SUNFLOWER DISEASE PREDICTION USING ARTIFICIAL INTELLIGENCE IN PERCEPT OF BANGLADESH

ZAKIA SULTANA
ROLL: BKH1711MSc110F
SESSION: 2017-2018

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

SUNFLOWER DISEASE PREDICTION USING ARTIFICIAL INTELLIGENCE IN PERCEPT OF BANGLADESH

ZAKIA SULTANA

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science in Information and Communication Engineering

Department of Information and Communication Engineering
NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

November 2022

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this research and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Master of Science in Department of Information and Communication Engineering of Noakhali Science and Technology University

____

DR. MD. ASHIKUR RAHMAN KHAN
PROFESSOR & CHAIRMAN
Department of Information & Communication Engineering
Noakhali Science & Technology University
Noakhali- 3814, Bangladesh

STUDENT'S DECLARATION

I hereby declare that this thesis has not been submitted elsewhere for the requirement of any kind of degree, diploma or publication.

Zakia Sultana
Dept. of ICE, NSTU
Roll No: BKH1711MSc110F
Session: 2017-18

ACKNOWLEDGMENTS

I am grateful and would like to express my sincere gratitude to my supervisor Professor Dr. Md. Ashikur Rahman Khan for his germinal ideas, invaluable guidance, continuous encouragement, and constant support in making this research possible. He has always impressed me with his outstanding professional conduct, his strong conviction for science, and his belief that an MSc program is only the start of a life-long learning experience. I appreciate his consistent support from the first day I applied to the graduate program to these concluding moments. I am truly grateful for his progressive vision about my training in science, his tolerance of my naive mistakes, and his commitment to my future carrier. I also would like to express very special thanks to my co-supervisor Zayed-Us-Salehin for his suggestions and co-operations throughout the study. I also sincerely thank you for the time spent proofreading and correcting my many mistakes.

My sincere thanks go to all my lab mates and members of the staff of the Information and Communication Engineering Department, NSTU, who helped me in many ways and made my stay at NSTU pleasant and unforgettable.

I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream, and sacrifice throughout my life. Special thanks should be given to my committee members. I would like to acknowledge their comments and suggestions which were crucial for the successful completion of this study.

ABSTRACT

Plant disease prediction and recognition in the early stage is one of the most essential needs to increase agriculture, which plays an important role in our country's economy and helps to feed a large population. Sunflower is a plant categorized as a low to medium drought-sensitive crop. But now a days sunflower production faced a sever crisis due to its many diseases. Farmers also lost their interest to the production of sunflower. But if proper action is received earlier, many serious disease will have no effect on the plants. It also improve the productively, quantity and quality of sunflower plants. Manual identification of sunflower disease is an exhausting task at time. Now Artificial intelligent is a great field and gained its popularity for the prediction of plant disease as early as possible. In this research paper different machine learning techniques such as Linear Regression (LR), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Extreme Gradient Boosting (XGBoosting) are applied in the image features that are extracted using digital image processing technique. And deep learning convolution neural network (CNN) techniques such as Residual Network 50 (ResNet 50), Visual Geometry Group 19 (VGG 19) and Mobile Net Version 2 (MobileNet V2) are used in the direct image for predicting the disease of sunflower. A total of 602 images which are collected from Bangladesh Agricultural Institute, gazipur, is used in this research work. Machine learning techniques GB shows 93.38 % accuracy on the other hand XGBoost Shows 92.56% accuracy but highest sensitivity 0.92. Deep learning model ResNet 50 Shows highest accuracy and sensitivity 96.19 % and 0.96. Among Machine learning and Deep learning models ResNet 50 shows highest accuracy and sensitivity, therefore ResNet 50 is the best model for predicting sunflower disease for this dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Detail
AI	Artificial Intelligence
ML	Machine Learning
CNN	Convolutional Neural Network
LR	Logistic Regression
RF	Random Forest
SVM	Support Vector Machine
GB	Gradient Boosting
KNN	K-Nearest Neighbor
XGBoost	Extreme Gradient Boosting
ResNet 50	Residual Network 50
VGG19	Visual Geometry Group Layer 19
MobileNet V2	Mobile Network Version V2
FPR	False Positive Rate
FNR	False Negative Rate

LIST OF STMBOLS

Symbol	Meaning
μ	Mu
σ	Sigma
α	Alpha
β	Beta
γ	Ghama
θ	Theta
ρ	Rho
Φ	Phi

CHAPTER 1

1. INTRODUCTION

1.1 Introduction

The introductory study of this research work is focused out in this chapter. Since the research work worked on the prediction of sunflower disease prediction. For that the background analysis on the sunflower disease is carried out first. The motivation then paddled out what is the reason for this research. The problem statement is then introduced and the research objectives are described on this basis. And finally the outline of this research is presented.

1.2 Background

Sunflower scientific name *Helianthus annuus,* is a blossomy annual plant in the family Asteraceae, grown for its seeds. A sunflower plant known as a seed of a productive oil crop. The flowers come in many colors (yellow, red, orange, maroon, brown), but they are commonly bright yellow with brown centers that ripen into heavy heads filled with seeds. A flower has hundreds or thousands of florets that is a tiny flower in a large flower head. The common sunflower is valuable from an economic as well as an ornamental point of view. The leaves are used as fodder, the flowers yield a yellow dye, and the seeds contain oil and are used for food. Sunflower oil cake is used for stock and poultry feeding. The oil is also used in soap and paints and as a lubricant. The seeds may be eaten dried, roasted, or ground into nut butter and are common in bird seed mixes. This cash crop is primarily cultivated in North and South America. In Bangladesh, it was first introduced as a garden plant, but in the 1980s some dwarf varieties were introduced to cultivate as an oil seed crop.

Roasted sunflower whole seeds are sold as snacks low in cholesterol and sodium. The seeds also improve brain power and digestion. Sunflower seeds can also be processed into sunflower butter which is used as an alternative to peanut butter. Sunflower oil extracted from sunflower seeds is a popular cooking oil used in most kitchens because of its various health benefits. Sunflower oil is rich in Vitamin E; thus it acts as an oxidant. The Vitamin E extracted from sunflowers also prevents asthma, arthritis, and colon cancer. The sunflower cake which remains after the seeds have been processed into oil is a nutritious feed for livestock. The Chinese use the fibers from the stem of the sunflower for fabric and paper [1]. The pith is the lightest known substance and is used in scientific laboratories for experiments. In the Netherlands, sunflowers have been used to reclaim marshes because of their ability to absorb more water [1]. Ukraine and Russia remain the top sunflower seed producers in the world. They produce almost half of the world's sunflower seeds. Ukraine produced 17.5 million metric tons of Sunflower seeds while Russia produced 15.5 million metric tons in 2022/2021 [1].

1.3 Motivation

The rate of sunflower production is decreasing day by day as it is affected by various diseases. Till now 30 sunflower diseases have been identified on wild or cultivated sunflowers, but only a few are of economic significance as far as causing yield losses [2]. Diseases of sunflowers include those caused by fungi, bacteria, viruses, and nematodes. Most diseases commonly affecting sunflowers in this region are caused by fungi, and are generally more severe in irrigated production. The most common fungi diseases that cause most are Downy Mildew, gray mold, leaf Rust, Charcoal rot, Leaf Scars and many others. Identification of diseases in plants has been carried out by experts through naked eyes based on their vivid experience. This method is also not practically feasible for large fields; sometimes this also does not provide accurate results. Accurate identification and control of disease in sunflower plants can be precisely acknowledged by automatic detection of the disease symptoms appearing on sunflower plant leaves. Detection of plant diseases using modern available techniques involves image processing, pattern recognition and some automatic classification tools. These are proving useful for farmers in improving the quality and quantity of crop products.

1.4 Problem Statement

Early detection of sunflower disease for preventing diseases spreading and improving crop production. In this work four diseases, downy Mildew, gray mold, leaf Rust, leaf scars and disease free sunflower can be detected using image processing and by six different classifiers and three deep learning classifiers. The disease affected region of the acquired image is segmented by using k-means clustering. Features are extracted by using GLCM from segmented images. For prediction Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, k-Nearest Neighbors, XGBclassifier have been used. Seven performance metrics have been estimated in this research for finding the best classifier of the diseases. This paper used three deep learning convolutional neural network model ResNet 50, VGG19 and Mobile Net V2 for predicting the sunflower disease and four performance metrices is considered for finding the best classifier. Finally this paper finds the best techniques among machine learning deep learning classifiers.

1.5 Objectives

This research paper worked on different machine learning and deep learning techniques and produce better result that will help in harvesting sunflower. The main purpose of this research is as follows

1. To early predicting sunflower disease for reducing vast production loss and improve crop planting.

2. To detect different sunflower disease different machine learning and deep learning techniques such as Logistic Regression, Random Forest, Gradient Boosting, Support Vector machine, K- Nearest Neighbors, XGBoost classifier, ResNet 50, Mobile Net V2, VGG 19 is used.

3. To introduce the best machine learning model and CNN model based on accuracy and sensitivity of that model among different models.

4. To Rank the extracted features based on the contribution to produce model output using Anova F1 features ranking algorithm.

5. To find the best techniques among all the techniques.

1.6 Research Outline

This thesis consists of 5 chapters and these chapters are organized as follows, in chapter 1 it provides a detailed discussion on the importance of the work that has been done and why the current topic is selected for thesis. In the next that is chapter 2 is about some related works of this thesis is described in this section. It gives the information about dengue outbreaks of different countries using different classification algorithms and techniques. Next is chapter 3 which introduces the dataset and the methods and explains deep understanding about the algorithms and evaluating criteria. Then the chapter 4 describes experimental plans and prediction test results on evaluation criteria and shows comparison of different algorithm's performance. Finally, in Chapter 5 5, paper will end up as allowing to know why this study is a valuable for sunflower cultivation in the conclusion.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, different past works related with sunflower data collection, disease prediction method and prediction accuracy are described in short. These three things which are related to dengue disease prediction are discussed in literature review. And lastly, the summary of this chapter which is the outcome of literature review is described briefly.

2.2 Literature Review

Automatic methods for an early detection of plant diseases are vital for precision crop protection. Several automatic approaches are used in the agricultural field including investigation of plant disease and pests. Many research works have been done on the detection and prevention of plant disease in general and sunflower disease in particular. In this paper I introduced different research work on sunflower disease along with different plant diseases.

V. Singh [3] developed a system for detecting leaf disease in sunflowers utilizing image segmentation. The classification of the leaf diseases was performed by the Particle Swarm optimization algorithm. They didn't conclude any specific size of image data but obtained 92.7% accuracy.

A method for identifying the leaf diseases of sunflowers using Random Forest classifier was proposed by Jun Liu et al [4]. They adopted k-means clustering and watershed algorithms for segmenting the diseases affected region and image processing technique to extract the 19 features. The obtained accuracy was 95.00%.

Sinan Uguz and Nase Uysal [5] introduce a system for classifying olive leaf disease using deep learning. They used transfer learning methods on VGG16 and VGG19 architectures, as well as proposed CNN architecture. This paper shows augmented data producing the highest success value in trained models was 95%, whereas without data augmentation the highest value was 88%. This paper also focuses on Adam, AdaGrad, Stochastic gradient descent and RMS Prop optimization algorithms' effect on the network's performance and finds that Adam and SGD optimization algorithms were generally observed to generate superior results.

Sammy V. Militante, Bobby D. Gerardo and Ruji P. Medina [6] brings a system to identify disease of sugarcane using machine learning through computer vision using deep learning techniques. They found 95% accuracy by applying deep learning Convolutional neural network technique on 13,842 sugarcane image dataset of disease infected leaves.

Four modified deep learning models are developed by Zhaohua Huang et al. [7] for grape leaf disease detection and classification based on a developed grape leaf dataset. In this paper VGG16, MobileNet, and AlexNet transfer learning techniques have been used. They show better accuracy and performance improvement compared with pretrained models. Also an ensemble model improves final detection and classification accuracy.

For early detection of tomato plant leaf disease Surampalli Ashok et al. [8] established a system using image processing techniques based on Image segmentation, clustering, and open-source algorithms. This paper proposed a method and found 98.18 % accuracy compared with automated capturing systems using Deep Neural Networks (AlexNet) and ANN technique.

S. Gayathri et.al [9] developed a deep CNNs called LeNet to discover the tea plant diseases from leaf image sets. This model was applied to detect 4 diseases of the tea plant and produce 85% to 98% accuracy for detecting these 4 types of disease and showed a high accuracy level compared with traditional backpropagation networks.

An automated method was introduced by T. Rumpf et.al., [10] for the early detection and differentiation of sugar beet diseases based on Support Vector Machines and spectral vegetation indices. This paper focus on discriminating diseases from non-diseased sugar beet leaves, differentiate between the diseases Cercospora leaf spot, leaf rust and powdery mildew and identify diseases even before specific symptoms become visible. The discrimination between healthy sugar beet leaves and diseased leaves resulted in classification accuracies up to 97% compair with dicision trees and artificial neural networks. And SVM shows the multiple classification between healthy leaves and leaves with symptoms of the three diseases accuracy higher than 86%.

A system was introduced by Edna Chebet Too et.al, [11] for fine-tuning and evaluation of state-of-the-art deep convolutional neural network for image-based plant disease classification. This paper produced a comparison of the deep learning models such as VGG 16, Inception V4, ResNet with 50, 101 and 152 layers and DenseNets with 121 layers architecture. 38 different classes including diseased and healthy images of leaves of 14 plants from plantVillage used for this experiment. They proved that DenseNets has a tendency to consistently improve in accuracy with growing number of epochs, with no signs of overfitting and performance deterioration. DenseNets requires a considerably less number of parameters and reasonable computing time to achieve state-of-the-art performances. It achieves a testing accuracy score of 99.75% to beat the rest of the architectures.

A computer vision approach was proposed to extract leaf part from whole image by subtracting complex background and to introduce a deep learning convolutional neural network (CNN), SoyNet, for soybean plant diseases recognition using segmented leaf images by Aditya Karlekar and Ayan Seal [12]. The proposed model achieves identification accuracy of 98.14% with good precision, recall and f1-score. Six popularly used deep learning CNN models namely, VGG19, GoogleLeNet, Dense121, XceptionNet, LeNet, and ResNet50 were compared with the proposed model based on state-of-the-art methods and the proposed method outperforms nine state-of-the-art methods.

A novel plant leaf disease identification model based on a deep convolutional neural network was proposed by Geetharamani G and Arun Pandian J [13]. 39 different classes of plant leaves augmented images were used to train this deep Convolutional neural network model. The proposed model achieves better perfor- mance when using the validation data compared with popular transfer learning approaches. The proposed model achieves 96.46% classification accuracy. The proposed model is also tested with respect to its consistency and reliability.

Miaomiao Ji et al., [14] proposed a united convolutional neural networks (CNNs) architecture based on an integrated method for detecting grape leaf disease. UnitedModel is designed to distinguish leaves with common grape diseases i.e., black rot, esca and isariopsis leaf spot from healthy leaves. The combination of multiple CNNs enables the proposed UnitedModel to extract complementary discriminative features. The UnitedModel has been evaluated on the hold-out PlantVillage dataset and has been compared with several state-of-the-art CNN models. The UnitedModel achieves an average validation accuracy of 99.17% and a test accuracy of 98.57%, which can serve as a decision support tool to help farmers identify grape diseases.

Akash Sirohi and Dr. Arun Malik [15] introduced a hybrid model of deep learning to classify the sunflower diseases, i.e. Alternaria leaf blight, Downy mildew, Phoma blight, and Verticillium wilt. The combination of VGG-16 and MobileNet and stacking ensemble learning technique were used to make this hybrid model. The proposed model gave 89.2% accuracy, which is better than the other models.

A Convolutional Neural Network model was created by Mercelin Francis and C. Deisy [16] to perform plant disease detection and classification using apple and tomato leaf images of healthy and diseased plants. The model consists of four convolutional layers each followed by pooling layers. Two fully connected dense layers and sigmoid function is used to detect the probability of presence of disease or not. This model shows 87% accuracy and the overfitting problem is identified and removed setting the dropout value to 0.2. And also run on GPU Tesla to evaluate its speed of performance and accuracy.

The extraction of features based on texture and color features for one disease in sunflower named leaf rust was performed by D. Penghui et al [17]. They just experiment to extract features.

A. Rajbongshi et al [18] proposed a system for recognizing the disease of Rose applying the MobileNet model with transfer learning technique. They utilized four type of diseases in the rose and obtained 95.63% accuracy.

I. Gogul and V. S. Kumar [19] presented a technique for the recognition of flower species Inception-3 and Xception model with transfer learning technique. The feature extraction was performed by transfer learning and obtained accuracy was 90.60%.

