficulties due to the high expense of medical supplies and equipment as well as a lack of medical knowledge.

Environmental factors and genes are usually to cause for skin cancer. Most populations around the world still lack access to the requisite techniques for the early diagnosis of these diseases. The suggested paper here offers a method to identify distinct forms of these disorders. The system returns a negative result if there is no cancer present. A dataset that could be utilized to test and train the model was made available by the International Skin Imaging Collaboration (ISIC).

1.1 Objectives of Thesis

The objectives of our thesis are as follows:

- To classify skin cancer or not by using skin images
- To propose robust and effective skin cancer classification system
- To support medical workers when decision making concern with skin cancer
- To make easily skin cancer diagnosis and to take early treatment for the patient
- To learn support vector machine, GLCM and image preprocessing processes
- To learn support vector machine, LBP, GLCM and image preprocessing processes

1.2 Motivation of Thesis

Skin is the largest area of the human body and skin cancer is one of the most dreaded cancers now. Nearly all skin cancers can be cured if found and treated early. If the disease is not known and left untreated, it can easily spread to the whole body. It can be very difficult to treat, and it can affect the life expectancy. Although it has been treated, it can be found that the hard leaves remain on the body unfortunately. Therefore, early detection and prompt treatment of skin cancer are crucial. By accurately diagnosing the disease and assisting clinical decision-making, early classification of skin lesions may improve the likelihood of curing cancer before it spreads.

Software systems can predict outcomes more correctly with the use of machine learning (ML), a type of artificial intelligence (AI), without needing to be explicitly told to do so. Machine learning algorithms use historical data as input to forecast new output values.

These factors serve as the driving force for this study, which examines how well Support Vector Machine, a type of machine learning, performs in classifying skin lesions from dermoscopic pictures.

1.3 Organization of the Thesis

This thesis is organized in five chapters. They are as follows:

Chapter 1 presents introduction of the proposed system, objectives of thesis, motivation

Chapter 2 presents the Machine Learning and Image Processing and

Chapter 3 describes the Support Vector Machine

Chapter 4 expresses the design and implementation of the proposed system and finally

Chapter 5 presents the conclusions of this thesis, Advantages and Limitations of the System

CHAPTER 2

THEORETICAL BACKGROUND

Image processing is a method for spotting patterns and other elements in images. A few of the most common applications for image processing include face detection, PET scanning, X-ray imaging, medical CT, UV imaging, cancer cell image processing, and many others. The addition of image processing to the field of medical technology has improved the diagnostic system. Image enhancement, classification, and correction and restoration are all kinds of image processing. Many various applications, such as handwriting analysis, image identification, computer-assisted medical diagnosis, and more, use pattern recognition.

2.1 Image Processing

Image processing is a technique for applying various procedures to an image in order to improve it or extract some relevant information from it. It is a kind of signal processing where the input is an image and the output can either be another image or features or characteristics related to that image. Imaging technology is one of those that is developing swiftly.

Image processing will include the following three steps:

- Uploading the image using image-acquisition software;
- Examining and modifying the image;
- A report or image analysis-based output whose results can be adjusted.

The processing of digital images through the use of a digital computer is referred to as digital image processing. Image processing is done to make an image better or to retrieve important information from it. Image processing is the process of analyzing images using computer algorithms. In this system, it will be found that skin cancer is classified by machine learning methods after using image processing.

The medical photographs that were obtained from a trustworthy and reputable source. The image pre-processing and segmentation processes are carried out using the installed PYTHON program. An algorithm for segmentation thresholding is utilized. Feature extracted are used in Local Binary Pattern and Gray Level Co-occurrence Matrix. SVM classifier computes the feature values of test and database images and classifies in accordance with those results. The dataset, which includes photos from the two categories of skin diseases, was trained with 1440 images and tested with 360.

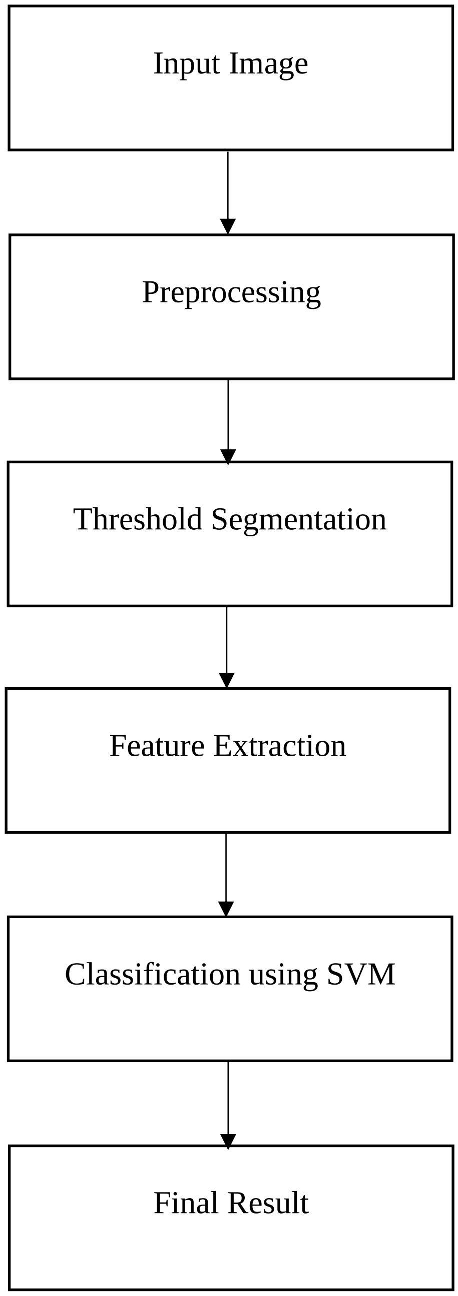


Figure 2.1 Propose System Block Diagram

2.2 Image Preprocessing

Before being used for model training and inference, pictures must first undergo image preprocessing. This includes, but is not limited to, resizing, orienting, and color adjustments. The pace of model inference and model training may both be accelerated by image preprocessing. Reducing the size of input images will greatly speed up model training time without significantly affecting model performance if the input images are extremely huge.

Before image data is ready to be used in a computer vision model, preprocessing is a necessary step to clean it up. Preprocessing is necessary for both technical and performance reasons.

The model might not function as expected if all of the images are not the same size. If the image is not the same size as the model that are constructing in code using a library like Tensorflow, then will probably run into a problem.

The propose system, preprocessing is used in three steps. In the pre-processing stage; it takes three steps converts the image from rgb to grayscale; image enhancement; to remove unwanted noise. After that, image segmentation from image to wound and skin is done using the Thresholding approach. Image processing is done in three parts: image preprocessing, image enhancement and feature extraction. The direct use of a classification algorithm on dermatology-related image data is not recommended. Skin images could have a variety of unwanted noise. Each image in the dataset must have its noise level decreased.

2.2.1 RGB to Grayscale

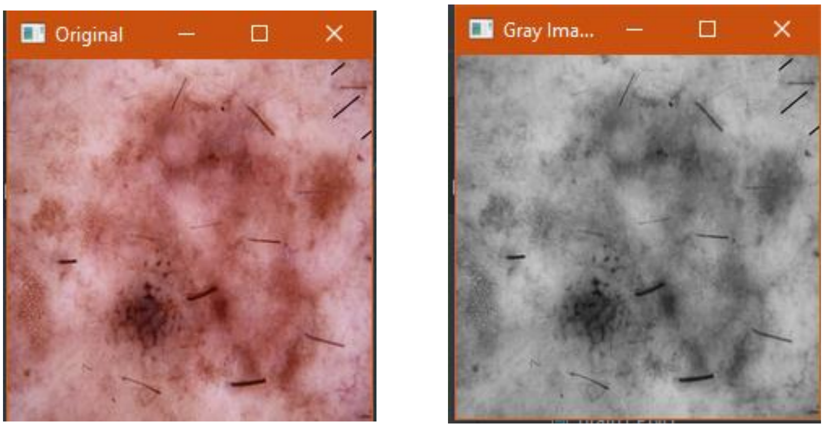
The brightness of the 8-bit image ranges from 0 to 255, with 0 being black and 255 denoting white. Pure red is encoded as (255,0,0), pure green as (0,255,0), and pure blue as (0,255,0). (0,0,255). The first number in every RGB encoding represents the amount of red, the second value the amount of green, and the final value the quantity of blue. The three numbers' ranges are 0 to 255. Black, white, and all the shades of gray in between are how grayscale images are represented. Any gray value can be represented as an RGB value by a group of three equal numbers. Black is (0,0,0), white is (255,255,255), and medium gray is (127,127,127). The higher the numbers, the lighter the gray. The purpose of converting color images to grayscale is due to the use of grayscale in the GLCM algorithm.

$$\text{Grayscale intensity} = 0.299r + 0.587g + 0.114b. \tag{2.1}$$

A shade of dark purple has an RGB value of (100, 0, 150). The weighted average is

Grayscale intensity = 0.299(100) + 0.587(0) + 0.114(150),

Grayscale intensity = 47.



(a) Original Image **(b) RGB to Grayscale Image**

Figure 2.2 Original RGB Image and Grayscale Image

2.2.2 Image Enhancement

The quality of an image will be improved using image enhancement techniques. In image enhancement, there are many different approaches for enhancing image quality. The more popular methods are contrast stretch, density slicing, edge enhancement, and spatial filtering. After the image has been adjusted for geometric and radiometric errors, image enhancement is tried. Different image enhancing techniques are used on each band of a multispectral image. Image Enhancement Techniques is one of the most important stages in the identification and interpretation of medical pictures.[5] The benefits of frequency domain image improvement include the use of better domain qualities, manipulation of a picture's frequency coefficient, and minimal complexity of computations. Enhancement and modification of digital images is done using filtering techniques. Image filters are used for edge recognition, noise reduction, and blurring. Image filters are mostly used to reduce high and low frequencies (using smoothing techniques) (image enhancement, edge detection). A 2-D convolutional operator is used in this filter. Images used to be blurry. It also eliminates noises and details. This system use gaussian kernel.

Images are blurred and noise and detail are removed using gaussian filtering (g). The Gaussian function in two dimensions is:

$$G(x) = \frac{1}{\sqrt{2\pi\theta^2}} e^{\frac{-(x^2+y^2)}{2\theta^2}} \tag{2.2}$$

Where θ is the standard deviation of the distribution. It is a given that the distribution has a mean of 0. The 2D distribution is utilized by the gaussian filter as a point-spread function. Using the image, one can achieve the 2D gaussian distribution function. Convolution is used to generate Gaussian smoothing, which uses the "point-spread" function for this 2-D distribution. The image is stored as a set of distinct pixels. Similar to how the mean filter blurs an image, Gaussian smoothing has the same effect. The Gaussian's standard deviation determines the degree of smoothing.

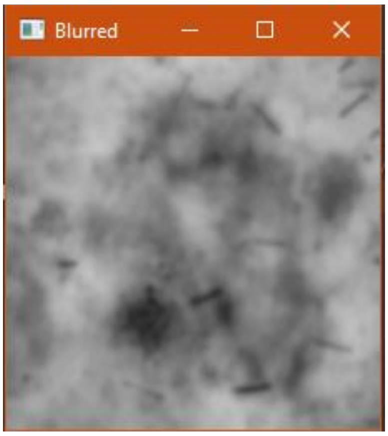


Figure 2.3 Image Enhancement using Gaussian Filter

2.2.3 Noise Removal

The image pre-processing of removing unwanted noise from image is called noise removal. Noise causes differences in pixel values. Image processing of medical images and the research of anatomical structure had made the use of noise removal techniques a necessary practice. The Salt and Pepper, Speckle, Gaussian, and Poisson noises are the ones that impact medical images the most frequently. The medical images took for comparison include images, in gray scale and RGB. The median filter performs better for removing salt-and-pepper noise and Poisson Noise for images in gray scale, and Weiner filter performs better for removing Speckle and Gaussian Noise for the blurred noise as suggested in the experimental results.