S.Ramesh and D.vydeki [20] recognise and classify paddy leaf disease using the Optimized Deep Neural Network with Jaya Optimization Algorithm (DNN_JOA). They compared the optimized algorithm experiment result with ANN, DAE and DNN. The proposed method achieved high accuracy of 98.9% for the blast affected, 95.78% for the bacterial blight, 92% for the sheath rot, 94% for the brown spot and 90.57% for the normal leaf image.

A novel rice diseases identification method based on deep convolutional neural networks (CNNs) techniques is introduced by Yang Lu et al. [21]. CNNs are trained to identify 10 common rice diseases which are captured from the rice experimental field and 500 natural images of diseased and healthy rice leaves are used for this work. Under the 10-fold cross-validation strategy, the proposed CNNs-based model achieves an accuracy of 95.48%.

Mr. Pramod S. landge et al., [22] developed an image processing algorithms that can recognize problems in crops from images, based on colour, texture and shape to automatically detect diseases or other conditions that might affect crops and give the fast and accurate solutions to the farmer with the help of SMS. The design and implementation of these technologies will greatly aid in selective chemical application, reducing costs and thus leading to improved productivity, as well as improved produce.

Huiqun Hong et al,. [23] Deep Learning architecture is used with transfer learning which reduces the size of the training data, the time and the computational costs. Five deep network structures of Resnet50, Xception, MobileNet, ShuffleNet, Densenet121_Xception were applied in this research paper. Compared to the five convolutional neural networks, the parameters and the average accuracy are different. The best recognition accuracy of Densenet_Xception is 97.10%, but the parameters of Densenet_Xception are at most, the recognition accuracy of ShuffleNet is 83.68%, and the parameters are small.

The summary of past works is described in short in the table 2.1 which is shown in below:

Table 2.1: Summary of literature

Sl No.	Author, Year	Implemented Methods	Remark
[1]	V. Singh, 2019	Particle Swarm optimization algorithm	• Couldnot include any specific size of image. • Ovtained 92% accuracy
[2]	Jun Liu et al., 2019	Random Forest classifier	• k-means clustering and watershed algorithms for segmenting the diseases affected region • 95% accuracy was found
[3]	Sinan Uguz and Nase Uysal, 2021	VGG16 and VGG19 architectures, Adam, AdaGrad, Stochastic gradient descent (SGD) and RMS Prop optimization algorithms	• Augmented data produce 95% accuracy • Without data augmentation 88% accuracy • Adam and SGD produce superior result.
[4]	Sammy V. Militante et al., 2021	Deep learning techniques	• 13,842 sugarcane image dataset • Convolutional neural network shows 95% accuracy
[5]	Zhaohua Huang et al., 2020	VGG16, MobileNet, and AlexNet and an Ensemble Method	• Good prediction found • Ensemble method shows good prediction
[6]	Surampalli Ashok et al., 2020	Deep Neural Networks (AlexNet) and ANN technique	• System proposed by image processing, segmentation, clustering • proposed method found 98.18 % accuracy
[7]	S. Gayathri et al., 2020	CNN model LeNet	• 4 tea leaf disease was selected • Produce 85% to 95 % accuracy.

			• Which is better than traditional back propagation network.
[8]	T. Rumpf et.al., 2010	SVM model	• SVM shows 97% accuracy than Decision Tree and ANN. • SVM shows multiple classification accuracy 86%
[10]	Edna Chebet Too et.al, 2018	Comparison between VGG 16, Inception V4, ResNet with 50, 101 and 152 layers and DenseNets with 121 layers architecture	• 38 different class disease and healthy leaves of 14 plant is used. • DesNet shows 99.75% accuracy
[11]	Aditya Karlekar and Ayan Seal, 2020	SoyNet CNN model	• Proposed model found 98.14% accuracy • Model compared with VGG19, GoogleLeNet, Dense121, XceptionNet, LeNet, and ResNet50 based on state-of-the-art method. • Proposed method outperforms nine state-of-the-art methods
[12]	Geetharamani G and Arun Pandian J, 2019	Deep CNN model	• Proposed model achieves better performance than popular transfer learning approces. • Proposed model achieves 96.46% accuracy.
[13]	Miaomiao Ji et al., 2019	UnitedModel combination of multiple CNNs enables	• UnitedModel has been evaluated on the hold-out PlantVillage dataset and has been compared with several state-of-the-art CNN models • UnitedModel achieves an average validation

			accuracy of 99.17% and a test accuracy of 98.57%
[14]	Akash Sirohi and Dr. Arun Malik, 2021	VGG-16 and MobileNet hybrid model	• Alternaria leaf blight, Downy mildew, Phoma blight, and Verticillium wilt sunflower disease was used • Found 89.2% accuracy.
[15]	Mercelin Francis and C. Deisy	CNN model consist of four convolutional layers each followed by pooling layers and two fully connected dense layers with sigmoid function.	• 87% accuracy and the overfitting problem is identified and removed setting the dropout value to 0.2 • Used GPU Tesla to evaluate its speed of performance
[16]	D. Penghui et al., 2015	Feature extraction of leaf rust	• Extraction of features based on texture and color features for one disease in sunflower named leaf rust
[17]	A. Rajbongshi et al, 2020	A system with MobileNet and transfer learning technique	• Used 4 types of disease • Found 95.63% accuracy
[18]	I. Gogul and V. S. Kumar, 2017	A technique with Inception-3 and Xception model with transfer learning technique	• feature extraction was performed by transfer learning • Accuracy 90.60%
[19]	S.Ramesh and D.vydeki, 2020	Optimized Deep Neural Network with Jaya Optimization Algorithm (DNN_JOA)	• The proposed model compared with ANN, DAE and DNN. • Proposed method achieved high accuracy of 98.9% for the blast affected, 95.78% for the bacterial blight, 92% for the sheath rot, 94% for the brown spot and 90.57% for the normal leaf image.

[20]	Yang Lu et al., 2017	convolutional neural network (CNN)	• 500 natural images of diseased and healthy rice leaves was used for this work • Under the 10-fold cross-validation strategy, the proposed CNNs- model achieves an accuracy of 95.48%.
[21]	Mr. Pramod S. landge et al., 2013	Image processing algorithms	• Can recognize problems in crops from images, based on colour, texture and shape to automatically detect diseases or other conditions that might affect crops and give the fast and accurate solutions to the farmer with the help of SMS
[22]	Huiqun Hong et al., 2020	Resnet50, Xception, MobileNet, ShuffleNet, Densenet121_Xception	• Reduces the size of the training data, the time and the computational costs • Densenet_Xception is 97.10%, but the parameters of Densenet_Xception are at most, the recognition accuracy of ShuffleNet is 83.68%, and the parameters are small

2.3 Research Summary

The analysis of literature review had broadened the scope of sunflower disease prediction issues in the perspective of Bangladesh. The information and findings collected from this chapter is used as a guidance to develop the proposed systems. This paper works on sunflower disease detection using machine learning algorithms by extracting numerical data from real time images by using image feature extraction technique and deep learning algorithms, applying on the real time images then finally introducing the best learning algorithm on the basis of disease recognition accuracy and sensitivity of the algorithms. This chapter has also demonstrated the necessities of an efficient technique which is able to predict sunflower disease and find the effective factors.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter presents an outline of research methods that were followed in the study. It provides information on the workflow diagram, data collection and preprocessing, different techniques for prediction, training and testing, performance evaluation terminologies used in the study, feature selection. The requirements of this work and the procedures that were followed to carry out this study are included.

3.2: Strategy of Work Frame

The main focus of this project work is to derive numerical data from the real time sunflower disease and disease free images using GLCM image texture feature extraction method. And different machine learning algorithm such as Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, k-Nearest Neighbors, Adaboost Classifier, XGBclassifier have been used for detecting downy Mildew, gray mold, leaf Rust, leaf scars and disease free sunflower plant. And convolution neural network (CNN) models ResNet 50, VGG 19 and MobileNet V2 are directly applied to the image for predicting downy Mildew, gray mold, leaf scars and disease free sunflower. Finally find the best algorithm that accurately identifies these diseases and disease free sunflower plants by measuring accuracy and sensitivity. The following figure 3.1 represents the working procedure of the research work.

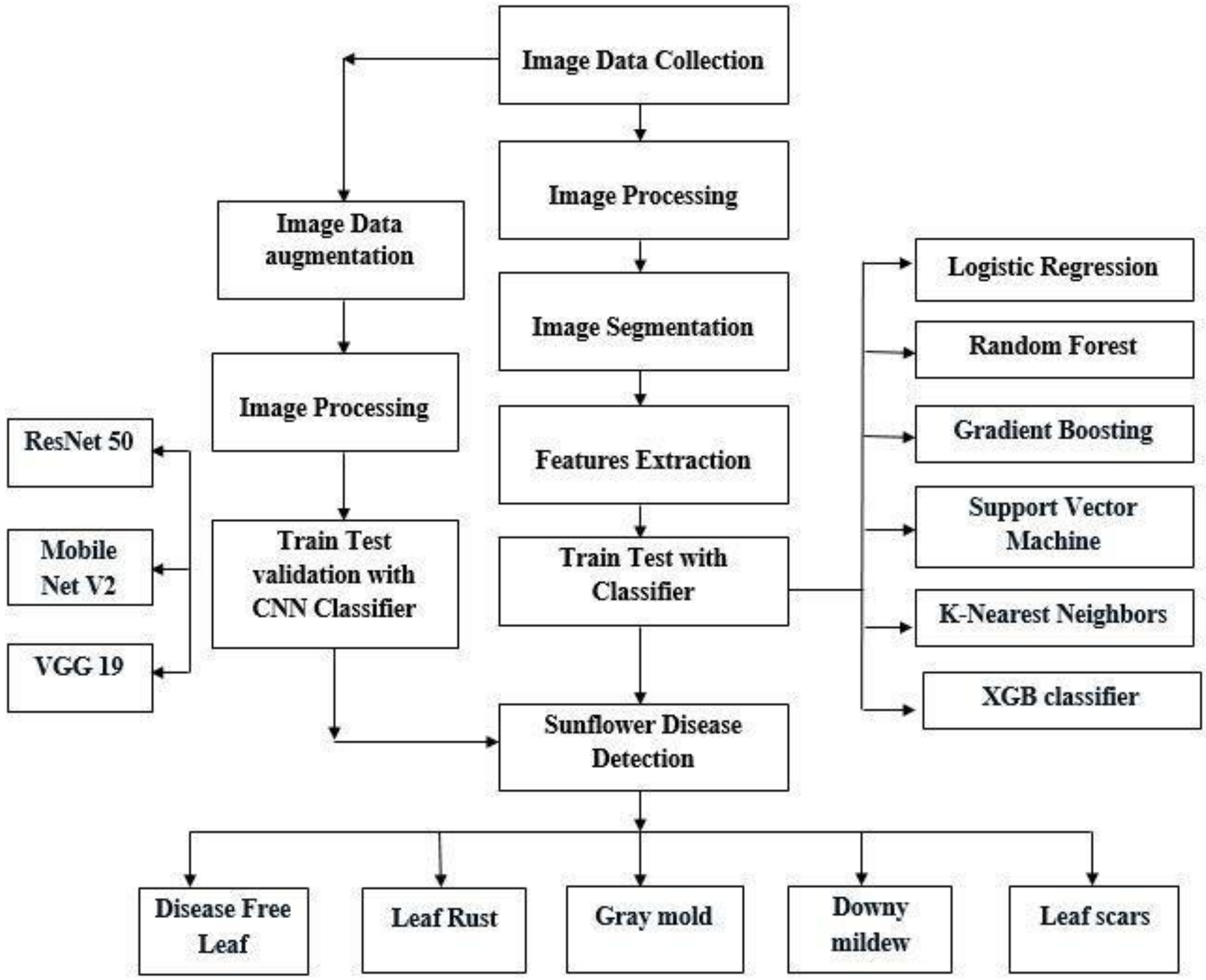


Figure 3.1: Block Diagram of the Overall Work

3.3 Image Collection and Processing

3.2.1 Image Collection

Image collection is the primary step for any computer vision system. This process is retrieving an image from any source. The quality and quantity of the images must be satisfactory for the research work and enough for the work. Otherwise the expected outcome may not be achieved. In this paper 5 types of image data of sunflowers were collected from the demonstration farm of Bangladesh Agricultural Research Institute (BARI) at Gazipur in cooperation with its one domain expert when the sunflower plants were about to bloom and the maximum diseases can be found.. These 5 types include 4 types of sunflower disease named Gray mold, Downy mildew, Leaf Scars, and Leaf Rust and the other type is disease free sunflower leaf or fresh leaf. The figure 3.2 shows the disease sunflower and disease free sunflower leaves. For this research work 100 gray mold, 115 Downy mildew, 86 Leaf Rust, 142 leaf scars and 159 disease free sunflower and sunflower leaves were used. And 1500 augmented image of gray mold, downy mildew, leaf scars and disease free types is for deep learning CNN models.

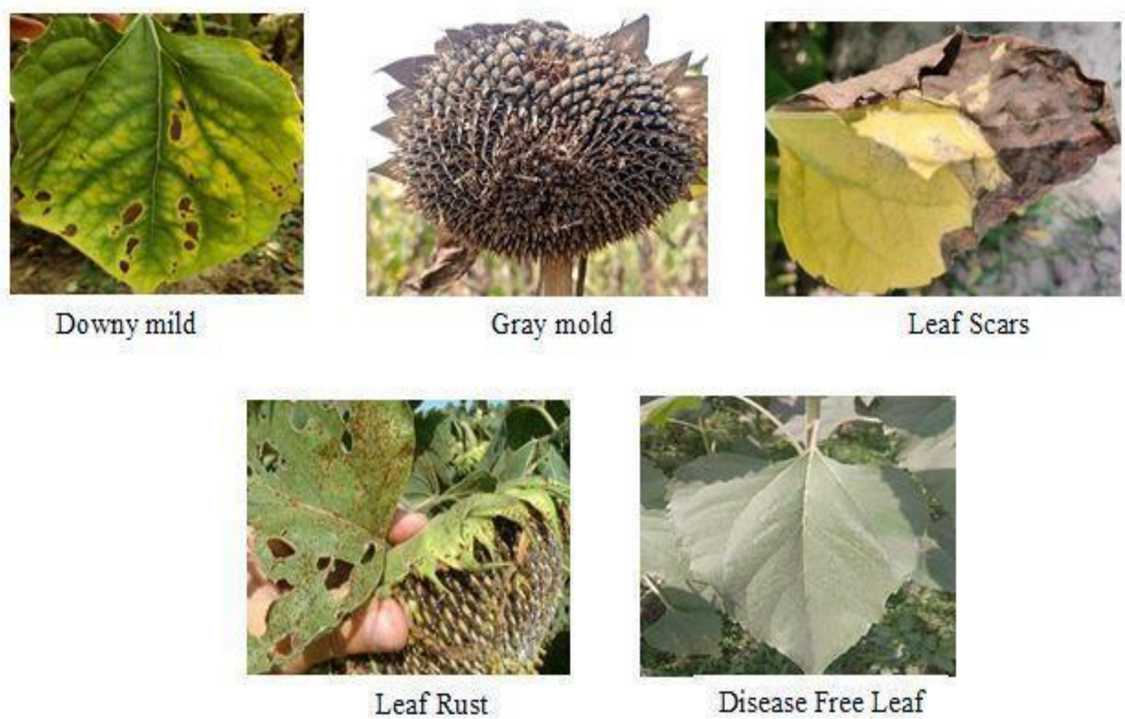


Figure 3.2: Sunflower diseases with disease free Sunflower

3.2.2 Data Processing

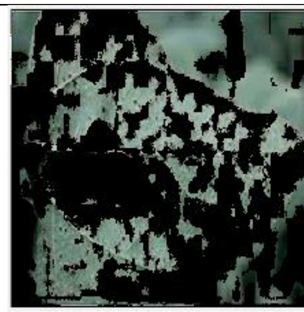
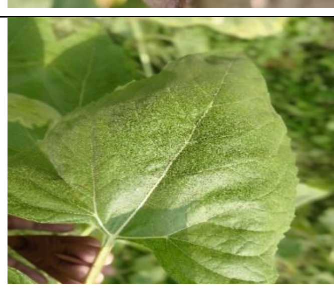
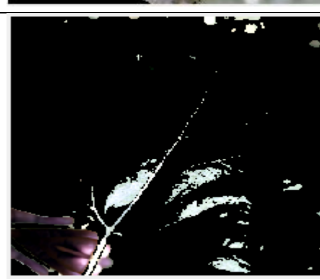
For this research work some pre-processing and enhancement operations are needed to obtain a good result. Sunflower leaf disease image acquisition under natural conditions, in the process of collection and transmission, and may be affected by external factors that cause image blur, degraded, affecting the accuracy of the subsequent identification of sunflower diseases. In this work images are pre-processed into two steps including image resizing, image contrast enhancement. Generally, the faster training of the machine learning models is possible by smaller image. For that image resizing is important. The contrast enhancement technique improves the information carried out.

i. Image segmentation

Image segmentation is a pre-processing task in the area of image analysis, computer vision, and pattern recognition. It is the process of dividing an image into various segments. Image segmentation is commonly used to determine objects and boundaries (lines, curves, etc.) in images. The purpose of image segmentation is to transform an image representation into something more meaningful and simpler to interpret. Sometimes the whole image is not important for work because some part of an image may not carry information. In that case image segmentation helps to find out the important parts of an image for further processing.