A median filter is used to remove unwanted noise. Median filter is the most commonly used filter. A nonlinear technique for removing noise from images is median filtering. The kernel can have a dimension of n x n and be designed to convolve or glide over an image that is m × m in size. The value of a specific pixel is replaced with the median value of the n x n kernel after this procedure obtains the median value of the n x n kernel on the image. The median filtering algorithm effectively reduces noise, however it has an undesirable level of temporal complexity.

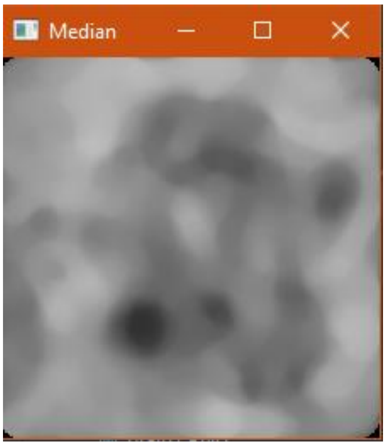


Figure 2.4 Noise removal using Median Filter

2.3 Image Segmentation

Image segmentation is the division of a digital image into several image segments, often referred to as image regions or image objects, in the context of digital image processing and computer vision. A digital image is divided up into different parts during the image segmentation process. Simplifying, altering, or separating a piece of an image into something more relevant and straightforward to analyze is the aim of segmentation. In image segmentation, researchers separate a picture into different sections that have common characteristics. It is the first step for image analysis. It would be practically difficult for someone to create computer vision without doing image segmentation. Segmentation is the removal of a required component area from a given image. Without performing picture segmentation, creating computer vision would actually be challenging. There are various image segmentation methods for segmenting images. To separate and group a certain collection of pixels from the image, a variety of image segmentation methods are utilized. Image segmentation yields either a collection of segments that collectively cover the full image or a collection of contours that are drawn directly from the image. With regard to a characteristic or computed feature like color, intensity, or texture, every pixel in a region is comparable.

For image segmentation, the proposed system threshold approach is employed. The simplest method of segmenting images is thresholding. There are three methods of thresholding.

Simple Thresholding

Adaptive Thresholding

Otsu's Binarization

Simple thresholding uses a global threshold value, meaning that it applies to every pixel in the image. With adaptive thresholding, the threshold value is determined for smaller regions, resulting in various threshold values for various locations. Similar to global thresholding, adaptive thresholding uses the differences in pixel intensities of each region to distinguish between desirable foreground image items and the background. In its most basic form, adaptive a binary image is produced via thresholding. that represents the segmentation from an input grayscale or color image. It is necessary to determine a threshold for each pixel in the image. The foreground value is assumed if the pixel value is over the threshold; otherwise, the background value is set. In situations when there are non-uniform lighting circumstances, adaptive thresholding is a highly helpful tool to have. However, it is more computationally expensive than Otsu's approach or basic thresholding.

The process for converting a grayscale image into a binary image is dependent on a threshold value. Separating images into two parts, the background and foreground, is considered a good idea. The original image should be split into two segments, one of which should have a threshold value of less than or equal to the other, while the other should have a background. A portion of the original image should be divided into two segments, one of which should have less than or equal to the threshold, and the other of which should have a background. The most often used segmentation techniques are thresholding approaches, and this system uses straightforward thresholding.

Using a straightforward threshold criterion, decision-making groups can assess options efficiently without depending on any kind of individual comparison of alternatives. Binary Thresholding is the fundamental Thresholding method. The same threshold value is used for each pixel. The pixel value is set to 0 if it is less than the threshold; otherwise, it is set to the maximum value.

Researchers have used image segmentation in the medical field to navigate during surgery, quantify tissue volumes, locate and identify cancer cells, and simulate surgeries virtually. The medical industry can use image segmentation for a variety of purposes. It assists in locating afflicted areas and developing treatment plans for them.

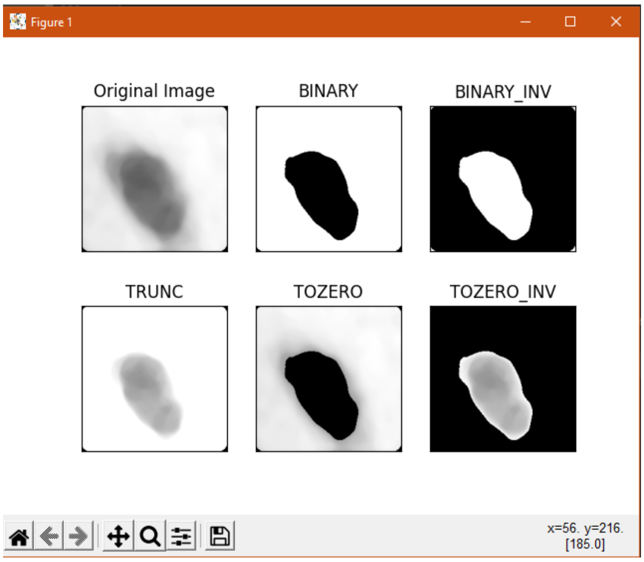


Figure 2.5 Image Segmentation using Simple Thresholding

2.4 Machine Learning (ML)

Supervised learning and unsupervised learning are the two categories of categorization techniques. Different applications, including business intelligence, Facebook's recommendation engines, human resource information systems (HRIS), and self-driving automobiles, are using machine learning. It can be categorized into three types:

(a) Supervised learning

(b) Unsupervised learning and

(c) Reinforcement learning.

2.4.1 Supervised Learning

According to Gartner, a business consulting firm, in 2022, enterprise IT professionals predict that supervised learning will remain the most widely used machine learning technique. In order to enable the system to alter the model and produce outputs that are as close to the intended outcome as in practical. In this type, feeds historical input and output data into the algorithms with processing applied in between each input/output pair. Neural networks, decision trees, linear regression, and support vector machines are examples of common algorithms used in supervised learning.

Because users input the algorithm information to aid in learning while it is being "supervised," this type of machine learning is known as "supervised" learning. The remaining details you provide are used as input features, and the output users provide to the system is identified as data. Sales forecasting, inventory optimization, and fraud detection are just a few of the business uses for supervised learning that are successful. Several instances of use cases include:

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, they are mostly used in solving classification problems. In machine learning, it is mostly employed to solve classification issues.

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary.

SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme cases, are the basis for the SVM algorithm. Consider the diagram below, where a decision boundary or hyperplane is used to categorize two distinct categories:

- Spam Filters
- Demand Forecasting
- Price Prediction
- Image Recognition

2.4.2 Unsupervised Learning

Unlike supervised learning, which relies on people to help the machine learn, unsupervised learning doesn't use the same labeled training sets and data. The computer instead searches the data for less obvious patterns. When it comes to pattern recognition and using data to inform decisions, this type of machine learning is particularly useful. Hidden Markov models, k-means, hierarchical clustering, and Gaussian mixture models are common unsupervised learning algorithms.

Consider the situation where one were uncertain as to which clients had defaulted on loan payments, using the example from supervised learning. Instead, the machine would evaluate the data after receiving borrower information to find trends among the borrowers before grouping people into other categories.

This type of machine learning is often used to create predictive models. Furthermore, frequent uses include clustering and association, which reveal the relationships between the groups. Segmentation creates a model that classifies objects together based on particular attributes. Several instances of use cases include:

- Amazon e-commerce websites
- Recommender System
- Anomaly detection
- Customer segmentation
- Preparing data for supervised learning

2.4.3 Reinforcement learning

Reinforcement learning is the machine learning technique that most closely matches human learning. The algorithm or agent being used learns by interacting with its surroundings and getting rewards, whether positive or negative. Common algorithms include deep adversarial networks, Q-learning, and temporal difference.

Users may review customer data using a reinforcement learning system, keeping in mind the bank loan client example. If the system classifies users as high-risk and they default, the algorithm benefits. If users don't default, the algorithm gives them a negative reward. Both instances, in the end, help machine learning by increasing its awareness of the problem and its surroundings.

Because reinforcement learning takes more computational capacity than most enterprises have, Gartner observes that the majority of ML platforms lack these features [2]. When a condition can be completely replicated, is stationary, or has a lot of relevant data, reinforcement learning can be applied. Working with unlabeled data sets is thought to make this sort of machine learning easier to employ because it requires less management than supervised learning. There are still applications for this kind of machine learning. Examples of certain uses are as follows:

- application in self-driving cars (AWS Deep Racer)
- industry automation (Deep mind cool Google Data Centers)
- educating robots to learn rules by providing it with raw video images that computers can use to copy the behaviors that observe

2.5 Related Work

With the advent of Deep Learning and Machine Learning, various methods have been proposed to easy skin cancer classification task in literature. The research works related to this research in literature are discussed in this section.

In [1] One of the worst malignancies is skin cancer. In this study, nine different forms of skin cancer have been classified. Additionally, deep convolutional neural networks' (CNN) effectiveness and capacity are seen.

The most common disease in the world has historically been skin cancer. This dataset, which comes from kaggle.com, has 25,780 images of healthy and cancerous tissue. Each image was assigned a category based on the ISIC description. All files are reduced to 224 x 224 pixels because scaling is one of the primary steps in data preprocessing. Figure 2 displays a few of the benchmark dataset's photos. The train, validation, and test portions of the dataset have been partitioned in an acceptable ratio. Given the imbalance in our dataset, F1 score might be a more realistic statistic.

In [2] Skin cancer is one of the most serious health problems in the world due to its high prevalence relative to other cancer forms. Melanoma has historically been a rare kind of cancer, but in the last 50 years, cases have dramatically increased over the world. The ABCD acronym, which stands for Asymmetry, Border irregularity, Color variegation, and Diameter (ABCD), was developed in 1985 by a research team at New York University as a straightforward but powerful tool to inform the general public about the early detection of melanoma.

The main goal of this research is to create a cutting-edge Convolutional Neural Network (CNN) model to classify photos of skin lesions into the appropriate cancer kinds. The dataset made available by the International Skin Imaging Collaboration is used to train and test the model (ISIC). In order to address the overfitting issue, a huge number of images must be used to train the CNN from scratch.

In [3] The technology for detecting skin cancer is largely divided into four fundamental components, starting with the collection of dermoscopic image data sets and databases, image pre-processing that includes hair removal, noise removal, sharpening, resizing, contrast stretching, and segmentation that allows for the separation of the zone of interest from the given image.

The classification that is used most frequently. Support vector machines, feedforward artificial neural networks, and deep convolutional neural networks are examples of algorithms. The medical photos were obtained from a trustworthy and dependable source. 1500 photos from the indicated two forms of skin cancer and the four categories of skin illnesses make up the dataset used to train the algorithm. Additionally, it is discovered that the SVM classifier has reduced computing complexity and is more accurate.

In [4] For doctors, this technology has exceptional and economical value. The system is divided into two categories. In the first system, the processes of picture acquisition, preprocessing, and classification are present, whereas they are only present in the second system. The findings of this study demonstrate that the accuracy of the model utilizing (NB) without any pretreatment is on average 70.15%, whereas accuracy with preprocessing is 69.69%.

Supervised learning and unsupervised learning are the two categories of categorization techniques. A unique approach in the area of unsupervised learning is clustering. A neural network's realization algorithm can either be supervised or unsupervised. The proposed system uses 3297 datasets, some of which are used for testing and others for training. In actuality, there are (1800) photographs of cases of benign carcinoma and more than (1497) appearance cases of malignant skin caser type. The full ISIC archive was mined for data for the entire dataset. All of the photographs have been downsized to 224*224*3 RGB, a lower resolution.

In [5] The Support Vector Machine (SVM) algorithm and image processing techniques are used in the diagnosing methodology. A dermoscopy image of skin cancer is captured, then it is pre-processed using a variety of techniques to reduce noise and improve the image. The segmentation of the image is then done using the Thresholding approach. The GLCM approach must be used to extract some visual features. The classifier receives these features as input. SVM, or support vector machines, are employed for classification. It designates whether the displayed image is malignant or not.

Skin cancer is a fatal condition. The basic layers of skin are three. First-layer squamous cells, second-layer basal cells, and third-layer melanocytes make up the skin's outermost layer, which is where skin cancer first manifests itself. Basal and squamous cell carcinomas are also referred to as non-melanoma malignancies. Skin cancer that isn't melanoma always responds to therapy, and it seldom metastasizes to other skin tissues. If it is not discovered in the early stages, it quickly spreads to other parts of the body by invading surrounding tissues.

In [6] The development of computer-aided diagnostics that use artificial intelligence (AI) to diagnose skin cancer has garnered a lot of attention. There is an urgent need for AI systems to help physicians in this field because to the rising prevalence of skin malignancies, low awareness among a growing population, and a lack of competent clinical experience and resources. In order to differentiate between malignant and benign skin lesions in several picture modalities, including dermoscopic, clinical, and histopathological images, researchers have created AI solutions, specifically deep learning algorithms.

CHAPTER 3
Skin Cancer Classification

The suggested cancer categorization system's methods are detailed in detail in this chapter. Local binary patterns and the gray level co-occurrence matrix approach were employed for feature extraction. In gray level co-occurrence matrix method contrast, dissimilarity, energy, correlation and homogenity are extracted from cancer images. In cancer classification method, training data and testing dataset is calculated using SVM.

3.1. Feature Extraction

Deep learning and machine learning both require feature extraction. The technique of turning raw data into numerical features that can be handled while keeping the information in the original data set is known as feature extraction. Compared to using machine learning on the raw data directly, it produces better outcomes. A linear combination of the current features results in new features when using the feature extraction technique. When compared to the values of the original characteristics, the new set of features will have different values. The fundamental objective is to use fewer features to collect the same information.

When a large data set is present and there is a need to conserve resources without losing any crucial or pertinent data, the approach of feature extraction is helpful. The amount of redundant data in the data collection is decreased with the aid of feature extraction. The data reduction stage speeds up the learning and generalization phases of the machine learning process and makes it easier for the computer to create the model.

The technique of turning raw data into numerical features that can be handled while keeping the information in the original data set is known as feature extraction. By extracting features from the input data, feature extraction improves the accuracy of learnt models. Feature extraction is a method use to remove unwanted areas. Both basic features, such the extraction of color, texture, and shape, and domain-specific features can be found in image features. LBP and GLCM are used to extract texture features during picture analysis.

3.1.1 Local Binary Pattern (LBP)

The features of an image can be extracted using a variety of texture descriptors. LBP, or local binary pattern, is a straightforward and invariant grayscale texture descriptor measure for

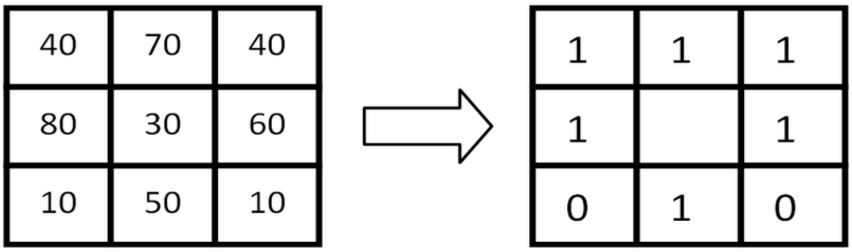
classification. In LBP, each pixel generates a binary code by thresholding the pixels in its immediate vicinity to either 0 or 1 depending on the value of the center pixel.LBP is a technique used for face representation and classification in images. The texturing operator known as Local Binary Pattern (LBP) labels the pixels in an image by thresholding the area around each pixel and treating the result as a binary number.
Formally, given a pixel at (x_c, y_c), the resulting LBP can be expressed in decimal form as:

$$LBP_{P,R}(x_c, y_c) = \sum_{P=0}^{P-1}(i_p - i_c)2^P \tag{3.1}$$

where i_c and i_P are respectively gray-level values of the central pixel and P surrounding pixels in the circle neighborhood with a radius R, and function s(x) is defined as

$$S(x) = \begin{cases} 1 \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases}$$

Where P is the number of neighborhood pixels, i_p represents the i^{th} neighboring pixel, and c represents the center pixel.



Pattern= 11110101
LBP=1x2^0+ 1x2^1+ 1x2^2+ 1x2^3+ 0x2^4+ 1x2^5+ 0x2^6+ 1x2^7=245

Figure 3.1 Calculation of Local Binary Pattern

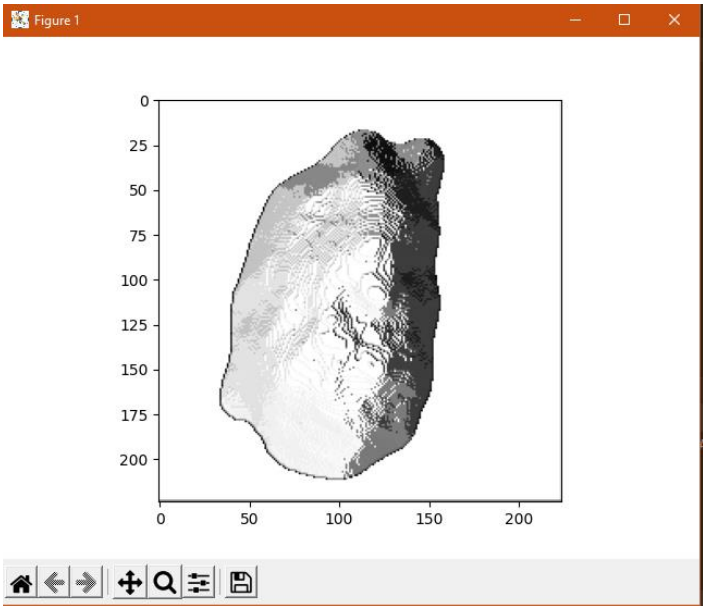


Figure 3.2 Sample Output Image of LBP

3.1.2 Gray Level Co-occurrence Matrix (GLCM)

A histogram is used in the feature extraction procedure, and the Gray Level Co-Occurrence Matrix is used to determine the outcomes (GLCM). The texture categorization notion is applied by GLCMs. The homogeneity value is used to categorize the concept of texture. Every pixel that makes up the image has its homogeneity value determined.

Statistical texture features of second order can be extracted using the Gray Level Co-occurrence Matrix (GLCM) approach. The Gray Level Co-occurrence Matrix (GLCM) is used to determine how the gray level in a picture varies from one another. The Gray Level Co-occurrence Matrix (GLCM) texture features are used in image classification, detection, clustering, diagnosis, identification and recognition problems. There are many various extraction methods depending on the type of features such as color, geometric, statistical and texture.[7] Image processing is been crucial by feature. The different features of an image are domain specific features, or shape, texture and color [8].

The number of rows and columns in GLCM exactly matches the number of image gray levels. In propose system are used contrast, dissimilarity, homogeneity, energy and correlation. The co-occurrence matrix that Haralick developed is frequently used to extract texture information. The correlations between picture pixels are used to construct the co-occurrence matrix of an intensity image. An LL matrix with the entries being the number of occurrences of a pair of pixels with brightness of a and b, distinct pixels separated by d, some distance, and pixels in a specific direction is generated for an image with a k-bit size and L=2k brightness levels.The textual characteristics of the second statistic is calculated after calculating the matrix.[6]

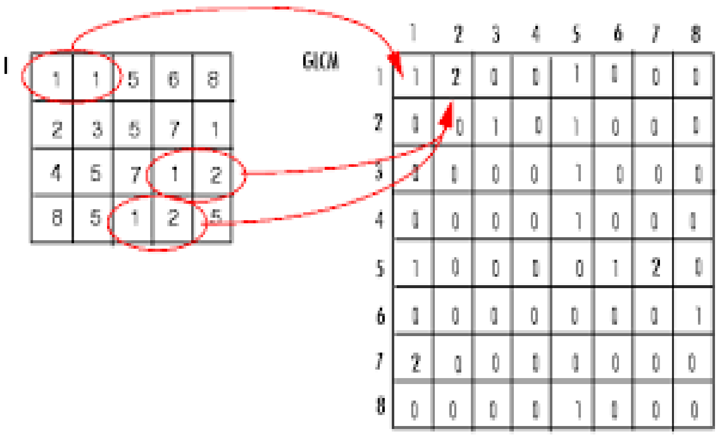


Figure 3.3 Extraction a Co-occurrence Matrix with 8 Intensity Levels

For instance, Figure 3.2 shows a 2-bit image with 8 intensity levels and a co-occurrence matrix with 8 columns and 8 rows. The matrix elements are the pixel occurrences number with gray intensity levels, i and j, which are represented by a displacement of 1 pixel in the direction of zero degrees.[6]

In various orientations, GLCM depicts the relationship between the neighboring pixels (j) and the reference pixel I Co-occurrence matrices in GLCM are built in four distinct orientations (0, 45, 90, 135). The pixel relationship is calculated horizontally and to the right (0). The two-element vector [row offset, col offset] in each row of the array designates one offset. The row offset indicates how many rows separate the target pixel from its neighbor. The amount of columns separating the target pixel from its neighbor is the col offset.

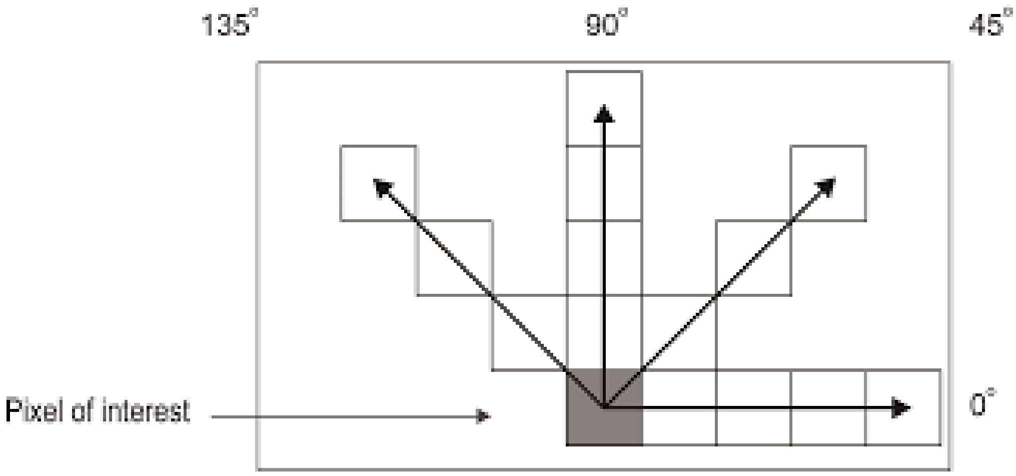


Figure 3.4 Four Different Directions with Displacement 3 between Two Pixels

1. Contrast

In the matrix of the co-occurrence of different grey levels, contrast is a local fluctuation. It can be compared to a linear dependence between the gray levels of nearby pixels. [9]

$$\text{Contrast} = \sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2 \quad (3.2)$$

In (1) where,

i and j mean the horizontal and vertical cell coordinates

p means the cell value.