Image segmentation methods fall into different categories: Region based segmentation, Edge based segmentation, and Clustering based segmentation, Thresholding, artificial neural network, feature-based segmentation. Clustering of an image is one of the good techniques, which is used for segmentation of images. After extraction of features, these features are put together into well-separated clusters based on each class of an image. The clustering algorithm aim is to develop the partitioning decisions based on the initial set of clusters that is updated after each iteration. In this paper k means unsupervised clustering method is used to perform the segmentation on sunflower images. The K-Means clustering technique is a well-known approach that has been applied to solve low-level image segmentation tasks. This clustering algorithm is convergent and its aim is to optimize the partitioning decisions based on a user-defined initial set of clusters that is updated after each iteration clusters. In this work, k=3 is used for segmenting an image which implies that it will distinguish 3 clusters in the image. Thus, the part of the object is selected from the background of the image. Great quality of segmentation always guides to extract great quality of feature extraction. The feature extraction procedure from the original image is presented in Table 3.1.

Table 3.1: Feature extraction Procedure from original image

Class Name	Original Image	Segmented Image	Extracted Value
Gray Mold			1.6435, 0.8052, 0.4020, 0.8929, 44.4398, 70.9354, 3.5067, 9.04, 6299.4835, 3.3548, 1.3203
Downy Mildew			0.9177,0.7423, 0.3319, 0.8918, 30.5238, 46.6168, 3.8846, 9.5388, 2633.8259, 3.8914, 1.3815
Leaf Rust			1.0600, 0.8405, 0.3245, 0.8914, 49.9113, 63.8092, 3.9721, 10.223, 2786.8199, 1.8212, 0.7028
Leaf Scars			0.4235, 0.9437, 0.6258, 0.9738, 28.8488, 63.799, 2.2975, 5.9921, 1931.3405, 5.9752, 2.0885
Disease Free			8.8519, 0.8063, 0.8345, 0.9732, 13.6389, 51.5177, 1.1270, 4.2376, 3940.2797, 16.4441, 3.8570

ii. Feature Extraction

Feature extraction is to extract important information from the image for subsequent image recognition. Appropriate characteristic parameters are particularly important, which are the important basis of disease identification. The common disease features are color feature, texture feature and shape feature. In this research work texture features are used to identify the diseases. The statistical method and the structure method are two main methods to describe texture characterization of an image [17]. Texture features exist in the image, which are the attributes of the image itself, the internal organization structure and the description of the internal pixels. In this paper, the gray level co-occurrence matrix (GLCM) method is selected to study the texture characteristics of the disease image. Gray level co-occurrence matrix (GLCM) is a popular texture-based feature extraction method. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image. The GLCM determines the frequency of combinations of the pixel brightness values, that it represents the frequency formation of the pixel pairs [23]. The GLCM properties of an image are expressed as a matrix with the same number of rows and columns as the gray values in the image. The elements of this matrix depend on the frequency of the two specified pixels. Both pixel pairs can vary depending on their neighborhood. These matrix elements contain the second-order statistical probability values depending on the gray value of the rows and columns. The gray level co-occurrence matrix is calculated when the distance $d = 1$ and the angle are 0, 45, 90 and 135 respectively. Textures features that was calculated for this research work from GLCM matrices are homogeneity (H), contrast (C), correlation (ρ), energy (E), entropy (S) and statical features Mean(μ), Standard deviation (σ), Variance(σ^2), Kurtosis (κ), Skewness (γ). Let the co-occurrence matrix be $P_{i,j}$ and the size of the matrix is $N \times N$. Each element (i,j) represents the frequency by which pixel with gray level i is spatially related to pixel with gray level j. The number of rows and columns is equal to the number of quantized gray levels, Ng, in the image. The matrix element $p(i, j)$ is the set of second order statistical probability values for changes between gray level i and j at a particular displacement distance (d) and angle (θ).

(a) Contrast

Contrast, considering the local homogeneity of the image. It is a measure of linear dependencies in the image. A high correlation indicates a high number of linear structures in the image. .Digital Image Correlation is an optical method that employs tracking & image registration techniques for accurate 2D and 3D measurements of changes in images. This is often used to measure deformation, displacement, strain and optical flow, but it is widely applied in many areas of science and engineering. Correlation is computed as

$$C = \sum_i \sum_j |i - j|^2 \, p(i,j) \tag{3.1}$$

(b) Correlation

Correlation measures the linear dependency of grey levels of neighboring pixels. Digital Image Correlation is an optical method that employs tracking & image registration techniques for accurate 2D and 3D measurements of changes in images. And the correlation is computed as:

$$\rho = \frac{\sum_{i,j=0}^{Ng-1} (i-\mu_i)(j-\mu_j) p(i,j)}{\sigma_i \sigma_j} \tag{3.2}$$

(c) Energy

The energy function measures the smoothness of an image. It provides the sum of squared elements in the GLCM. The homogeneous scene contains only a few gray levels. With highly correlated pixels, the energy value is very large. The energy feature of an image 'E' is defined in equation (3) as follows

$$E = \sum_{i,j} p(i,j)^2 \tag{3.3}$$

(d) Homogeneity

The homogeneity is commonly referred to as inverse different movement. It measures image homogeneity that takes high values for low contrast images. Homogeneity decreases as contrast increases with constant energy. The homogeneity feature of an image 'H' is defined in equation (4) as follows:

$$H = \sum_i \sum_j \frac{1}{1+ |i-j|^2} p(i,j) \tag{3.4}$$

(e) Entropy

The entropy feature measures the degree of uniformity between pixel within the image and randomness is used to characterize the texture of the image. Entropy is strongly but inversely correlated to energy as images with larger gray levels has high entropy. The entropy function of an image 'E' is defined in equation (5) as follows:

$$S= \sum_{i=0}^{N_g-1} \sum_{j-0}^{N_g-1} p(i,j) log\ (p(i,j)) \tag{3.5}$$

(f) Mean

The mean equation will computed from the GLCM matrices as follows:

$$\mu_i = \frac{\sum_{i,j=0}^{N_g-1} i\ (p_{i,j})}{} \tag{3.6}$$

$$\mu_j = \sum_{i,j=0}^{N_g-1} i\ (p_{i,j}) \tag{3.7}$$

(g) Variance

The variance equation will computed from the GLCM matrices as follows:

$$\sigma_{\boldsymbol{i}}^{\mathbf{2}}= \sum_{\mathrm{i,j=0}}^{\mathrm{N-1}} \mathrm{p_{i,j}}(\mathrm{i} - \mathrm{\mu_i})^2 \tag{3.8}$$

$$\sigma_{\boldsymbol{j}}^{\mathbf{2}}= \sum_{\mathrm{i,j=0}}^{\mathrm{N-1}} \mathrm{p_{i,j}}(\mathrm{j} - \mathrm{\mu_j})^2 \tag{3.9}$$

(h) Standard Deviation

The standard equation equation will computed from the GLCM matrices as follows:

$$\sigma_i = \sqrt{\sigma_i^2} \qquad \sigma_j = \sqrt{\sigma_j^2} \tag{3.10}$$

(i) Kurtosis

It Measures the stability of the distribution, which relates to the normal distribution

$$\kappa = \sum_{i,j=0}^{N_g-1} \frac{(p_{i,j})^4}{N_g\,\sigma^4} \tag{3.11}$$

(j) Skewness

Skewness can be measured as:

$$\gamma = \frac{1}{\sigma^3} \sum_N \big((i,j) - \mu\big)^3\, p(i,j) \tag{3.12}$$

iii) Normalization

Before training, testing by the classifiers, the normalization of the data set is needed. Normalization is the process of organizing a database to reduce redundancy and improve data integrity. Normalization (or scaling) is one of the main parts for learning process because real data obtained from experiments and analysis most times are distant from each other. If data is not normalize, the importance of each inputs between (0,1) or (-1,1) are not equally distributed by thus naturally large values become dominant according to less values during classifiers training. The effect is great because the common activation functions such as sigmoid, hyperbolic tangent and Gaussian produce result that ranges between [0,1] or [-1,1]. It is important to normalize the values to be in that range. Moreover the dissimilar rage of the inputs can distort the neural network result.

The common normalization approach includes Statistical normalization (using mean and standard deviation) and Min-Max Normalization. Therefore the experimental data are normalized according to Equation in order to train the neural network effectively. Normalization confirms that the neural network was trained effectively without any particular variable significantly distorting the result.

$$X_N = 2\,\frac{x - x_{min}}{x_{max} - x_{min}} - 1 \tag{3.13}$$

Where, X_N is the normalized value of the real variable, x is the measured value of the variable and x_{min} is the minimum value of the real variable, x_{max} is the maximum value of the real variable.

3.3 Distinct Techniques for Disease Prediction

Different machine learning and deep learning CNN networks techniques was used for predicting sunflower disease. Logistic regression, random forest, gradient boosting, support vector machine, k-nearest neighbor, xgboost techniques is used for identifying the sunflower disease by using extracting features. And CNN network such as resnet 50, mobiloe net v2, vgg 19 used direct image for predicting the sunflower disease. Different supervised and unsupervised learning techniques are used in this research work.

3.3.1 Machine Learning Models

i) Logistic Regression

Logistic regression is a supervised machine learning algorithm developed for learning classification problems. Logistic regression is a member of generalized linear models. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The sigmoid equation and logistic function is

$$f(x) = \frac{1}{1+e^{-x}} \tag{3.14}$$

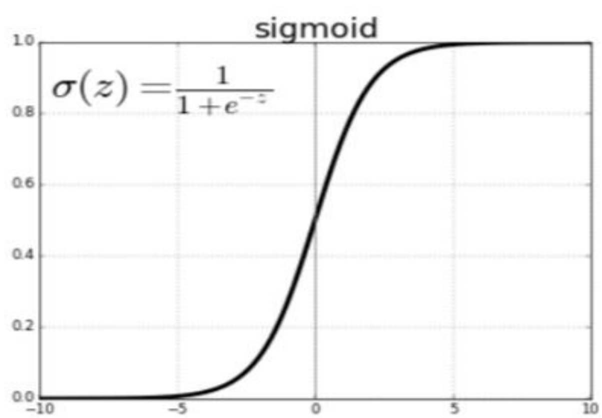


Figure 3.3: Logistic function

ii) Random Forest

Random forests are a combination machine learning algorithm. Which are combined with a series of tree classifiers, each tree casts a unit vote for the most popular class, then combining these results get the final sort result. A random forest is a classifier consisting of a collection of tree–structured classifiers {h (x, Θ_k), k=1,...} where the {Θ_k} are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x . The algorithm improves the classification performance by combining multiple decision trees to generate a classifier, and determines the prediction results by multiple backdrops to achieve the purpose of classification. As a statistical learning theory, a random forest has a good tolerance for sample classification and regression [23]. In Breiman's RF model, every tree is planted on the basis of a training sample set and a random variable, the random variable corresponding to the kth tree is denoted as Θ_k , between any two of these random variables are independent and identically distributed, resulting in a classifier h(x, Θ_k), where x is the input vector. After k times running, we obtain classifiers sequence {$h_1(x)$, $h_2(x)$,…. $h_k(x)$}, and use these to constitute more than one classification model system, the final result of this system is drown by ordinary majority vote, the decision function is and use these to constitute more than one classification model system, the final result of this system is drown by ordinary majority vote, the decision function is

$$H(x) = \arg \sum_{i=1}^{k} I\ (h_i(x) = Y) \tag{3.15}$$

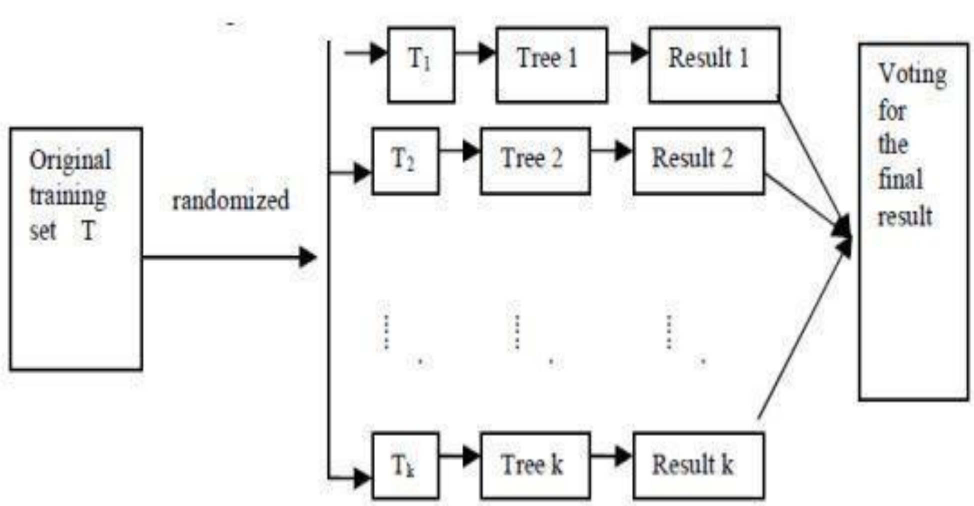


Figure 3.4: Random forest schematic

Here H(x) is combination of classification model, h_i is a single decision tree model, Y is the output variable, I(-) is the indicator function. For a given input variable, each tree has the right to vote to select the best classification result. Random forest voting process is shown in figure below.

iii) Gradient Boosting

Gradient Boosting (GB) is another efficient method for solving non-linear classification and regression problems. It discards all weaker predictors and picks the stronger one. It is an improved version of the decision tree where every successor comparatively analyzed to build a set of the optimally satisfying structure of the tree by using the structure score, gain calculation, and increasingly refined approximations. The Gradient Boosting Classifier depends on a loss function. A custom loss function can be used, and many standardized loss functions are supported by gradient boosting classifiers, but the loss function has to be differentiable.

iv) Support Vector Machine

The support vector machine (SVM) is a universal constructive learning procedure based on the statistical learning theory. The Support Vector Machine (SVM) is implemented using the kernel Adatron algorithm. The kernel Adatron maps inputs to a high-dimensional feature space, and then optimally separates data into their respective classes by isolating those inputs which fall close to the data boundaries. Therefore, the kernel Adatron is especially effective in separating sets of data which share complex boundaries. SVMs can only be used for classification, not for function approximation.

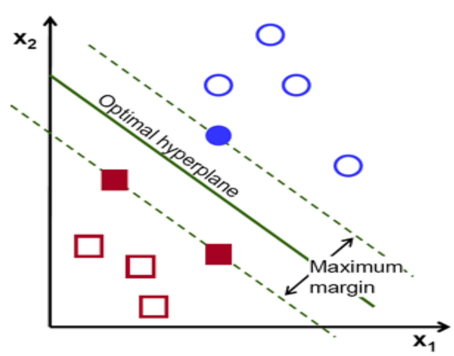


Figure 3.5: Support Vector Machine hyper plane

The Adatron is an online algorithm for learning perceptrons which has an attractive fixed point corresponding to the maximal-margin consistent hyper-plane when this exists. Simple example of the SVM algorithm using the hyper plane shown below.

v) K-Nearest Neighbors (KNN)

KNN is a simple model and one of the most widely used machine learning processes for classification problems, pattern recognition, and regression. KNN gets neighbors among data using Euclidean distance between points of data. This algorithm is used for classification and regression problems. The value of k (where k is a constant defined by the user) will find all of the similar existing feature cases with the new case and surround all cases to find the new case for a similar category. This algorithm involves finding the k nearest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the found data points to it. Bellow figure represents the basic idea of the KNN algorithm for classifying new data point when the value of k equal to 3 and 5. Solid border consist of 2 triangle and 1 square class and dotted circle consist of 3 square and 2 triangle. Using the Euclidean distance, the algorithm will classify the new point green dotted circle among this 2 classes.