If the neighboring pixels are very similar in their grey level values, then the contrast in the image is very low. In case of texture, the grey level variations indicate the variation of texture itself. If the contrast value is high, it means the image has heavy texture and low contrast values for smooth, soft textures images.[9]

2. Homogeneity

The GLCM's homogeneity refers to the consistency of the non-zero entries. [9]. It weights values by the inverse of contrast weight [9]

$$\text{Homogenity} = \sum_{i,j=0}^{N-1} \frac{Pi,j}{1+(i-j)^2} \quad (3.3)$$

Any texture's GLCM homogeneity is high if it concentrates along the diagonal, which means that many of the pixels have the same or nearly the same value for the grey scale. The GLCM homogeneity decreases as the changes in grey values increase, increasing the GLCM contrast. The homogeneity range is [0, 1]. If there is no variation in the image, homogeneity is equal to 1, and if there is little variance, homogeneity is high. Therefore, ideal repetitive structures are found in high homogeneity textures, whereas highly variable texture elements and spatial arrangements are found in low homogeneity textures. An „inhomogeneous texture" refers to an image that has almost no repetition of texture elements and spatial similarity in it is absent.[9]

3. Dissimilarity

Dissimilarity is a metric used to describe how different the pairings of grey levels are in an image. It is the closest to Contrast despite the weight difference. Dissimilarity rises quadratically in contrast to contrast [9]

$$\text{Dissimilarity} = \sum_{i,j=0}^{N-1} Pi,j|i-j| \quad (3.4)$$

Because they calculate the same parameter with different weights, it is assumed that these two measures will react similarly for the same texture. Dissimilarity will never yield slightly higher values than contrast. When the reference and neighboring pixels' grey levels are at the outermost limits of the texture sample's range of potential grey levels, the degree of dissimilarity, which runs from [0,1], reaches its maximum. [9]

4. Correlation

The linear relationship between the gray levels of adjacent pixels is measured through correlation. Digital Image Correlation is an optical approach that makes use of tracking and image registration methods to quantify changes in images accurately in 2D and 3D.

$$\text{Correlation} = \sum_{i,j=0}^{N-1} Pi,j \left[\frac{(i-\mu i)(j-\mu j)}{\sqrt{(\theta i^2)(\theta j^2)}}\right] \qquad (3.5)$$

5.Energy

Energy is a measure of local homogeneity and therefore it represents the opposite of the Entropy. Basically this feature will tell us how uniform the texture is

$$\text{Energy} = \sum_{i,j=0}^{N-1} Pi,j^2 \qquad (3.6)$$

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

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

In this thesis, contrast, homogeneity, dissimilarity, correlation and energy of a normalized GLCM transformed from a skin cancer image are used as texture features in features extraction of the system.

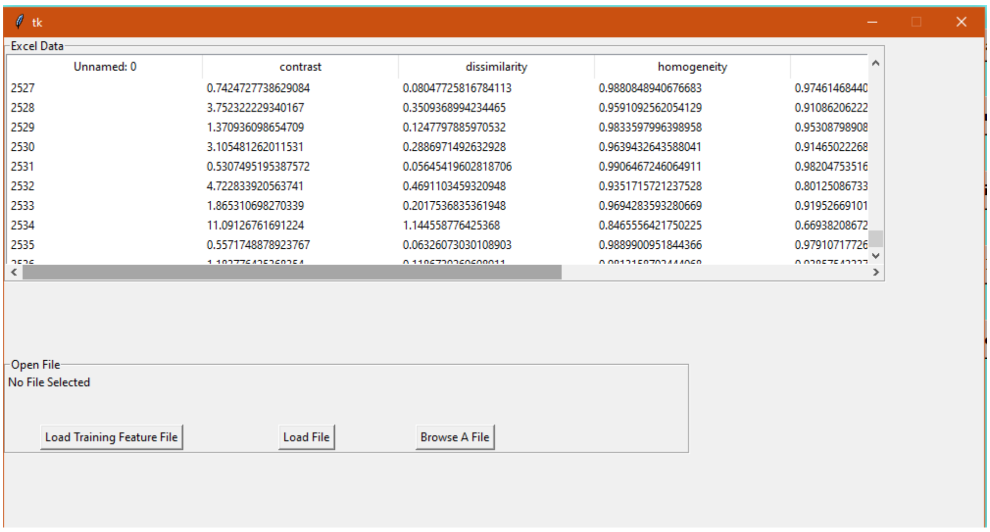


Figure 3.5 Output Result of GLCM

The value of features extracted by GLCM are presented in Figure 3.5.

3.2 Support Vector Machine

To address two-group classification issues, supervised machine learning models called support vector machines (SVM) use classification methods. After being given sets of labeled training data for each category, an SVM model may categorize new text.

People offer two key advantages over more recent algorithms like neural networks: greater speed and improved performance with fewer samples (in the thousands). The approach is thus ideally suited for text classification tasks, where it is typical to only have access to datasets with a few thousand or so tagged samples.

The following algorithm can be used to define the SVM working flow: SVM works by mapping the data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. After identifying a boundary between the categories, the data are processed so that the boundary can be seen as a hyperplane. It is therefore possible to estimate the group to which a new record should belong using the characteristics of fresh data.

The aforementioned normalized characteristics were created as a features dataset and were used to evaluate skin cancer images using SVM to classify cancer. Due to its simplicity and ease of use, the SVM method is one of the most often used machine learning techniques for classification. It is also used as the default classifier in many domain-specific situations.

A collection of mathematical operations known as the kernel are used by SVM algorithms. Data is inputted into the kernel, which then transforms it into the desired form. Different kernel functions are used by various SVM algorithms. There are various forms of these functions. SVM kernel functions include radial basis function (RBF), sigmoid, gaussian, polynomial, linear, nonlinear, and polynomial. Describe the kernel functions for vectors, text, pictures, graphs, and sequence data. RBFs are the most popular kind of kernel functions.

3.3 Advantages and Disadvantage of Support Vector Machine in Machine Learning

The data is classified by SVM (Support Vector Machine) utilizing a hyperplane, which serves as a decision boundary between several classes. Support Vectors are extreme data points from each class. The ideal hyperplane with the greatest margin from each Support Vector is what SVM seeks to identify.

The non-linear data are categorized using kernel functions or methods. It then creates a hyperplane after converting non-linear data into linear data.
The following are SVM's benefits and drawbacks:

Support vector machines' benefits (SVM)

1. SVM has L2 Regularization as one of its regularization features. Therefore, it has strong generalization abilities that keep it from fitting too tightly.

2. Effectively handles non-linear data: SVM uses the Kernel Trick to effectively handle non-linear data. SVM can be used to tackle classification and regression issues.

3. Solve both classification and regression issues. While SVR (Support Vector Regression) is used for regression issues, SVM is used for classification issues.

4. Stability: The hyperplane and, by extension, the SVM are not significantly affected by a tiny change in the data. SVM model is stable as a result.

Support vector machines' drawbacks (SVM)

1. It is challenging to select an appropriate Kernel function: It is difficult to choose the best Kernel function (to handle the non-linear data). It could be difficult and complicated. When utilizing a high dimension kernel, users risk producing an excessive number of support vectors, which significantly slows down training.

2. High algorithmic complexity and memory requirements: SVM has a very high memory need. Because you must keep all the support vectors in memory and because the size of the training dataset causes this number to rise quickly, you require a lot of memory.

3. Requires Feature Scaling: Before using SVM, feature scaling of the variables is necessary.

4. Long training period: SVM requires a lot of time to learn on big datasets.

3.4 Parameter of the RBF Kernel

The parameters C and gamma must be taken into account while training an SVM with the Radial Basis Function (RBF) kernel. The parameter C, shared by all SVM kernels, trades off incorrect classification of training samples and decision surface simplicity. A low C smoothes the decision surface, whereas a high C attempts to correctly classify every training case. Gamma expresses the strength of a particular training example. The closer additional examples must be to be impacted by a bigger gamma.

Due of their resemblance to the Gaussian distribution, RBF kernels are among the most generally utilized types of kernelization. The similarity or degree of proximity between two points, X1 and X2, is calculated using the RBF kernel function. This kernel can be mathematically represented as follows:

$$K\,(X1, X2) = \exp\left(-\frac{||X1-X2||^2}{2\sigma^2}\right) \tag{3.7}$$

Where,

1. 'σ' is the variance and our hyperparameter
2. $||X_1 - X_2||$ is the Euclidean (L_2-norm) Distance between two points X_1 and X_2

The kernel equation can be re-written as follows:

$$K\,(X1, X2) = \exp\left(-\frac{d_{12}}{2\sigma^2}\right) \tag{3.8}$$

The RBF kernel can be as large as 1, which happens when d_{12} is 0, which occurs when all the points are equal, or when X1 = X2.

1. When the points are identical, there is no space between them, making them incredibly similar.
2. The kernel value is less than 1 and near to 0, indicating that the points are dissimilar, when the points are far apart.

Since it can be shown that when distance between the points grows, they become less similar, distance can be regarded of as an equivalent to dissimilarity.

It is crucial to determine the appropriate value of " σ " in order to determine which points should be treated similarly, as can be shown on a case-by-case basis.

When $\sigma = 1$, $\sigma^2 = 1$ and the RBF kernel's mathematical equation will be as follows:

$$K(X_1, X_2) = exp\left(-\frac{||X_1 - X_2||^2}{2}\right) \tag{3.9}$$

Because it resembles the K-Nearest Neighborhood Algorithm, RBF Kernel is well-liked. As RBF Kernel Support Vector Machines only need to store the support vectors during training and not the complete dataset, it has the benefits of K-NN and solves the space complexity issue.

The scikit-learn library implements the RBF Kernel Support Vector Machines, which have two associated hyperparameters: "C" for the SVM and "" for the RBF Kernel. Here, the relationship between and is inverse.

$$\gamma \ \alpha \ \frac{1}{\theta}$$

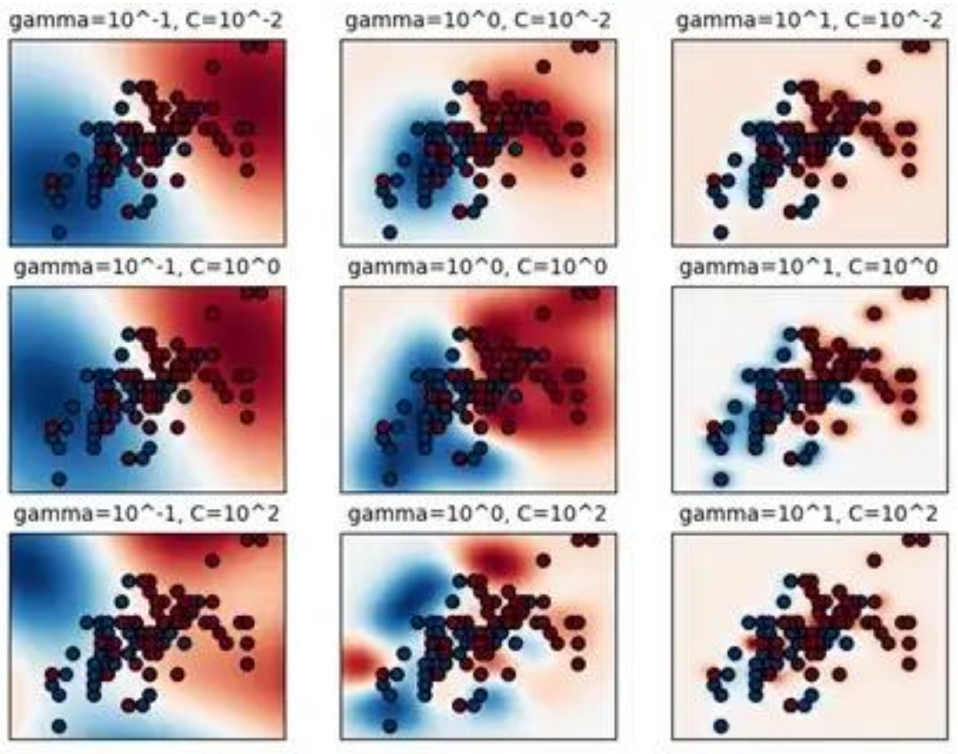


Figure 3.6 RBF Kernel SVM

From the figure, γ increases, i.e. σ reduces, the model tends to overfit for a given value of C.
To get the optimal Bias-Variance Trade off, it is crucial to identify the appropriate γ or σ and C value.

CHAPTER 4
SYSTEM DESIGN AND IMPLEMENTATION

This chapter describes the proposed SVM-based cancer classification system's design and implementation. In order to comprehend the suggested cancer classification system's planned process flow, an overview of the system design is first provided. The use of the PYTHON programming environment to implement each process in the suggested system is then detailed with user interfaces (UI). In this chapter, the performance analysis and the experimental findings are eventually presented in charts, figures, and tables.

4.1 Overview Design of the System

The overview design of the proposed skin cancer diagnosis system is shown in Figure 4.1. The proposed skin cancer diagnosis system is implemented as the skin cancer classification system by using SVM. The contribution of the proposed system is segmentation and texture features extraction and skin cancer diagnosis based on these extracted features using SVM as a classifier. In this system, three main steps are essential.

In the first step, pre-preprocessing, which consists of acquiring cancer images and image color converting, image enhancement and noise removal is performed. The cancer images are size of 224*224.

Using the extracted feature vectors, the features dataset is constructed in the second stage after picture segmentation and texture feature extraction from the cancer images, as explained in chapter 3 of the preceding chapter.

The system is then used to insert a cancer image that has been tested, and features are also extracted from this image in the final stage. Using SVM as a classifier, the tested cancer image is identified based on distances between its features vector and the features vectors in the features dataset.

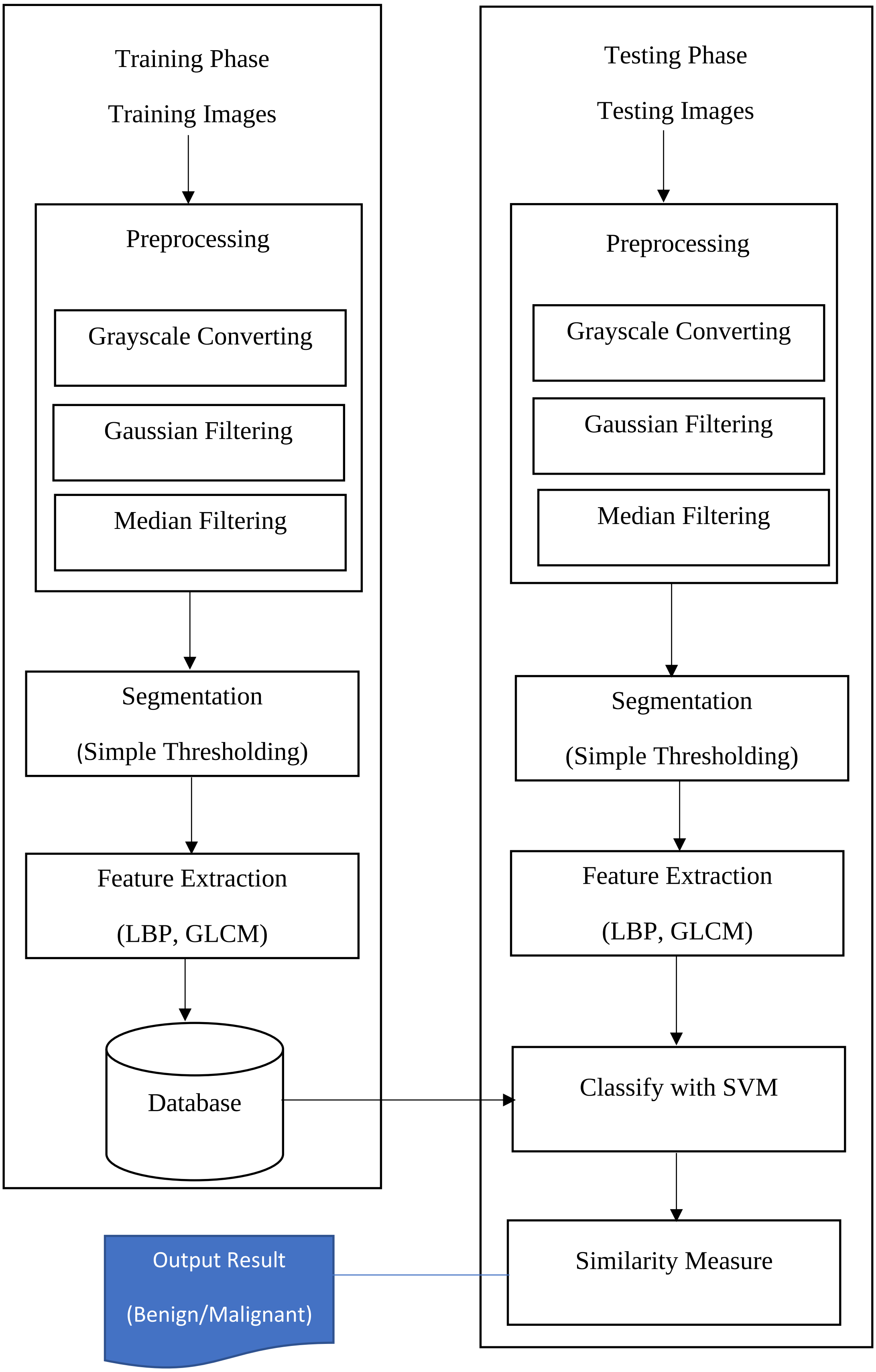


Figure 4.1 Overview Design of the System

4.2 Architecture of the System

The suggested system's architecture consists of three stages: pre-processing, picture segmentation and feature extraction, and cancer classification. Valid and practical properties are necessary for it to function as a categorization system. The segmentation of the image and features extraction completes this stage. In the stage of skin cancer classification, features from tested skin photos are collected, and they are compared with the features dataset using SVM to classify the skin cancer. Finally, it will become apparent if the system has cancer.

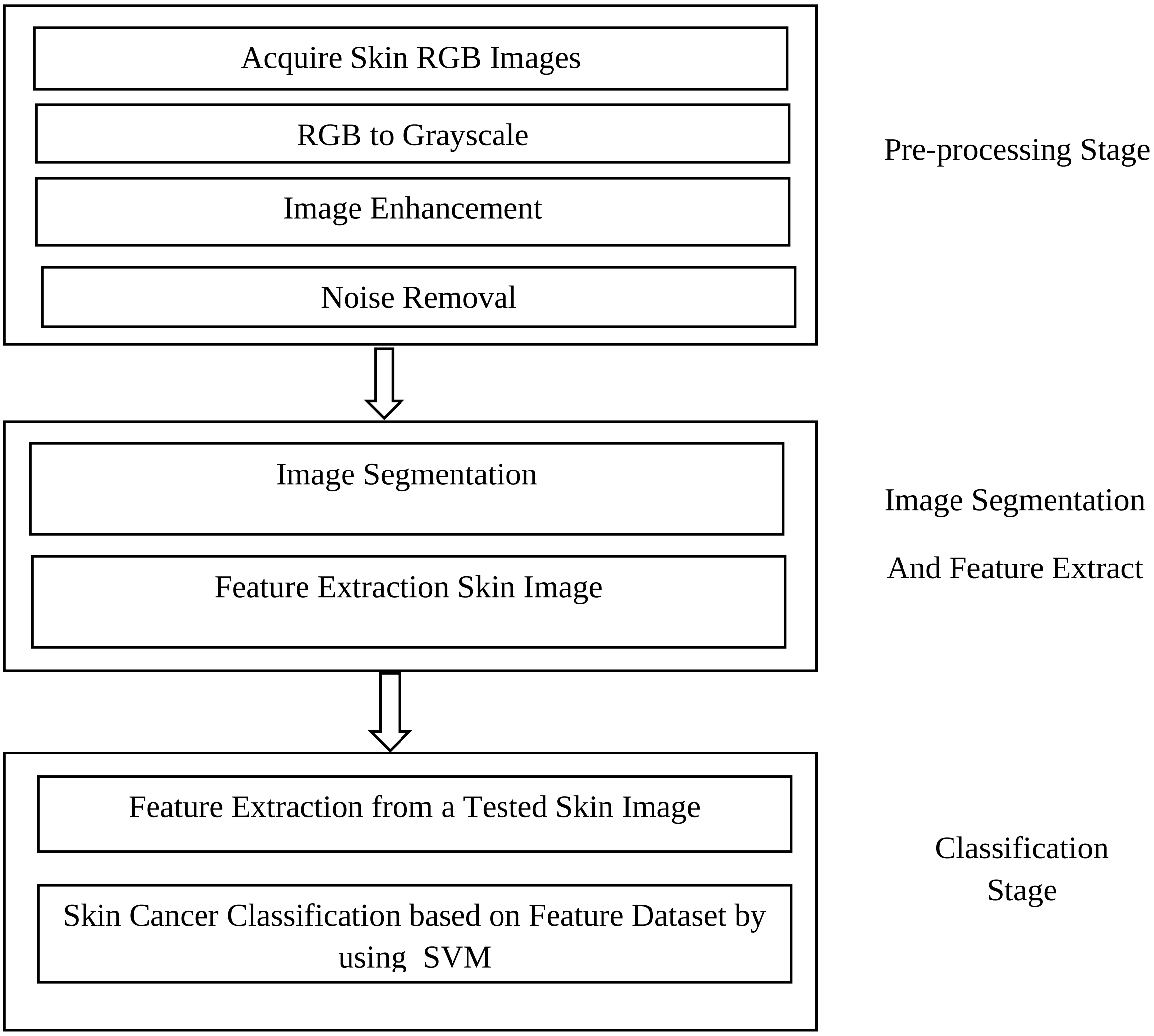


Figure 4.2 Architecture of the System

4.3 Implementation of the system

The overview architecture depicted in Figure 4.1 is used to create the cancer classification system for skin cancer using SVM. The system is put into use using the PYTHON programming language.

Four significant buttons make up the system's main graphical user interface (GUI): Features Cancer Classification as depicted in Figure 4.3.