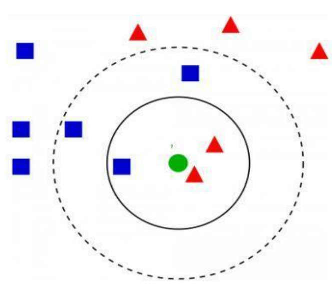


Figure 3.6: K=3 and K=5 data point

vi) Extreme Gradient Boosting Technique

XGBoost classifier is a gradient boosting method incorporating the regression tree. It uses the combination of weak learners to create a single strong learner. The idea is to continuously add trees and continuously perform feature splitting to grow a tree. The implementation of XGBoost offers several advanced features for model tuning, computing environments and algorithm enhancement.

$$y_i = \Phi(x_i) = \sum_{k=1}^{T} f_k(x_i),\ \ f_k \tag{3.16}$$

in which is the predicted label of ith sample, represents the ith sample, T represents the total number of trees, represents the kth tree model.

The objective function of XGBoost classifier is defined as

$$L(\Phi) = \sum_{i=1}^{n} l(y_i, y_i) + \sum_{k=1} \Omega(f_k) \tag{3.17}$$

There are two terms in the objective function. The first term is the loss function measuring the difference between the predicted value and the real value. The second term is the regularization term.

The model summary used in this research work of logistic regression, random forest, support vector machine, gradient boosting, K-Nearest Neighbors and xgboost classifier is shown in table 3.2.

Table 3.2: Machine Learning Classifier Model Summary

Classifier Name	Model Summary
Logistic Regression	Model: Multinomial Logistic function: logit(p) = ln(p) - ln(1-p).
Random Forest	n_estimators= 5 randomly chosen value= 0
Gradient Boosting	n_estimators=30 randomly chosen value= 42
Supporting Vector Machine	Karnel= ‘rbf’
K-Nearest Neighbors	algorithm="brute" n_neighbors=15
XGB classifier	objective ='multi:softprob' max_depth = 7

3.3.2 Convolutional Neural Network Model

Convolutional neural network CNN is a special type of ANN that is customized for image processing. Deep learning is a class of machine learning algorithms. Researchers have shown the capability of CNN to give high accuracy in image processing tasks. In conventional ANN, a single neuron in a layer takes its input from all neurons in the previous layer. The deep neural networks are composed with n number of layers with extraneous features. A deep CNN can extract higher level features progressively from the input images by the multiple layers used in a model. CNN general model shown in figure 3.7

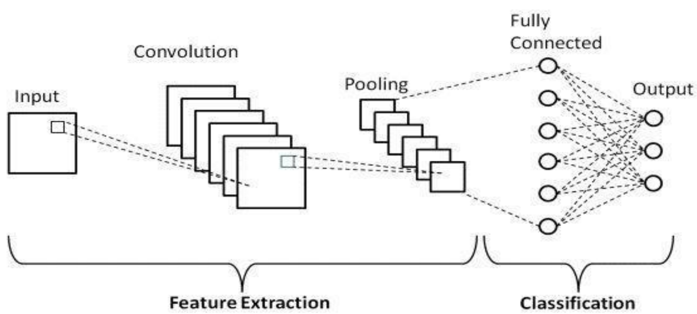


Figure 3.7: CNN general model

The CNN model shows input image, convolution layer, pooling layer, fully connected layers, activation function. The usages of these components vary depending on the network architecture. Figure (3.8) shows several phases in sunflower disease detection using a basic architecture of convolutional neural network. The model implementation is detailed in the proposed models.

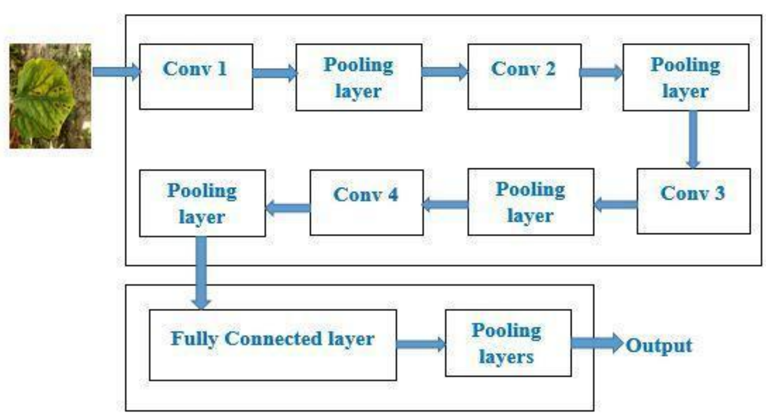


Figure 3.8: CNN several phases for disease detection

Convolution layer

The layer receives as input a RGB image or an output of another layer for further processing. Convolutional layers store the output of the kernels from the previous layer which consists of weights and biases to be learned [24]. The inputted image is abstracted to a feature map, also known as an activation map, after passing through a convolutional layer. The input is convolved by convolutional layers, which then pass the output to the next layer. This is analogous to a neuron's response to a single stimulus in the visual cortex.

Feature Map = Input Image x Feature Detector

Pooling Layer

Pooling layers reduces overfitting and lowers the neuron size for the down sampling layer [6]. A 2x2 filter size is used to the pooling layer to give an output through pooling type. The pooling layer reduces the size of the feature map and number of parameters, the training time, computation rate and controls overfitting [25]. Overfitting is a situation where the model achieves 100% or 99% on the training set but an average of 50% on the test data. To reduce the dimension feature map, the use of max pooling integrated with stride and ReLU are applied [26]. If there is no pooling, the output has the same resolution as the input. Hence the pooling layer reduces the size of the feature map [25], i.e., length and width is reduced but not the depth. This reduces the number of parameters and weights, lessens the training time and also reduces the cost of computation. There are 2 methods of pooling, one is max pooling which chooses the most significant element from the feature map. And the 2nd one is average pooling it chooses the average for each region of the feature map.

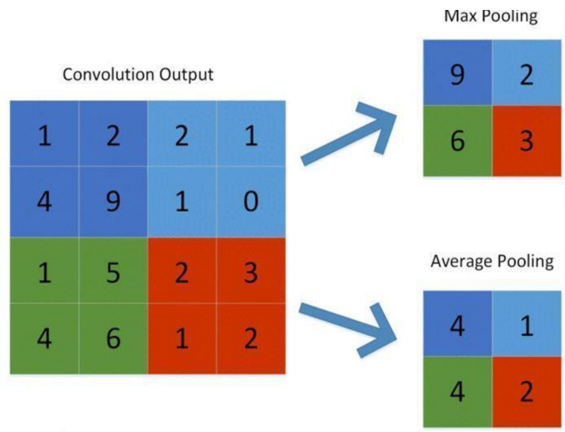


Figure 3.9: The average and max pooling

From this figure 3.9 the max pooling illustrates the max feature map and average pooling shows the average feature map.

Activation Layer

In this layer, a non-linear ReLU (Rectified Linear Unit) is applied constantly in each convolution layer.

Fully Connected Layers

Fully connected layers are used to analyze the class probabilities and the result is the input of the classifier. Softmax classifier is the well-known input classifier and recognition and classification of sugarcane diseases are applied in this layer.

(a) Residual Network 50

Residual network (ResNet) is a CNN architecture whose core building element is a residual block [27]. Figure 3.10 shows a residual block. It shows that They explicitly let the layers fit a residual mapping and denoted that as H(x) and they let the nonlinear layers fit another mapping F(x):=H(x)−x so the original mapping becomes H(x):=F(x)+x.

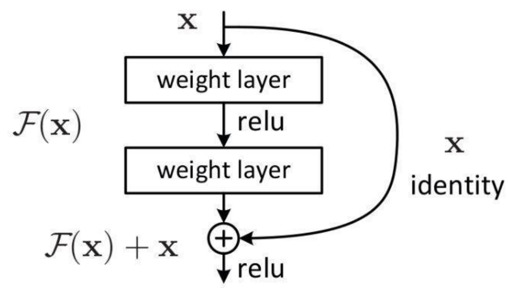


Figure 3.10: Basic ResNet residual block

A residual block makes the use of skip connections to address the degradation problem. Connections that skip one or more layers are shortcut connections. ResNet-50 is a 50-layer convolutional neural network, 48 convolutional layers, one MaxPool layer, and one average pool layer. The fundamental breakthrough with ResNet was it allowed us to train extremely deep neural networks with 150+layers. The ResNet50 takes the input as 256 dimensions, and the main advantage of ResNet50 architecture is the incorporation of skip connections, which helps solve the vanishing gradient descent problem.

(b) MobileNet Version 2

MobileNet is a CNN-based model that is extensively used to classify images. The main advantage of using the MobileNet architecture is that the model needs comparatively less computational effort than the conventional CNN model that makes it suitable for working over mobile devices and the computers that work over lower computational capabilities [28]. MobileNet architecture is equally efficient with a minimum number of features, such as Palmprint Recognition [29]. The architecture of MobileNet is depth-wise [30]. The fundamental structure is based on different abstraction layers, a component of different convolutions that appear to be the quantized configuration that assesses a regular problem complexity in-depth. The complexity of 1×1 is called point wise complexity. Platforms made in-depth are designed to have abstraction layers with structures in-depth and point through a standard, rectified linear unit (ReLU). The principle concept of the Mobile Net architecture is to substitute complected convolutional layer of size 3 × 3 that buffers the input data, accompanied by a convolutional layer of size 1 × 1 point wise that incorporates these filtered parameters to build a new component as shown in Figure 3.11 [31].

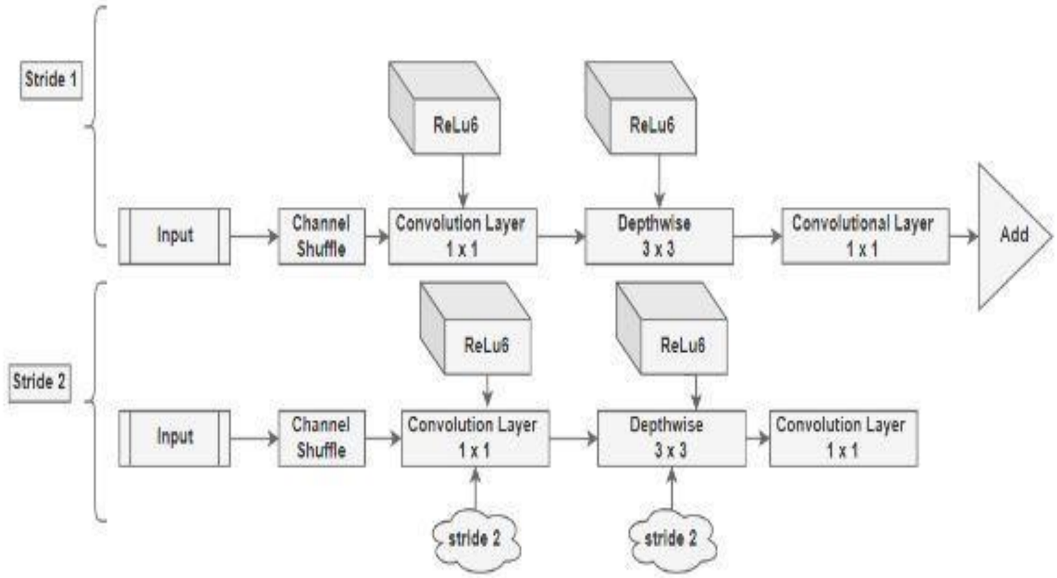


Figure 3.11: Basic Structure of Mobile Net V2

(c) Visual Geometry Group-19 Layers

VGG 19 is derived from VGG architecture, consisting of different layers such as convolution layers- 16, Fully connected layer-3, MaxPool layers-5, and SoftMax layer-1.The abbreviation of VGG is the Visual Geometry Group who created this network. It was designed to recognize large scale visual recognition. The main advantage of this algorithm is its code is open source, and we can easily apply transfer learning and make the network work for other architectures. The algorithm also learns collective small-sized kernel rather than one big kernel because the inclusion of more small-sized kernels makes the system learn complex features. This architecture was in the top 5 scoreboards. The reason for using this architecture in plant disease classification is its robustness to understand the complex features. The overall architecture of VGG 19 is shown in Figure 3.11.

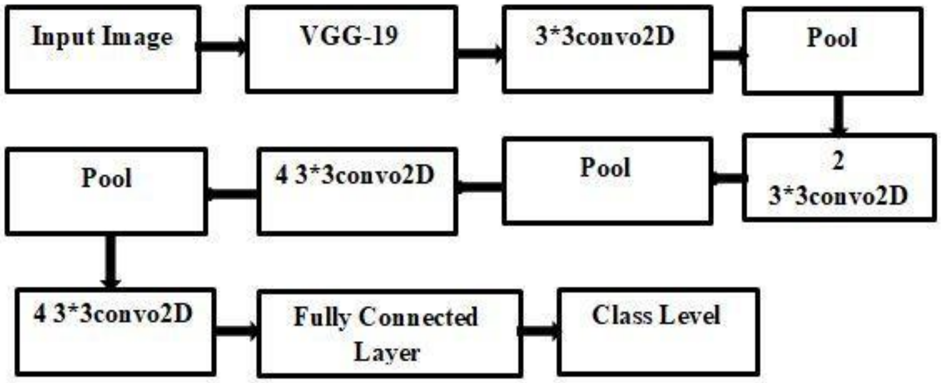


Figure 3.11: VGG19 disease detection steps

ResNet 50, VGG19 and Mobile Net V2 model summary used in this research work is shown in Table 3.3.

Table 3.3: CNN classifier model summary

CNN Classifier Name	Model Summary
ResNet 50	Danse Layer (unit: 128 Activation function: Rule) Epoc: 10
Mobile Net V2	Danse Layer (Activation function: Rule) Epoc: 10
VGG19	Danse Layer (units=1024, activation='relu') (units=512, activation='relu') Epoc: 10

3.4 Training and Testing

Once a network has been structured for a particular study, that network will be trained. After the appropriate training, the selected network has the ability to interconnect one value of output to a given input. For starting this process the initial weights are chosen randomly and then, the training or learning begins. For developing an appropriate network the networks will be trained many times. There are two approaches to training - supervised and unsupervised. Supervised training involves a mechanism of providing the network with the desired output either by manually grading the network's performance or by providing the desired outputs with the inputs.

Unsupervised training is where the network has to make sense of the inputs without outside help. Unsupervised training is used to perform some initial characterization on inputs. In supervised training, both the inputs and the outputs are provided. The network then processes the inputs and compares its resulting outputs against the desired outputs. Errors are then propagated back through the system, causing the system to adjust the weights which control the network. This process occurs over and over as the weights are continually tweaked. The set of data which enables the training is called the training set. The other type of training is called unsupervised training. In unsupervised training, the network is provided with inputs but not with desired outputs. The system itself must then decide what features it will use to group the input data. This is often referred to as self-organization or adaptation.

Testing of the data is used to confirm the expected result, i.e. when test data is entered the expected result should come and some test data is used to verify the software behavior to invalid input data. Test data is generated by testers or by automation tools which support testing. Data validation is intended to provide certain well-defined guarantees for fitness, accuracy, and consistency for any of various kinds of user input into an application or automated system. For machine learning techniques the 602 data features extracted from image are divided as training data (80%) and testing data (20%). Here trained each using 80.00% of the training data. Then the techniques are tested on the 20.00% of the testing dataset, and 5*5 confusion matrices are generated for each classifier. For performing CNN techniques the 1500 image data set is divided into 70% for training, 20% for testing and 10% for validation. Then the CNN models are trained with the 70% of training data. Then testing the model by 20 % testing data set. And finally, the model are validated using 10% validation data set.

3.5 Performance Evaluation Terminologies

For the performance measurement of the applied classifier, we have considered several performance metrics named accuracy, specificity, sensitivity, precision [34], false-positive rate (FPR), and false-negative rate (FNR).

3.5.1 Confusion Matrix

The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for Classification problems where the output can be of two or more types of classes. A confusion matrix is nothing but a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)" shown below.

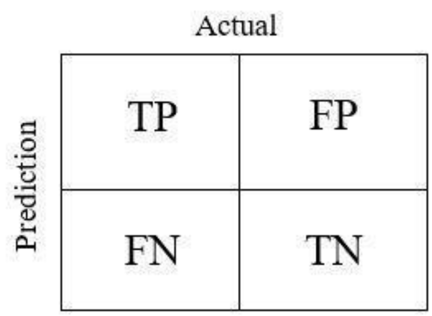


Figure 3.11: Confusion matrices

True Positives (TP): True positives are the cases when the actual class of the data point was true and the predicted is also true.

True Negatives (TN): True negatives are the cases when the actual class of the data point was False and the predicted is also false.