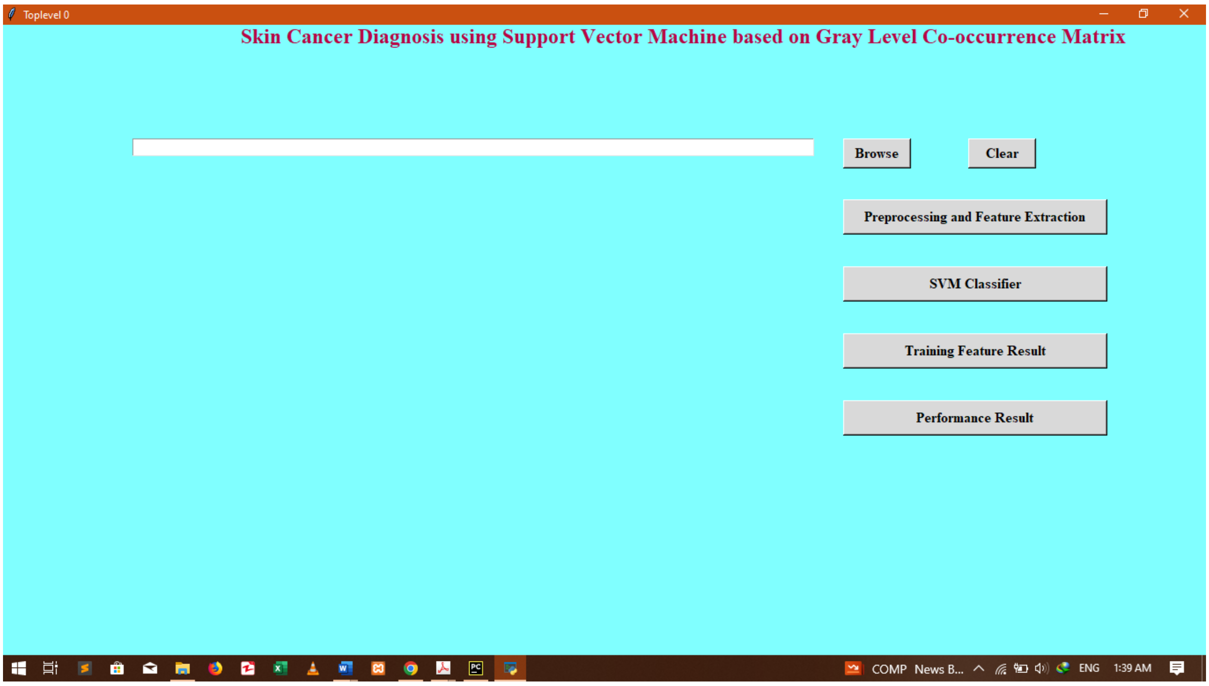


Figure 4.3 Main GUI of the System

Figure 4.4 Original image on GUI of the System

The main page of the proposed system is presented in Figure 4.3. Figure 4.4 depicted how to load the original image for testing.

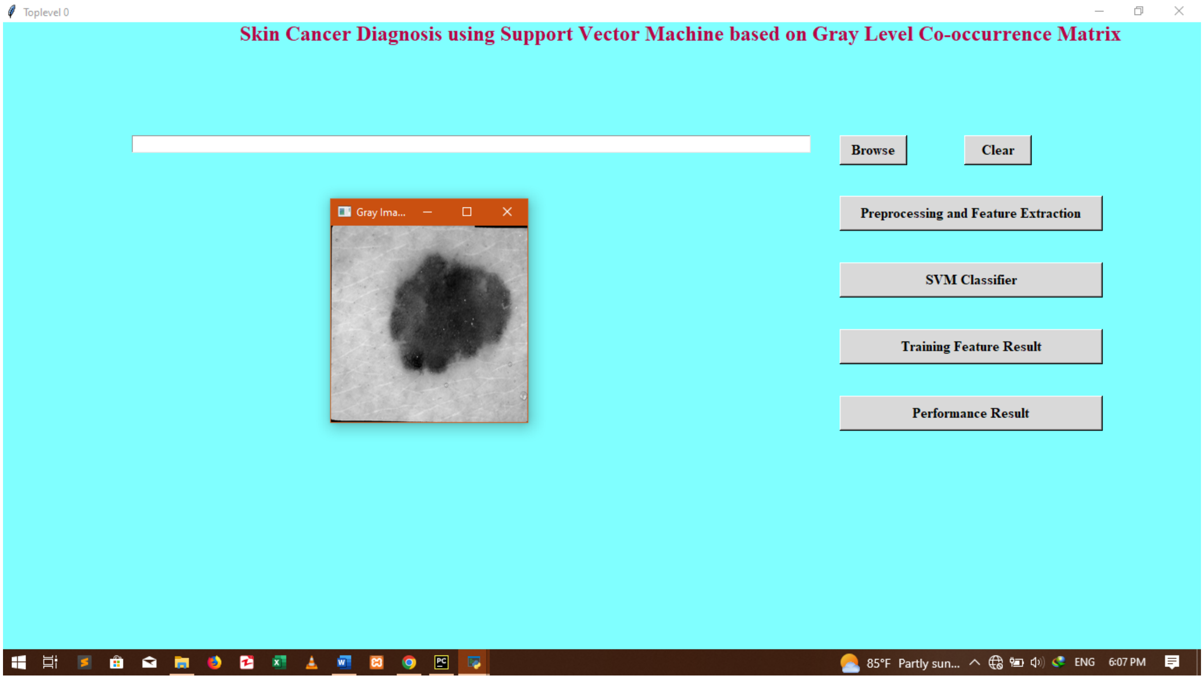


Figure 4.5 Gray Scale on GUI of the System

Figure 4.5 depicted how to convert from original image to Gray scale for testing.

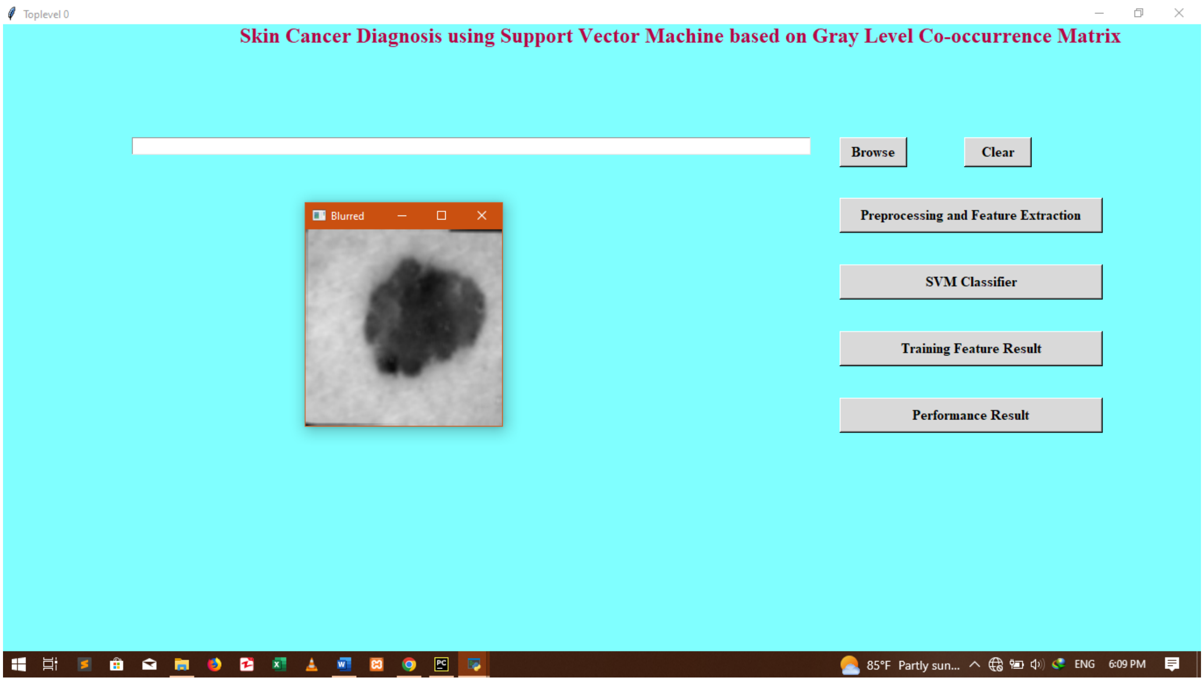


Figure 4.6 Blur image on GUI of the System

Figure 4.6 depicted how to change image smoothing for testing.

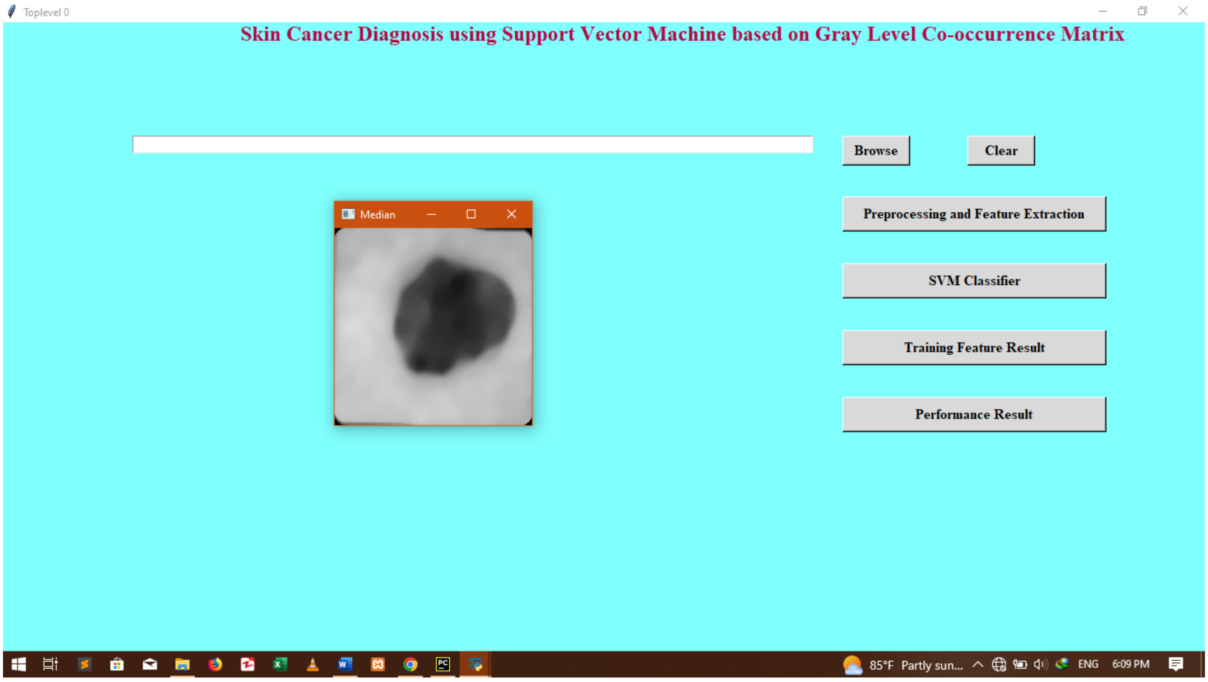


Figure 4.7 Median Image on GUI of the System

Figure 4.7 depicted how to remove unwanted noise for testing.