False Positives (FP): False positives are the cases when the actual class of the data point was false and the predicted is true. False is because the model has predicted incorrectly and positive because the class predicted was a positive one.

False Negatives (FN): False negatives are the cases when the actual class of the data point was true and the predicted is false. False is because the model has predicted incorrectly and negative because the class predicted was a negative one

3.5.2 Accuracy

It is the most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all predictions made.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \tag{3.19}$$

3.5.3 Precision

It is used in document retrievals, and may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula

$$\text{precision} = \frac{T\,P}{TP+FP} \tag{3.20}$$

3.5.4 Sensitivity

A sensitivity is essentially the ratio of true positives to all the positives in ground truth.

$$\text{sensitivity} = \frac{T\,P}{TP+FN} \tag{3.21}$$

3.5.5 Specificity

In contrast to recall or sensitivity, may be defined as the number of negatives returned by our ML model.

$$\text{Specificity} = \frac{T\,N}{N} \tag{3.22}$$

3.5.6 False Positive Rate (FPR)

The FPR is the percentage of negative cases in the data that were mistakenly reported as positive. The total number of negative cases wrongly reported as positive cases divided by the total number of negative cases (i.e. normal data) is described in this formula:

$$\text{FPR} = \frac{FP}{N} \tag{3.23}$$

3.5.7 False Negative Rate (FPR)

FNR: $\frac{FN}{N}$ (3.24)

Where, P = Positive class, N = Negative Class, TP = True positive, TN = True Negative, FP = False Positive, FN = False Negative.

3.5.8 F1-Score

It is used to measure test accuracy. It is a weighted average of the precision and recall. When F1 score is 1 it's best and on 0 its worst.

F1= $\frac{2*(Precision*recall)}{precision+recall}$ (3.25)

3.5.9 AUC – ROC curve

AUC - ROC is a graphical representation for performance measurement for classification problem. ROC is a probability curve and AUC represents degree or measure of area. It expresses how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease. The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

3.6 Extracting Feature Selection

The feature selection aims to identify the most significant features for predicting a diseases. Furthermore, feature selection helps to construct a more accurate model by eliminating or underrepresenting the less relevant features, minimizing training time and enhancing learning performance. Feature selection techniques are categorized as filter based, wrapper based, and embedded type. Filter-based techniques screen out features based on some specified criteria. Wrapper-based methods use a modeling algorithm that is taken as a black box to evaluate and rank features. The embedded methods have built-in feature selection approaches such as least absolute shrinkage and selection operator (Lasso) and random forest (RF) feature selection methods [35]. In this research paper ANOVA test is used for feature selection.

i. Analysis of Variance

Analysis of variance (ANOVA) is a well-known statistical method to determine whether there is a difference in means between two groups [36]. In this study, the ANOVA test was utilized to select the significant numerical features in predicting the occurrence of T2D. The ANOVA test uses the F statistic for feature ranking. The larger the value of the F statistic, the better the discriminative capacity of the feature [37]. The F value is calculated as:

$$F = \frac{\frac{SSB}{df_b}}{\frac{SSW}{df_w}} \tag{3.26}$$

Where SSB (sum of squares between groups) is the variation of group means from the total grand mean, and SSW (sum of squares within groups) is the sum of squared deviations from the group means and each observation. The degrees of freedom for mean square between and within is defined by df_b and df_w, respectively [38]. For all numerical features in the dataset, the F value was calculated using Equation (18) and the features with the larger value were selected.

3.7 Summary

This methodology chapter has discussed about the image collection, image processing and how to extract image data and pattern classification algorithms and basic terminologies. Performance analysis terminologies such as confusion matrix, precision, sensitivity etc. and feature selection technique also have included in this chapter

CHAPTER 4

IMPLEMENTATION AND ANALYSIS

4.1 Introduction

In this chapter, the experiment result of different machine learning and deep learning are discussed. Result produces from all the selected algorithms are compared with result from the evaluation process. Their performance was evaluated and selects a final model with maximum accuracy, sensitivity, specificity, f1 score.

4.2 Data Analysis

To evaluate the applied models, this paper took 602 coloring images of sunflowers and sunflowers' leaves. The images are obtained with different sizes and different angles that are used for the experiment. And 1500 augmented colored images of sunflowers and sunflowers leaves are used for deep learning classification algorithms.

The research paper works to predict the disease of sunflowers by machine learning algorithms and the disease of sunflowers is also predicted by deep learning algorithms. And finally identify the best algorithm among machine learning algorithms and deep learning algorithms.

4.3 Training Testing Results of Machine Learning and Conventional Neural Network Techniques

For this research paper Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, k-Nearest Neighbors, XGBoost classifier are used for classifying the disease of sunflower such as downy mildew, gray mold, leaf Rust, leaf scars and disease free sunflower.

4.3.1 Logistic Regression Result

The logistic algorithm shows 87.9% training accuracy and 83.47% testing accuracy for predicting the disease of sunflower and sunflower leaves.

Table 4.1: Logistic regression result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Gray mold	0.89	0.69	0.78	0.98	0.02	0.03
Downy mildew	0.74	0.92	0.82	0.91	0.08	0.08
Leaf Rust	1.00	0.50	0.67	1.00	0.0	0.5
Leaf scars	0.77	0.89	0.83	0.92	0.07	0.1
Disease-free	0.91	1.00	0.95	0.96	0.03	0.00

The overall experiment result analysis of logistic regression for detecting separate disease categories is shown below in the Table 4.1. The classifier has achieved the highest accuracy, sensitivity, precision, f1-score and the lowest FNR that are respectively 0.91, 1.00, 0.95 and 0.00 for disease free category. 0.98 highest specificity for gray mold and lowest FNR 0.0 for leaf rust. The confusion matrix of this algorithm shown in the below figure 4.1. Here we can see the separate diseases, how many are correctly detected and wrong detected.

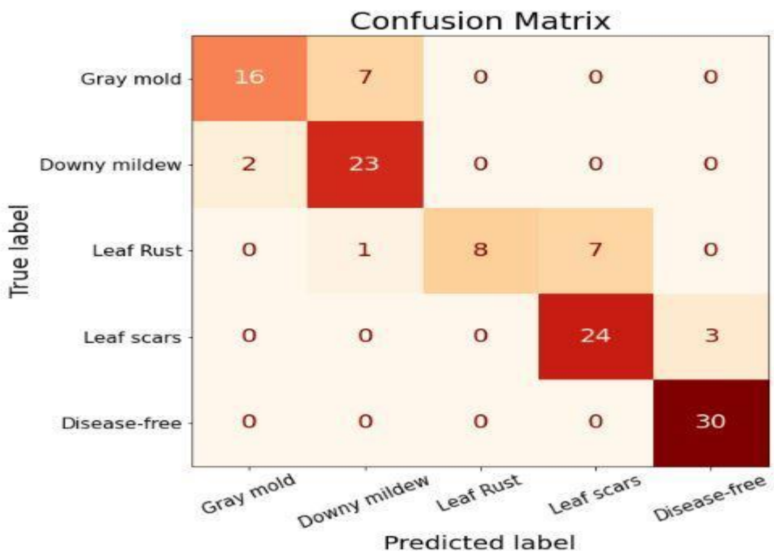


Figure: 4.1: Logistic Regression Confusion Matrices

In figure 4.2 is for ROC curve for measuring performance. The Y axis is stand for true positive rate and X axis is for false positive rate. Area under ROC class 0 is 0.99, class 1 is 0.92, class 2 is 093, class 3 is 0.94 and class 4 is 1.00.

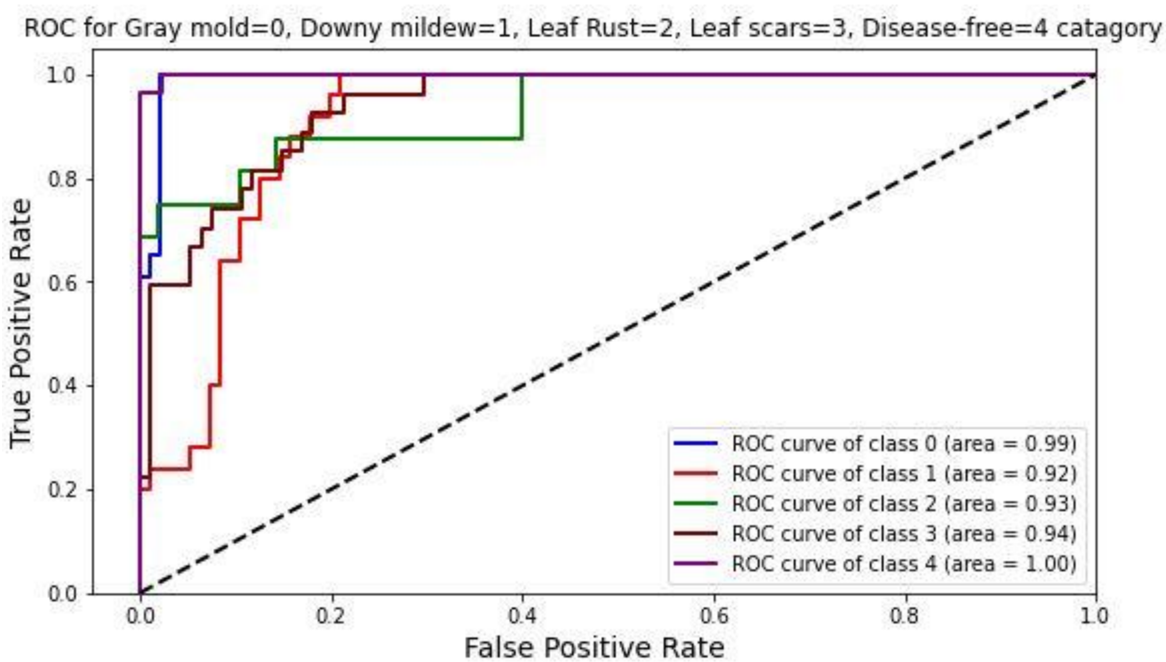


Figure 4.2: ROC Curve for measuring performance of Logistic Regression

4.3.2 Random Forest Result

In this research work random forest illustrate training accuracy 99% and 91.73% testing accuracy for detecting sunflower and sunflower leaf disease detection. Performance evaluation metrics for the Random Forest classifier are shown in Table 4.2. This classifier achieved the highest value for all performance parameters for disease free category. And the lowest precision, sensitivity, f1 score that are 0.87, 0.81, 0.84 and highest FPR, FNR that are 0.02, 0.18 for leaf rust. The confusion matrix of this algorithm shown in the below figure 4.2 shows the accurately predicted diseases and wrong predicted disease.

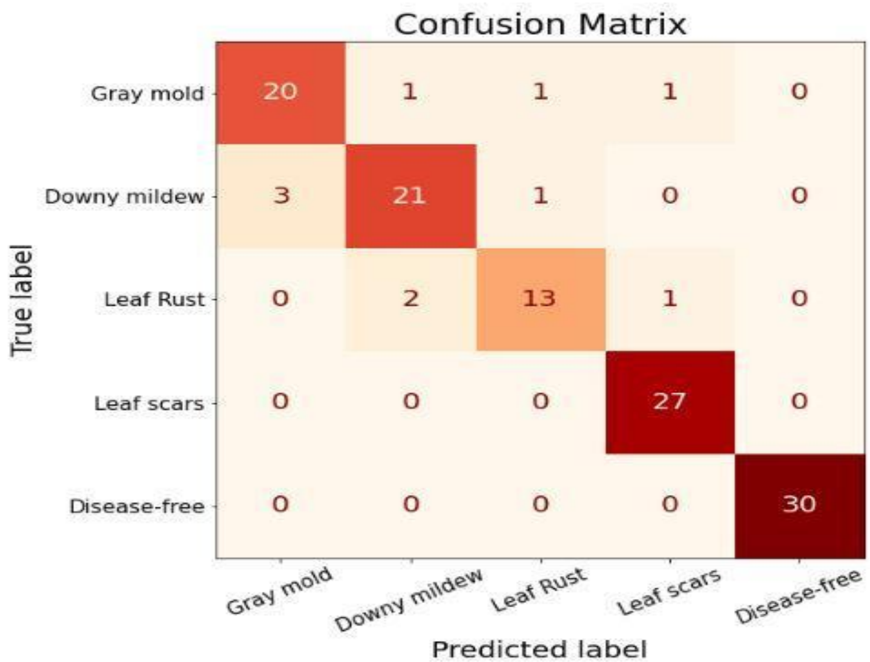


Figure 4.3: Confusion Matrices of Random Forest

Table 4.2: Random Forest result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Gray mold	0.87	0.87	0.87	0.97	0.03	0.13
Downy mildew	0.88	0.84	0.86	0.97	0.03	0.16
Leaf Rust	0.87	0.81	0.84	0.98	0.2	0.18
Leaf scars	0.93	1.00	0.96	0.97	0.02	0.00
Disease-free	1.00	1.00	1.00	1.00	0.00	0.00

In figure 4.4 is for ROC curve for measuring performance. The Y axis is stand for true positive value and X axis is for false positive value. The dashed line in each plot represents the perfect result – outputs = targets. In figure 4.4, area under ROC for class 0 is 1.00, class 1 is 0.97, class 2 is 0.96, class 3 is 1.00 and class 4 is 0.99.

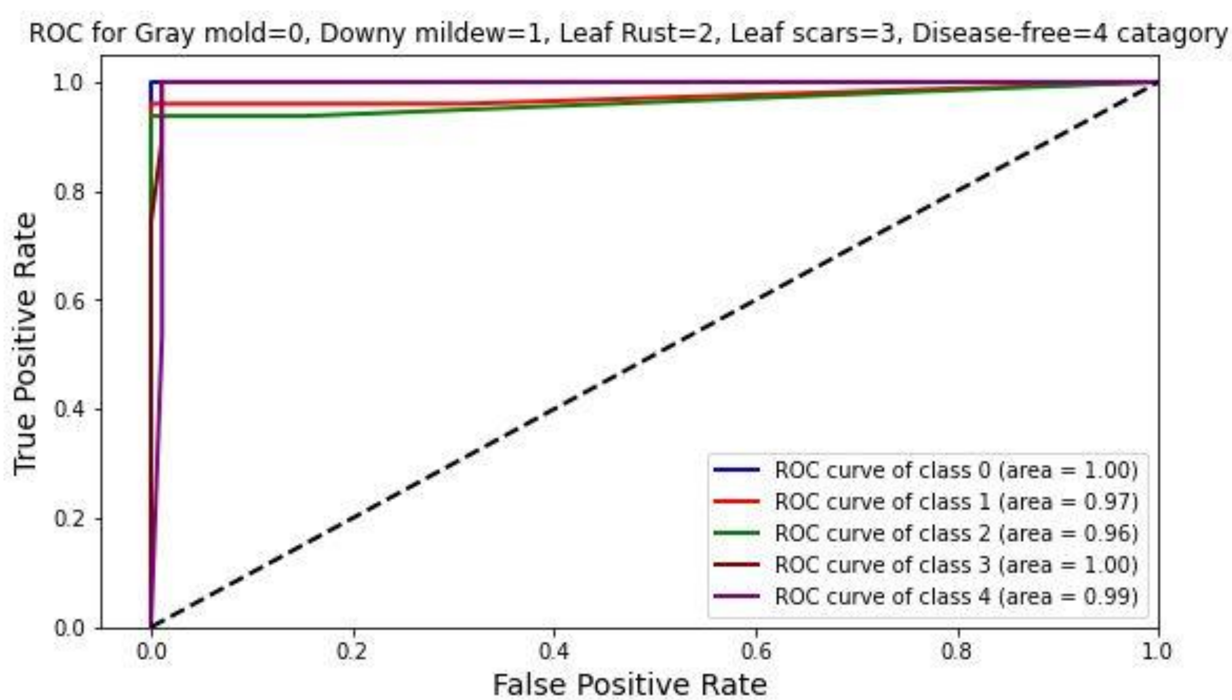


Figure 4.4: ROC Curve for measuring performance of Random Forest

4.3 Support Vector Machine Result

The support vector machine shows 83.3% training accuracy and 81.82% accuracy for predicting the disease of sunflower and sunflower leaves. Table 4.3 presents the overall performance for the support vector machine. The precision, sensitivity, specificity, f1 score and lowest FPR and FNR for disease free. Lowest sensitivity, f1 score that are 0.38, 0.55 and highest FNR 0.62 for leaf rust. Lowest precision and specificity and highest FPR for downy mildew

The confusion matrix of this algorithm shown in the below figure 4.5. Here we can see the correct detection and wrong detection of separate diseases. From this figure the TP, TN, FP, FN values are shown in 5×5 matrices. Following curve shown in figure 4.6 is for ROC curve for measuring performance. Y axis is stand for true positive rate and X axis is for false positive rate. The area under ROC for class 0 is 1.00, class 1 is 0.88, class 2 is 0.97, class 3 is 0.90 and class 4 is 1.00.