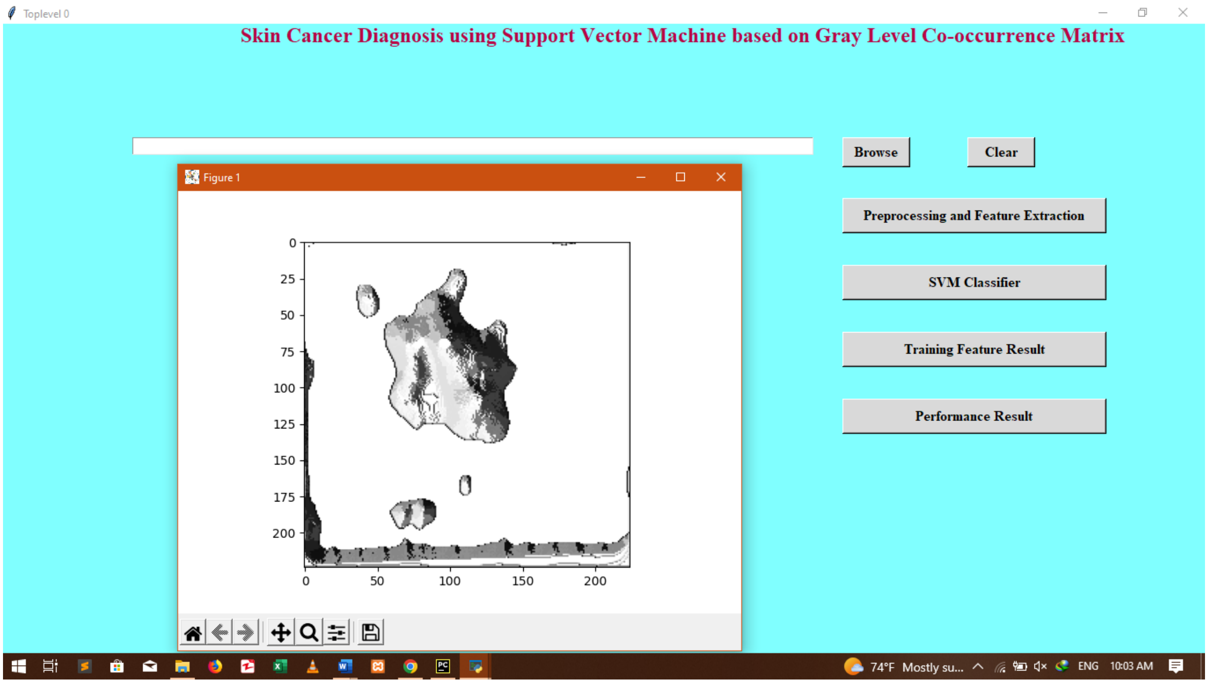


Figure 4.8 Local Binary Pattern on GUI of the System

Figure 4.8 depicted how to extracted feature from LBP for testing.

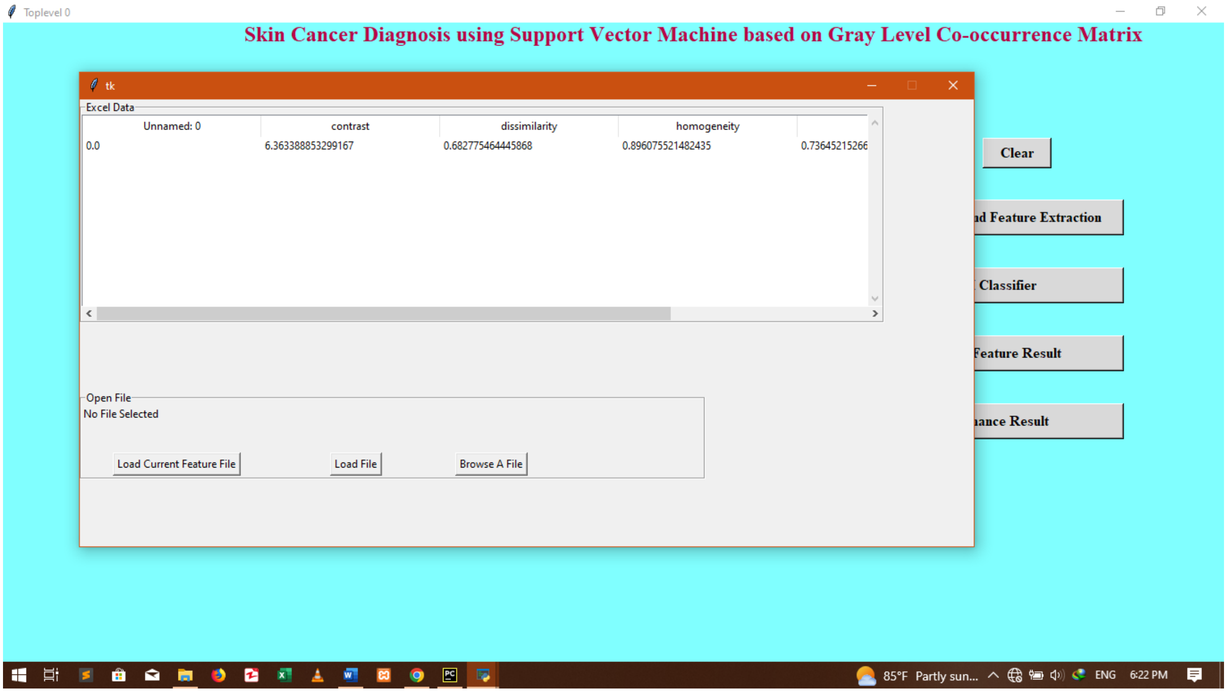


Figure 4.9 Input Image Tested Feature on GUI of the System

Figure 4.8 depicted how to extracted feature contrast, dissimilarity, homogeneity, energy, correlation from GLCM for testing.

4.3.1 Implementation of Features Extraction

Contrast, dissimilarity, homogeneity, energy, and correlation are recovered from each layer of segmented grayscale images during the features extraction process. Features extracted and their outcomes are as illustrated in Figure 4.10.

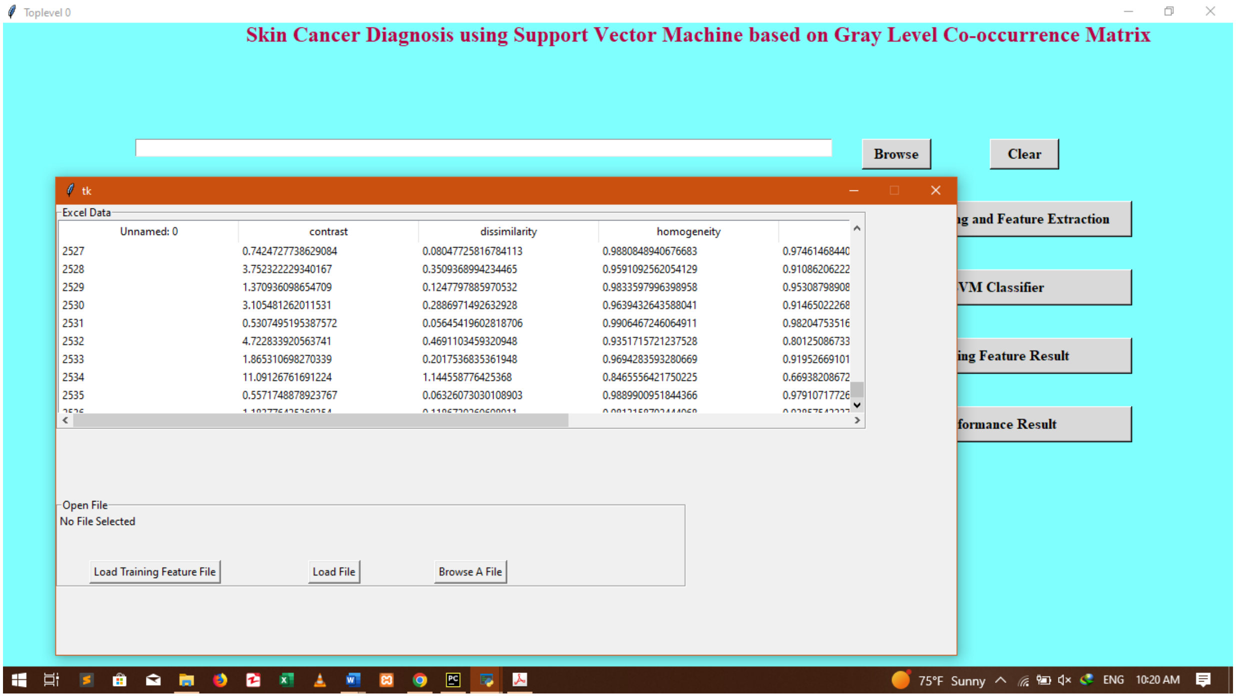


Figure 4.10 Training Dataset of the System

For eventual use in cancer classification, the extracted features are saved as features in an excel file with the (.xlsx) extension. Contrast, dissimilarity, energy, homogeneity, and correlation of each layer of a Grayscale image are the five features that each feature vector carries. The features dataset refers to the file that contains the extracted features. Dataset of features is visualized is as shown in Figure 4.10 and Figure 4.11

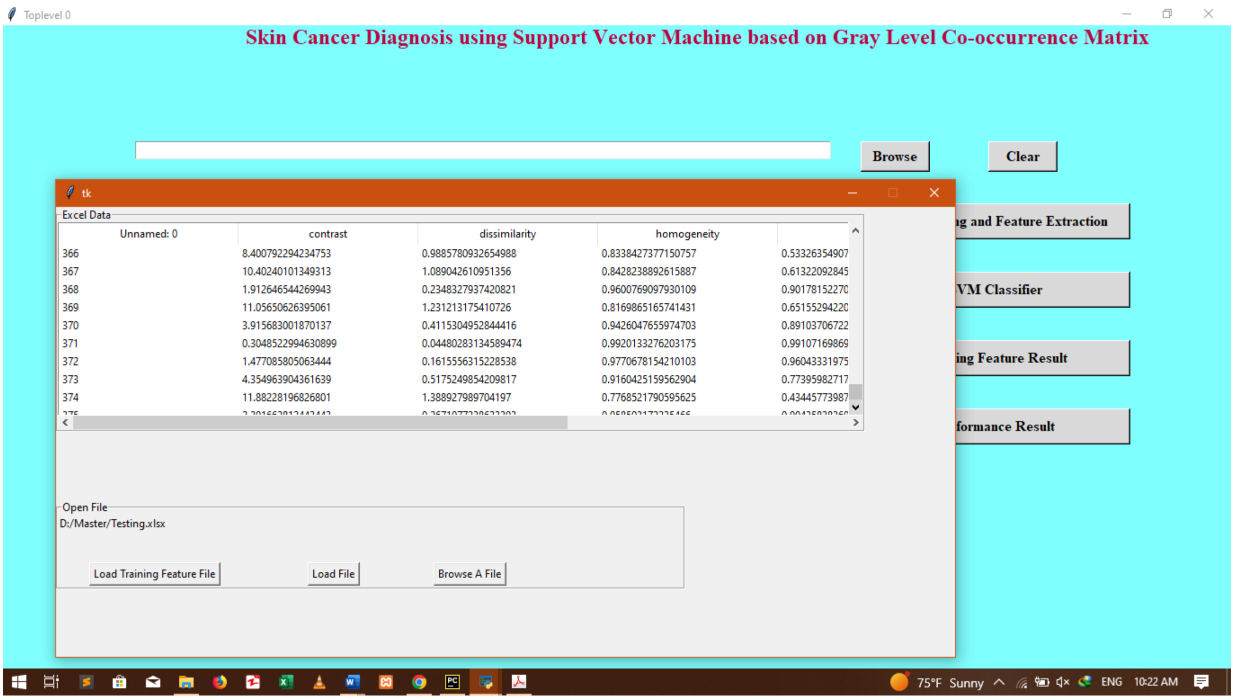


Figure 4.11 Testing Dataset of the System

4.4. Experimental Result

The following table contains the information of the images that are used to train and test for the proposed system:

Table 4.1 Dataset Description

Image	Train Data Set	Test Data Set
Size	224 * 224	224 * 224
Quantity	2536	375
Kind	Benign, Melignant	
Source	International Skin Imaging Collaboration (ISIC).	