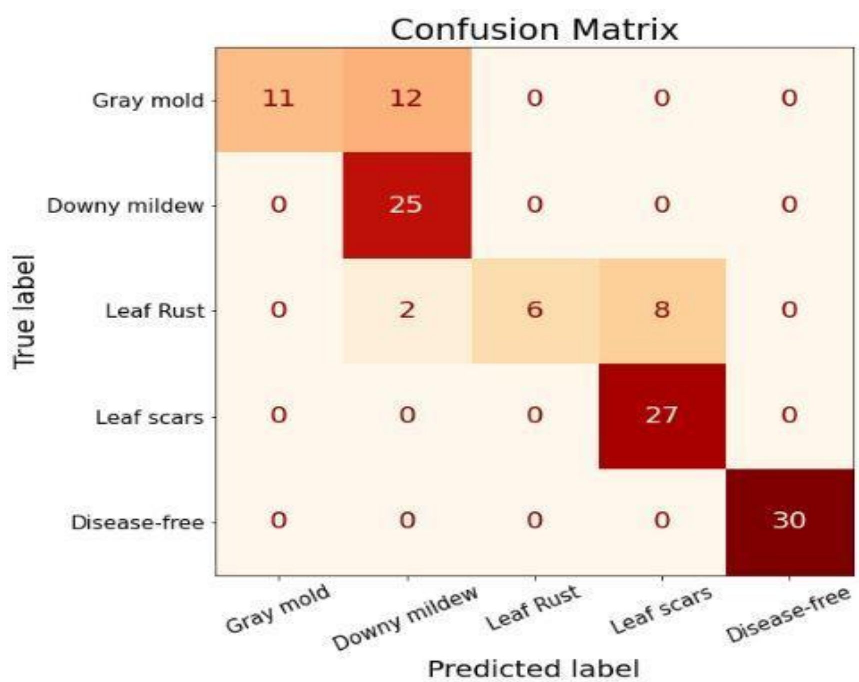


Figure 4.5: Support Vector Machine Prediction Confusion Matrices

Table 4.3: Support Vector Machine result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Gray mold	1.00	0.48	0.65	1.00	0.00	0.5
Downy mildew	0.64	1.00	0.78	0.85	0.14	0.00
Leaf Rust	1.00	0.38	0.55	1.00	0.00	0.62
Leaf scars	0.77	1.00	0.87	0.91	0.08	0.00
Disease-free	1.00	1.00	1.00	1.00	0.00	0.00

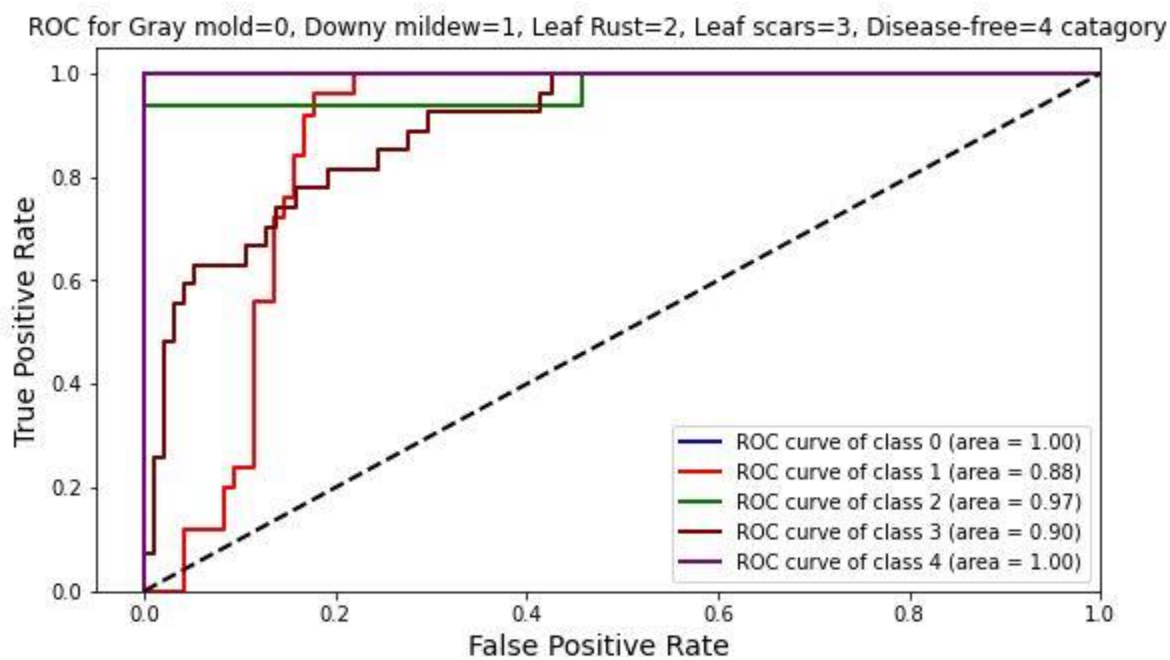


Figure 4.6: ROC Curve for measuring performance of Support Vector Machine

4.4 Gradient Boosting Result

In this research work gradient boosting algorithm illustrate 97.7 % training accuracy and 93.38% testing accuracy for detecting sunflower and sunflower leaf disease detection. The overall result analysis of gradient boosting for detecting separate disease categories is shown in Table 4.4. It shows highest precision, f1 score, specificity that are respectively 1.00, 0.98, 1.00 and lowest FPR 0.0 for gray mold category. Highest sensitivity 1.00, and lowest FNR 0.00 shows in disease free category. Lowest sensitivity 0.69, f1 score 0.79 and highest FNR 0.3 in leaf rust category. Lowest precision 0.90 in leaf scars category. The confusion matrix of this algorithm shown in the below figure shows the accurately predicted diseases and wrong predicted disease. The TP, TN, FP, FN is shown in this 5×5 figure. In figure 4.6 shows the ROC curve for measuring performance. Y axis is stand for true positive rate and X axis is for false positive rate. The area under ROC for class 0 is 1.00, class 1 is 1.00, class 2 is 1.00, class 3 is 0.97 and class 4 is 1.00.

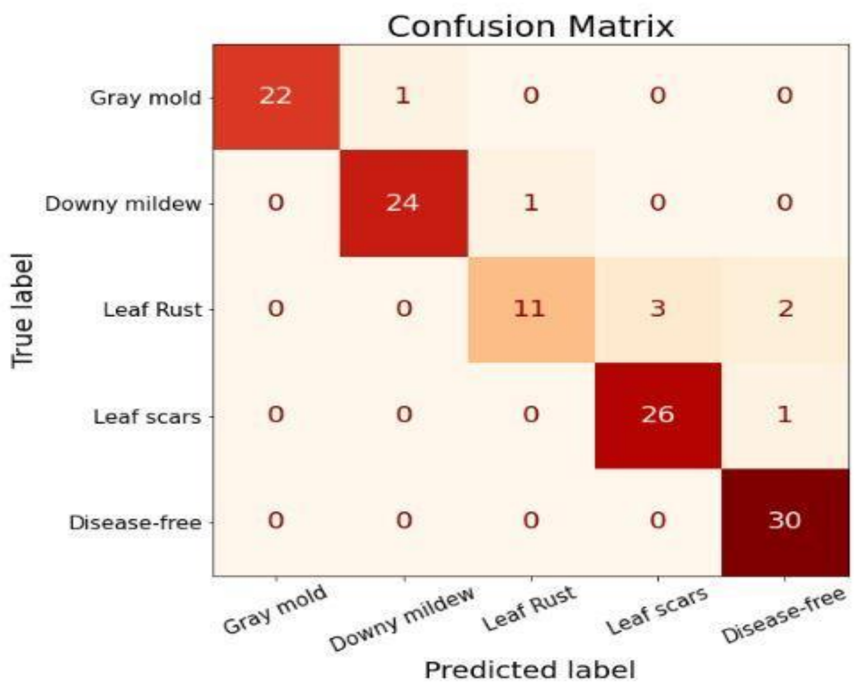


Figure 4.7: Gradient Boosting Confusion Matrices

Table 4.4: Gradient Boosting result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Gray mold	1.00	0.96	0.98	1.00	0.00	0.05
Downy mildew	0.96	0.96	0.96	0.98	0.01	0.04
Leaf Rust	0.92	0.69	0.79	0.99	0.01	0.3
Leaf scars	0.90	0.96	0.93	0.96	0.03	0.03
Disease-free	0.91	1.00	0.95	0.96	0.03	0.00

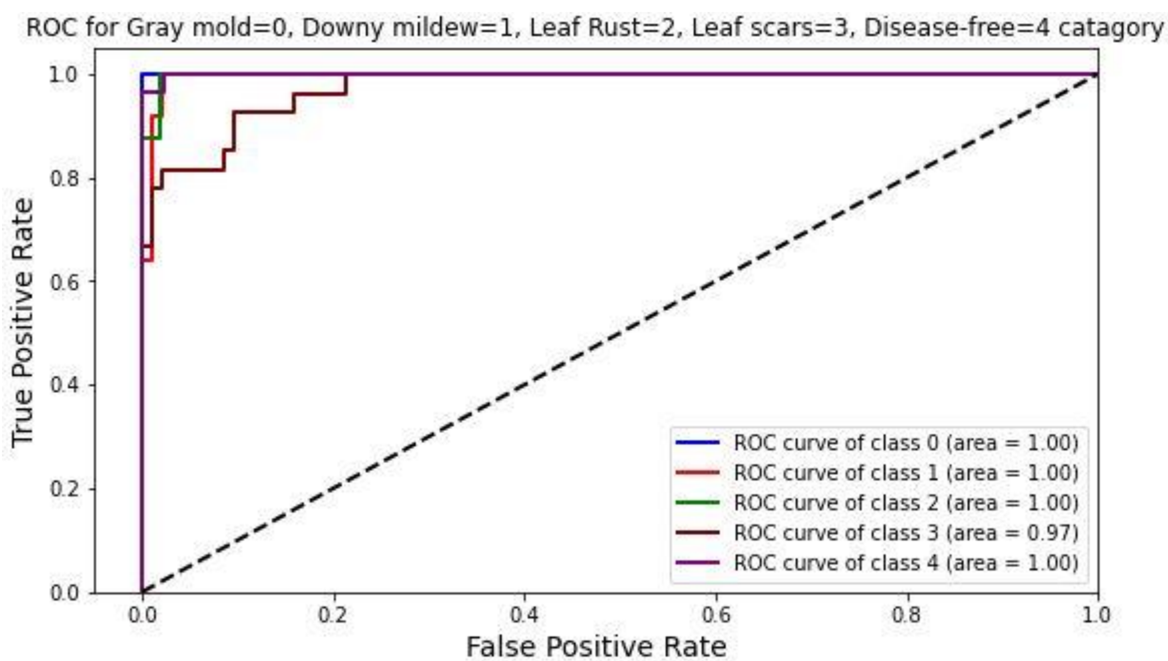


Figure 4.8: ROC Curve for measuring performance of Gradient Boosting

4.5 K-Nearest Neighbors Result

The k-nearest neighbor shows 90.00% training accuracy and 90.00% testing accuracy for predicting the disease of sunflower and sunflower leaves. The overall result analysis of k-nearest neighbor for detecting separate disease categories is shown below in Table 4.5. It shows the maximum sensitivity 1.00, f1 score 0.98 and lowest FNR for disease free sunflower. Maximum precision 1.00, specificity 1.00 and lowest FPR for gray mold and leaf rust. Lowest precision 0.74 and sensitivity 0.70 shown for respectively downy mold and gray mold.

The confusion matrix of this algorithm shown in the below figure 4.9. Here we can see the separate disease prediction correctly and wrongly. The TP, TN, FP, FN is shown in this 5×5 matrices figure. In figure 4.10 shows the ROC curve for measuring performance. Y axis is stand for true positive rate and X axis is for false positive rate. The area under ROC for class 0 is 1.00, class 1 is 0.99, class 2 is 1.00, class 3 is 1.00 and class 4 is 1.00.

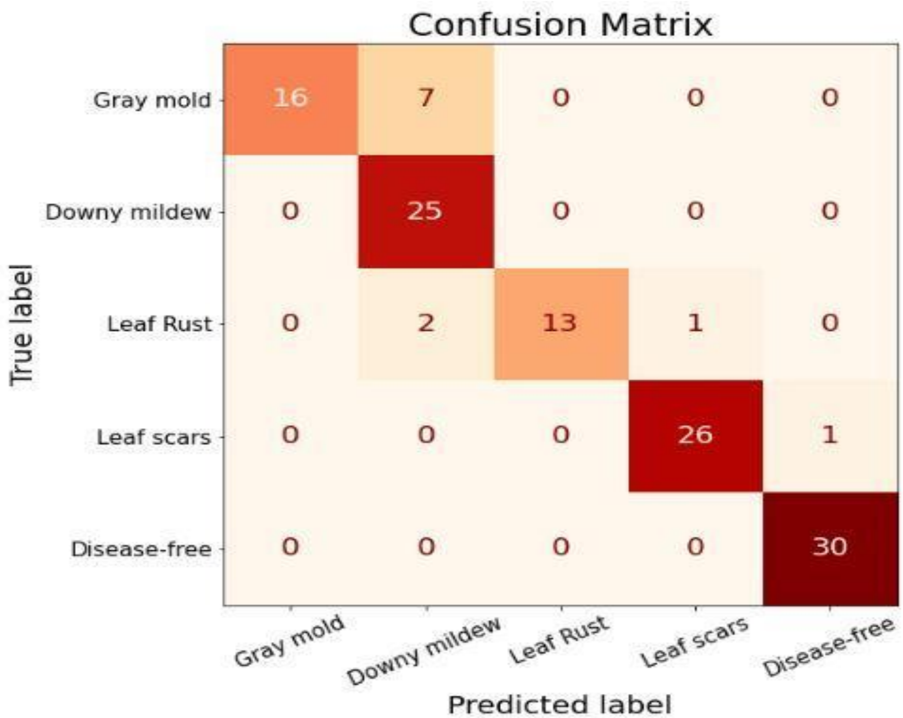


Figure 4.9: K-Nearest Neighbors Confusion Matrices

Table 4.5: K-Nearest Neighbors result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Gray mold	1.00	0.70	0.82	1.00	0.00	0.3
Downy mildew	0.74	1.00	0.85	0.9	0.09	0.00
Leaf Rust	1.00	0.81	0.90	1.00	0.00	0.18
Leaf scars	0.96	0.96	0.96	0.98	0.01	0.03
Disease-free	0.97	1.00	0.98	0.98	0.01	0.00

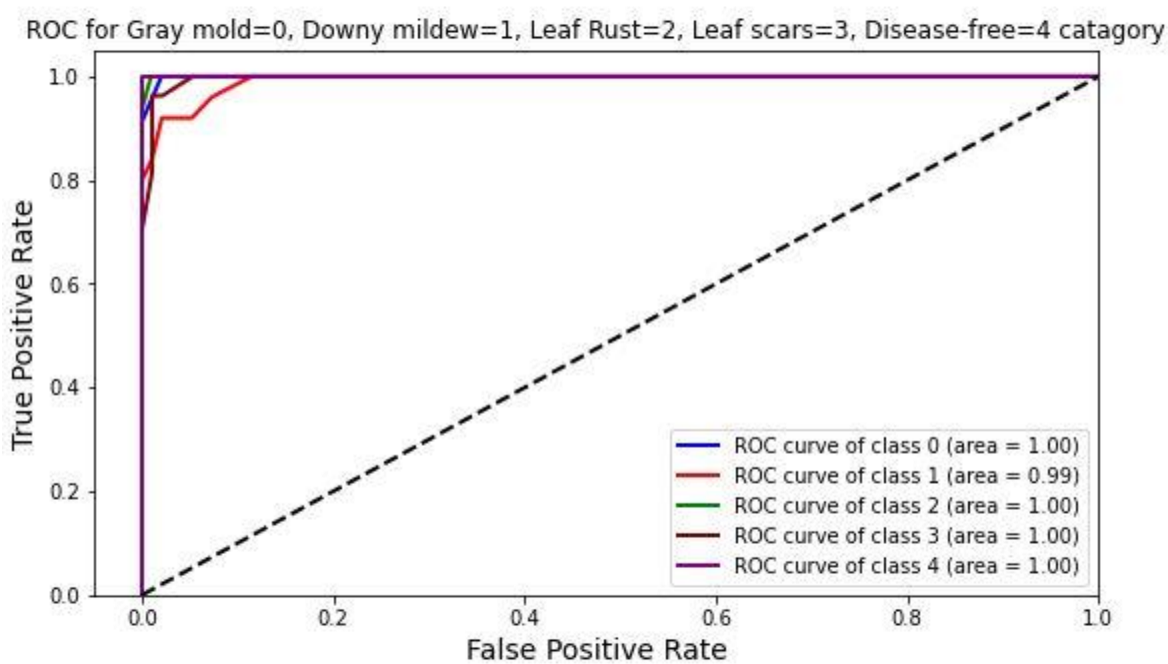


Figure 4.10: K-Nearest Neighbors ROC Curve

4.6 XGBoost Classifier Result

In this research work the XGBoost classifier illustrates 100% training accuracy and 92.56% accuracy for detecting sunflower and sunflower leaf disease detection. Table 4.6 shows the performance matrices for xgboost classifier. The maximum sensitivity 1.00, f1 score 0.97 and lowest FNR 0.00 shown for disease free. Highest precision 1.00 and lowest FPR. For downey mold shows the lowest result in all parameters shown for gray mold and leaf rust. The confusion matrix of this algorithm shown in the below figure shows the accurately predicted diseases and wrong predicted disease. The TP, TN, FP, FN is shown in this 5×5 matrices figure. In figure 4.10 shows the ROC curve for measuring performance. Y axis is stand for true positive rate and X axis is for false positive rate. The area under ROC for class 0 is 1.00, class 1 is 0.99, class 2 is 1.00, class 3 is 1.00 and class 4 is 1.00.

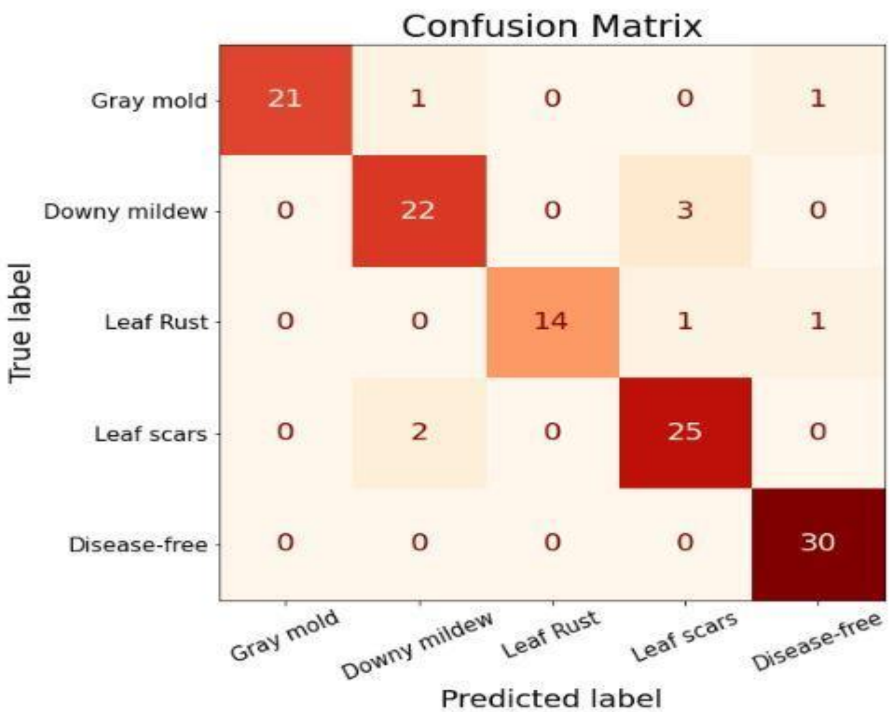


Figure 4.11: XGBoost Classifier Confusion Matrices

Table 4.6: Extreme Gradient Boosting result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Gray mold	1.00	0.91	0.95	1.00	0.00	0.08
Downy mildew	0.88	0.88	0.88	0.96	0.03	0.12
Leaf Rust	1.00	0.88	0.93	1.00	0.00	0.12
Leaf scars	0.86	0.93	0.89	0.95	0.04	0.07
Disease-free	0.94	1.00	0.97	0.97	0.02	0.00

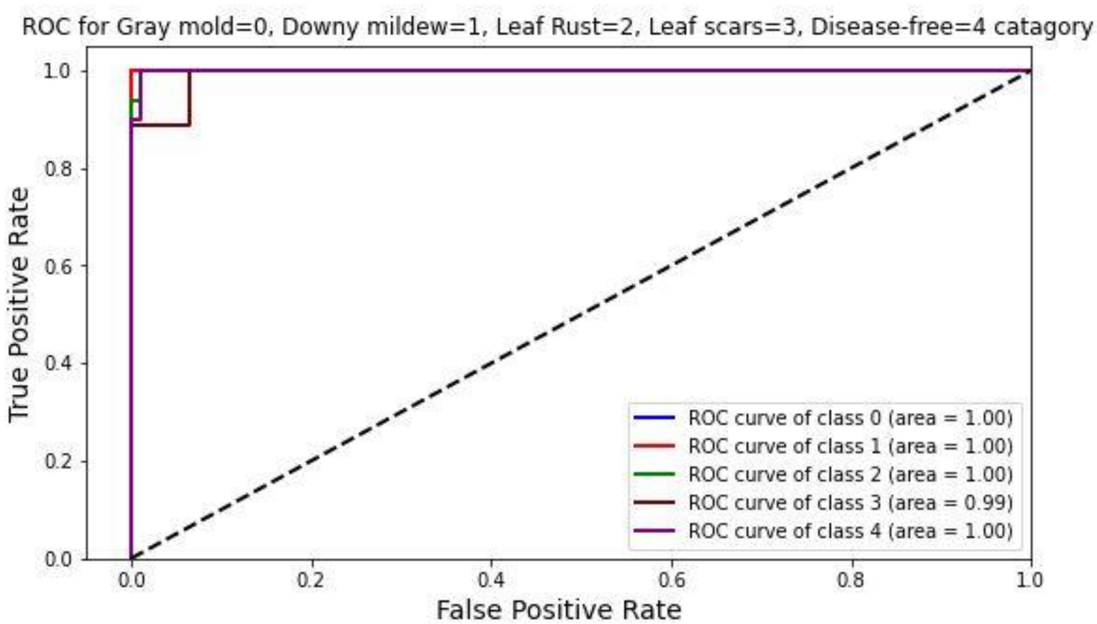


Figure 4.12: XGBoost Technique ROC Curve

4.7 Residual Network 50 Result

The ResNet 50 shows 96.19% accuracy and 0.09 loss for predicting the disease of sunflower and sunflower leaves. The overall result analysis of ResNet 50 for detecting separate disease categories is shown in Table 4.7. The highest precision, sensitivity and f1 score shown for gray mold and disease free. Lowest precision 0.89, lowest sensitivity 0.89 and fi score 0.94 shown for respectively leaf scars, downy mildew.

Table 4.7: ResNet 50 result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score
Gray mold	1.00	0.99	0.99
Downy mildew	0.91	0.96	0.93
Leaf scars	0.95	0.92	0.93
Disease-free	1.00	0.99	0.99

The confusion matrix of this algorithm shown in the below figure 4.7. Here we can see the separate disease, how many are correctly detected and wrong detected.

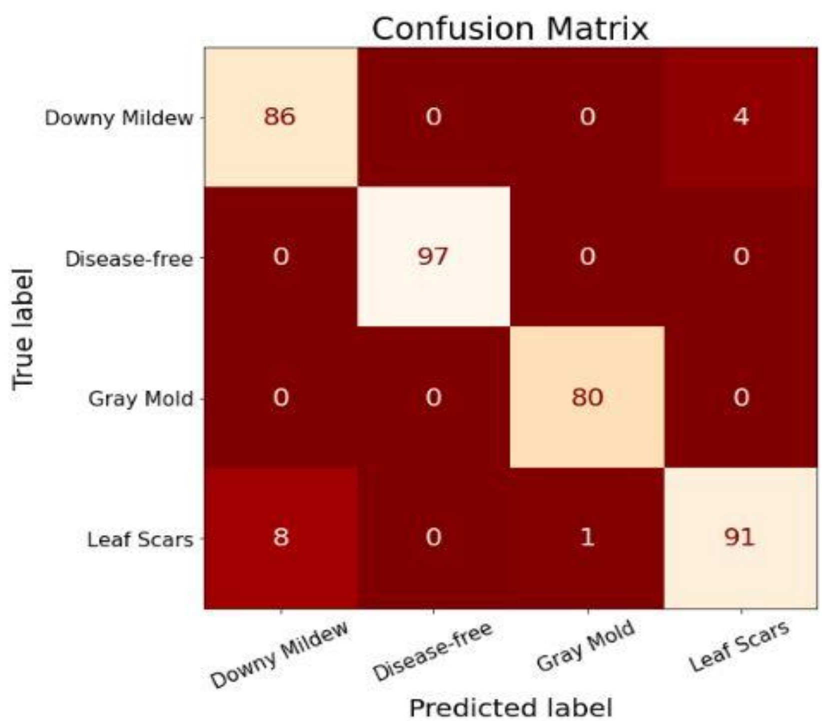


Figure 4.13: ResNet 50 Confusion Matrices

The ResNet 50 model accuracy and model loss figure 4.14 are also shown below. This figure shows how the model accuracy is changed with the epochs. And figure 4.15 illustrates the how the model is loosed with the epochs for validation data set.

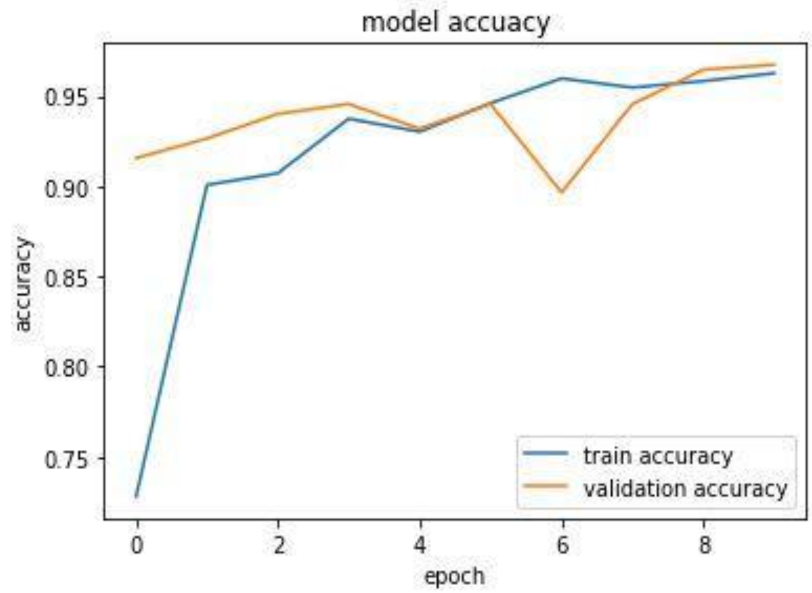


Figure 4.14: ResNet-50 Model Accuracy Train vs Validation

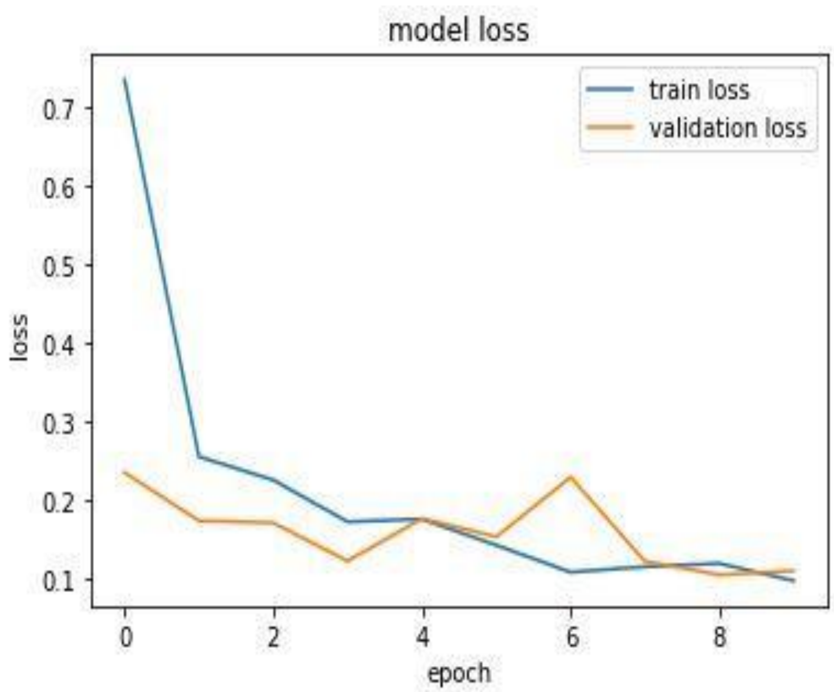


Figure 4.15: ResNet 50 Model Loss Train Loss vs Validation Loss

4.8 Visual Geometry Group-19 Layer Result

In this research work the VGG19 algorithm illustrates 92% accuracy and 0.20 loss for detecting sunflower and sunflower leaf disease detection. The overall result analysis of the VGG19 classifier for detecting separate disease categories is shown in Table 4.8. The highest precision 0.99, sensitivity 1.00 and f1 score 0.99 shown for gray mold. And the lowest precision 0.88, sensitivity 0.86 and f1 score 0.87 shown for downy mildew.

Table 4.8: VGG 19 result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score
Gray mold	0.99	1.00	0.99
Downy mildew	0.88	0.86	0.87
Leaf scars	0.88	0.90	0.89
Disease-free	0.98	0.97	0.97

The confusion matrix of this algorithm shown in the below figure shows the accurately predicted diseases and wrong predicted disease. The VGG19 model accuracy and model loss figure are also shown below. This figure shows how the model accuracy is changed with the epochs. And figure 4.15 illustrates the how the model is loosed with the epochs for validation data set.

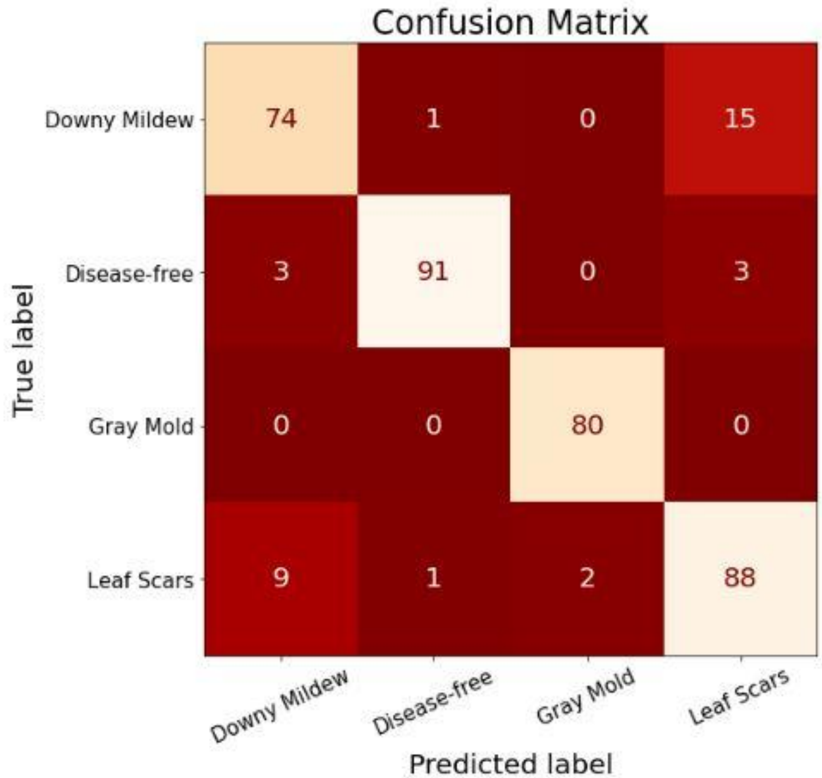


Figure 4.16: VGG 19 Confusion Matrices

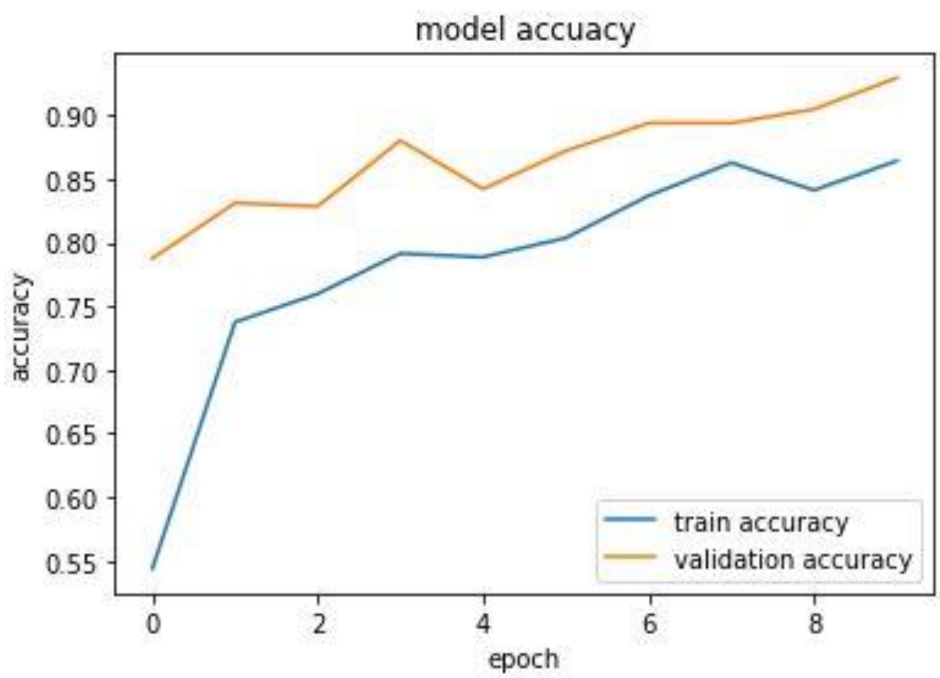


Figure 4.17: VGG 19 Model Accuracy Curve Training Accuracy vs Validation Accuracy