Table 4.2 Sample Images of Benign

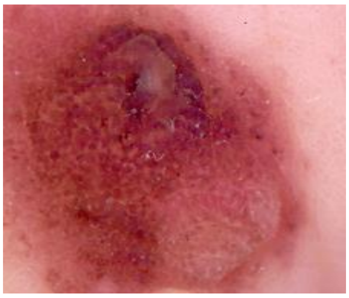
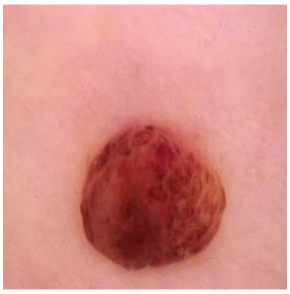
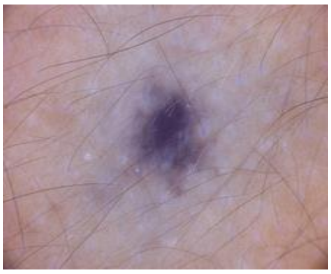
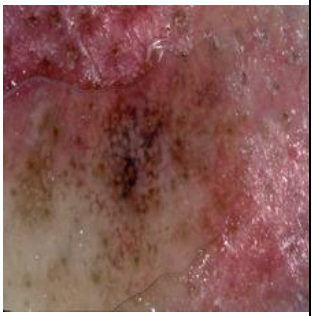
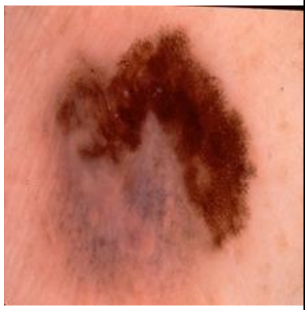
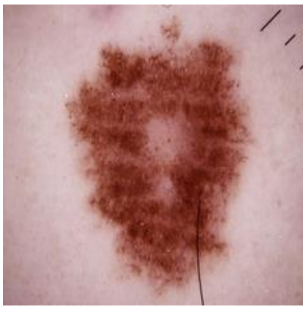
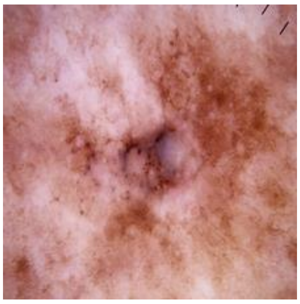
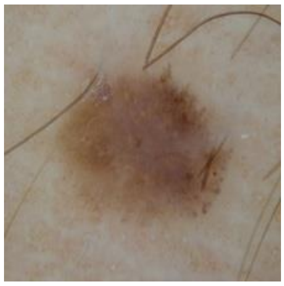
			
			

Table 4.3 Sample Images of Melignant

The performance evaluation of the system are printed as the following figure4.1 According to this experiment, the accuracy of the benign is better than malignant. The system accuracy is 0.86.

Accuracy = (TP + TN) / (TP + TN + FP + FN) (4.1)

Specificity = (TN) / (TN + FP) (4.2)

Sensitivity = (TP) / (TP + FN) (4.3)

Precision = (TP) / (TP+FP) (4.4)

Where,

- TP = the number of cases correctly identified as positive for disease
- FP = the number of cases incorrectly identified as positive for disease
- TN = the number of cases correctly identified as negative for disease
- FN = the number of cases incorrectly identified as negative for disease

The performance evaluation of the system are printed as the following figure 4.12.

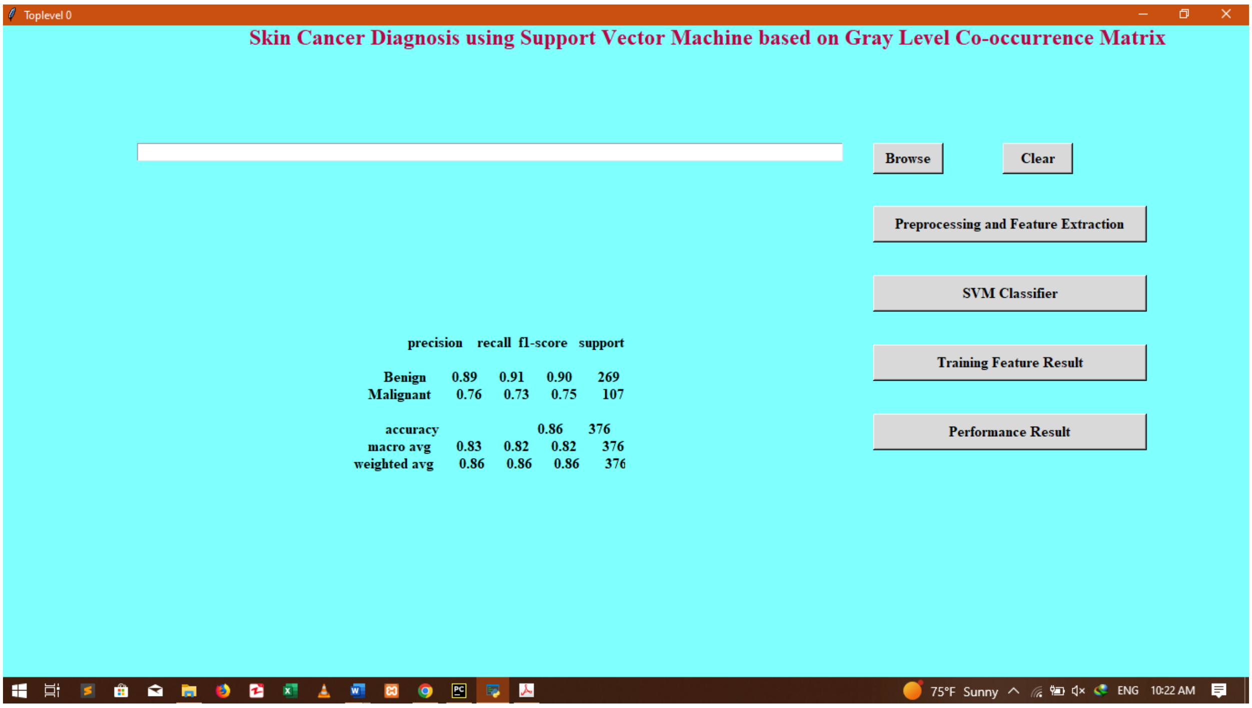


Figure 4.12 Evaluation Result

The performance evaluation of the system are printed as the following figure4.12

Table 4.4 Evaluation Result of the System

	Precision	Recall	F1-Score	Support
Benign	0.89	0.91	0.90	269
Malignant	0.76	0.73	0.75	107
Accuracy			0.86	376
Marco avg	0.83	0.82	0.82	376
Weighted avg	0.86	0.86	0.86	376

According to this experiment, the accuracy of the benign is better than malignant. The system accuracy is 0.86.

CHAPTER 5

CONCLUSION

Because melanoma skin cancer has been sharply increasing over the past few decades, a texture-based automated skin cancer diagnosis model has been presented. The image is segmented using the Simple Thresholding Algorithm after being preprocessed to boost resolution. Following that, LBP and GLCM are used to extract features from the filtered image. SVM is used to classify the texture features, and it can recognize melanoma from dermoscopy images. The experimental result shows that the accuracy result of 86%.

Gray level co-occurrence matrix and support vector machines can be used to quickly classify whether an image is carcinogenic or not as part of the skin cancer detection method that has been presented. Preprocessing, image segmentation, feature extraction, and classification are some of the phases in the process of identifying skin cancer. The SVM for lesion picture categorization was the main topic of this review. Both advantages and disadvantages of algorithms exist.

This research proposed the diagnosis of skin cancer using Support Vector Machine and the two feature extraction methods, LBP and GLCM. By using this system, the experimental result gets the accuracy of 86%. In this approach, features from LBP are first extracted, and only five features from GLCM are then employed, yet the testing showed that this achieved the right level of accuracy.

5.1 Advantage and Limitations of the System

The user only needs to snap a picture of a wound on the skin in order to use the proposed skin cancer classification system, which is straightforward but incredibly user-friendly and helpful. The system can easily detect whether a person has skin cancer and quickly treat it. The proposed skin cancer segmentation features are extracted from the skin and classified using SVM faster. The proposed skin cancer classification system reach up to 86% accuracy in skin cancer classification.

AUTHOR'S PUBLICATIONS

[1] Myint Myint Wai, Win Lelt Lelt Phyu "Skin Cancer Diagnosis using Support Vector Machine based on Gray Level Co-occurrence Matrix" The National Journal of Parallel and Soft Computing, (NJPSC 2022),2022

Reference

[1] A. Gebejes R. Huertas "Texture Characterization based onGrey-Level Co-occurrence Matrix" in March, 25. - 29. 2013

[2] B. Arivuselvam1, S. Tanisha2, S. Shalini3, V.S. Subhalaksmi4 "Skin Cancer Detection And Classification Using SVM Classifier" *Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021)*

[3] Danashree S. Kalel, Pooja M. Pisal, Ramdas P. Bagawade, "Color, Shape and Texture Feature Extraction for Content-based Image Retrieval System: A Study", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), ISSN (Online) 2278-1021, ISSN (Print) 2319 5940, Volume 5, Issue 4, April 2016.

[4] Hardik Nahata and Satya P. Singh "Deep Learning Solutions for Skin Cancer Detection and Diagnosis" in March 2020

[5] Md. Muzahidul Islam Rahi Farhan Tanvir Khan Mohammad Tanvir Mahtab "DETECTION OF SKIN CANCER USING DEEP NEURAL NETWORKS" in 24 March 2021

[6] Marcos A. M. Almeida" Classification Models for Skin Tumor Detection Using Texture Analysis in Medical Images" in 13 May 2020

[7] Mehwish Dildar 1, Shumaila Akram 2, Muhammad Irfan 3 , Hikmat Ullah Khan 4 , Muhammad Ramzan 2,5, Abdur Rehman Mahmood 6, Soliman Ayed Alsaiari 7, Abdul Hakeem M Saeed 8, Mohammed Olaythah Alraddadi 9 and Mater Hussen Mahnashi 10 "Skin Cancer Detection: A Review Using Deep Learning Techniques" in Accepted: 13 May 2021 Published: 20 May 2021

[8] Noortaz Rezaoana, Mohammad Shahadat Hossain, Karl Andersson "Detection and Classification of Skin Cancer by Using a Parallel CNN Model" 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)

[9] Ohood Fahdil Alwan "SKIN CANCER IMAGES CLASSIFICATION USING NAÏVE BAYES ALGORITHM" ISSN: 2776-0995 Volume 3, Issue 4, April, 2022

[10] P. Janani * , J. Premaladha and K. S. Ravichandran "Image Enhancement Techniques" Indian Journal of Science and Technology, vol 8(22), DOI: 10.17485/ijst/2015/v8i22/79318, September 2015

[11] Shwe Yee Win, Thin Lai Lai Thein"Soil Classification for Agriculture Crops using K-Nearest Neighbors (KNN)", The National Journal of Parallel and Soft Computing(NJPSC 2022),2022

[12] T. M. Mitchell, "Machine Learning", 1st Edition, McGraw-Hill Science/Engineering/Math, March 1997, ISBN: 0070428077.

[13] Zhongxin Gao "An Adaptive Median Filtering of Salt and Pepper Noise based on Local Pixel Distribution" International Conference on Transportation & Logistics, Information & Communication, Smart City (TLICSC 2018)