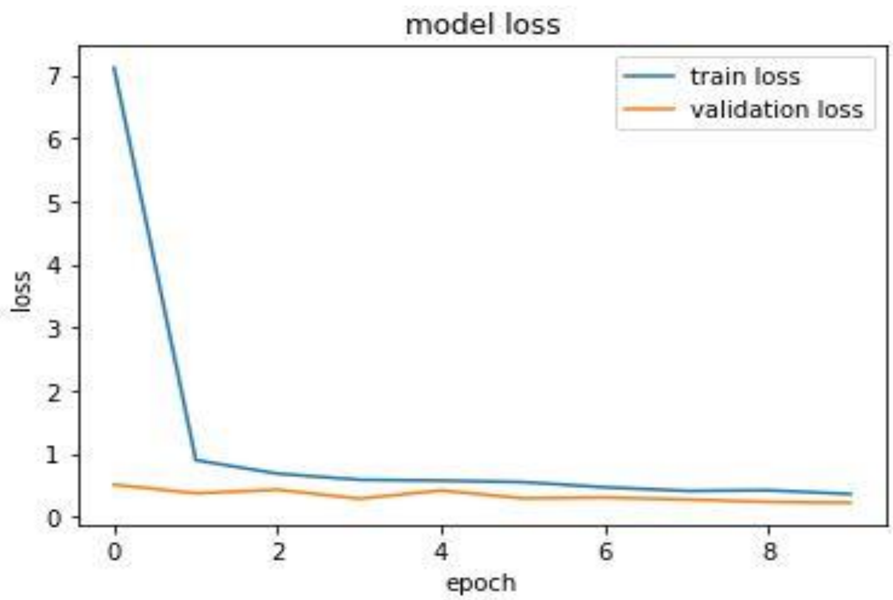


Figure 4.18: VGG 19 Model Loss Curve Training Accuracy vs Validation Accuracy

4.9 Mobile Network Version 2

The Mobile Net V2 shows 76.76% accuracy and 2.73 loss value for predicting the disease of sunflower and sunflower leaves. Table 4.9 presents the overall performance for the Mobile Net V2. Mobile Net illustrates the maximum sensitivity 0.97, f1 score 0.88 for gray mold and highest precision for disease free. Lowest performance parameters shown for leaf scars.

Table 4.9: Mobile Net V2 result for individual disease detection

Disease Name	Precision	Sensitivity	F1-score
Gray mold	0.80	0.97	0.88
Downy mildew	0.72	0.69	0.70
Leaf scars	0.71	0.60	0.65
Disease-free	0.82	0.84	0.83

The confusion matrix of this algorithm shown in the below figure. Here we can see the separate disease correct detection and wrong detection. The MobileNet V2 model accuracy and model loss figure are also shown below. This figure shows how the model accuracy is changed with the epochs. And figure 4.15 illustrates the how the model is loosed with the epochs for validation data set.

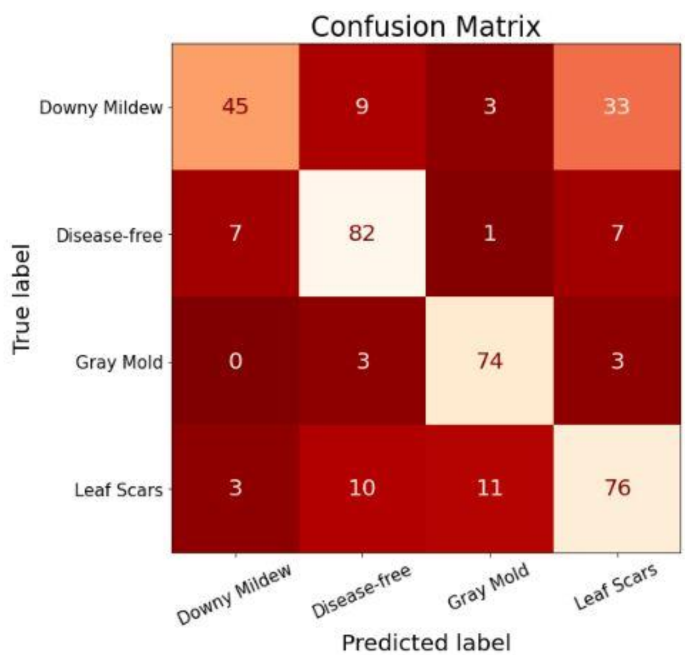


Figure 4.19: Mobile Net V2 Confusion Matrices

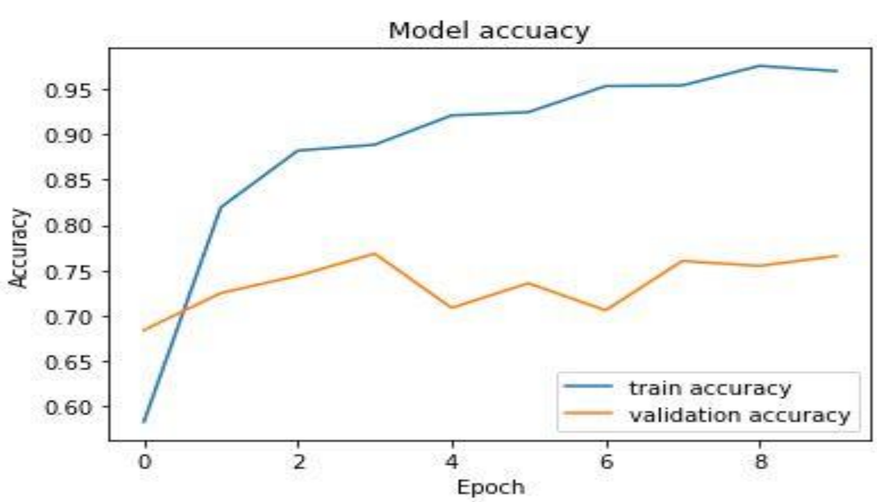


Figure 4.20: MobileNet V2 Model Accuracy Curve Training Accuracy vs Validation Accuracy

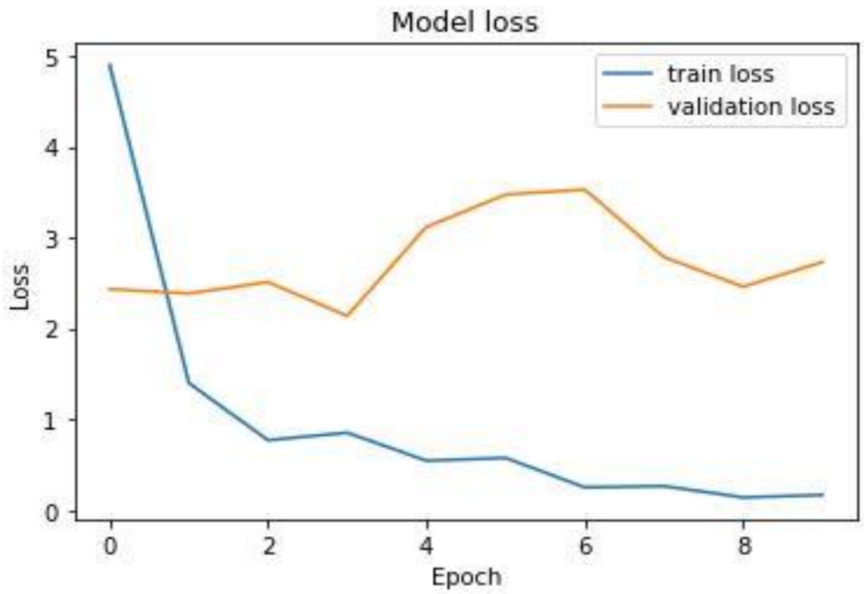


Figure 4.21: MobileNet V2 Model Loss Curve Training Accuracy vs Validation Accuracy

4.4 Discussion

In order to evaluate the overall performance of these classifiers seven aspects of these classifiers will be considered. The accuracy, precision, sensitivity, f1-score, specificity, FPR, FNR these classifiers will be considered for exploring the best machine learning classifier for predicting sunflower disease.

4.4.1 Comparative Analysis of Different Machine Learning Techniques

Comparison of different performance terminologies like confusion matrix, ROC curve, Performance table, performance histogram etc. of different machine learning techniques are shown below.

Table 4.10 shows the accuracy, precision, sensitivity, F1-score, specificity, FPR, FNR of machine learning techniques. From Table 4.10, it is noticed that gradient boosting shows the maximum accuracy, precision, specialty and minimum FNR and FPR that values are respectively 93.38%, 0.94, 0.98, 0.01 0.08. On the other hand the sensitivity and f1 score of gradient boosting respectively 0.91 and 0.92 is lower than xgboost classifier that are respectively 0.92 and 0.93. But the xgboost techniques shows 92.56% accuracy and the other performance parameter is lower in this techniques. Random forest illustrate 2nd maximum accuracy, sensitivity, fi score and specialty, the values are respectively 91.73%, 0.90, 0.91, 0.97 and lower FPR values is 0.02, FNR value is 0.09. K-Nearest Neighbors gives 2nd highest precision 0.93 than random forest but the other parameter is lower than random forest. Logistic regression shows highest accuracy 83.47% and sensitivity 0.80 and lowest precision than support vector machine but other parameters are almost same but lower than other classifiers.

Table 4.10: Machine learning prediction result

Classifier Name	Accuracy	Precision	Sensitivity	F1-score	Specificity	FPR	FNR
Logistic Regression	83.47%	0.86	0.80	0.81	0.96	0.042	0.19
Random Forest	91.73%	0.91	0.90	0.91	0.97	0.02	0.09
Gradient Boosting	93.38%	0.94	0.91	0.92	0.98	0.01	0.08
Support Vector Machine	81.81%	0.88	0.77	0.77	0.95	0.04	0.2
K-Nearest Neighbors	90.00%	0.93	0.89	0.90	0.97	0.02	0.1
XGB Classifier	92.56%	0.94	0.92	0.93	0.98	0.01	0.08

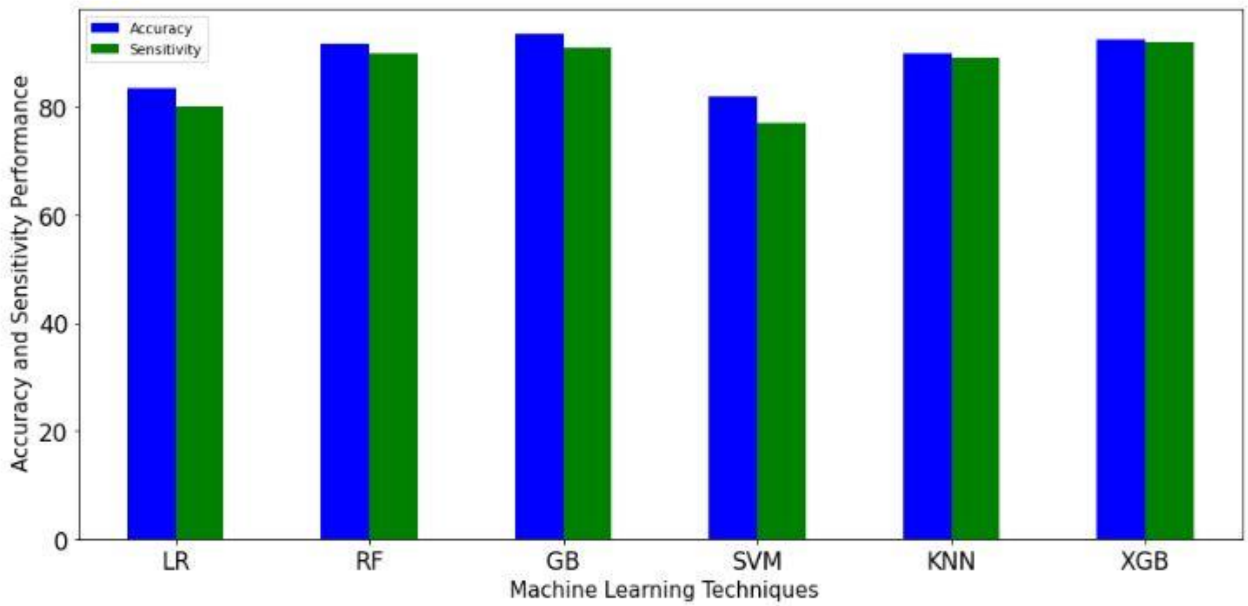


Figure 4.22: Comparison chart of accuracy and sensitivity of different ML techniques

Figure 4.22 shows accuracy and sensitivity histogram of al machine learning techniques. The figure shows that gradient boosting accuracy is high but sensitivity is low. Xgboost accuracy and sensitivity is almost same and maximum

Therefore from the overall discussion it showed that the gradient boosting classifier shows highest accuracy but lower sensitivity than the xgboost classifier and xgboost shows higher accuracy and these both are the best for predicting sunflower disease.

One of the main tasks of the research is to show the prediction with new samples. Since the main goal of the research is to find out the best algorithm which will give better accuracy than the early prediction system of sunflower disease. Here, 20 target prediction against each sample target is considered to understanding the procedure shown in table 4.11.

Table 4.11: Target and Predicted value of new samples of XGB technique

Sample No	Target	Predicted	Remarks
1	1	1	Right
2	3	3	Right
3	4	4	Right
4	0	0	Right
5	1	1	Right
6	2	2	Right
7	0	0	Right
8	4	4	Right
9	0	0	Right
10	3	1	Wrong
11	0	0	Right
12	1	1	Right
13	1	1	Right
14	4	4	Right
15	3	3	Right
16	2	3	Wrong
17	0	0	Right
18	1	1	Right
19	1	1	Right
20	0	0	Right

4.4.1 Comparative Analysis of Different Machine Learning Techniques

Table 4.12: CNN classifiers result

CNN Model Name	Accuracy	Loss	Average Precision	Average Sensitivity	Average F1-score
ResNet 50	95.09%	0.09	0.96	0.96	0.96
VGG 19	92.91%	0.20	0.93	0.92	0.93
Mobile Net V2	76.76%	2.73	0.76	0.77	0.76

Comparison of different performance terminologies like confusion matrix, Performance table, performance histogram etc. of different machine learning techniques are shown below. Table 12 illustrates the accuracy, precision, sensitivity, F1-score of convolution neural network models. Table 4.9 shows the ResNet 50, VGG 19 and Mobile Net V2 experiment results. The ResNet 50 shows the maximum accuracy 96.19%, average precision 0.96, average sensitivity 0.96, average f1 score 0.96 and minimum loss 0.097. Mobile Net V2 shows the minimum accuracy 76.76%, average precision 0.76, average sensitivity 0.77, average f1 score 0.76. And maximum loss 2.73. VGG 19 shows the average result between ResNet 50 and Mobile Net V2. Figure 4.23 shows accuracy, precision, sensitivity and f1 score histogram of all CNN models. The figure shows that resnet 50 shows highest accuracy, sensitivity, fi score than others CNN models. From the overall discussion, ResNet 50 is the best model for classifying sunflower disease.

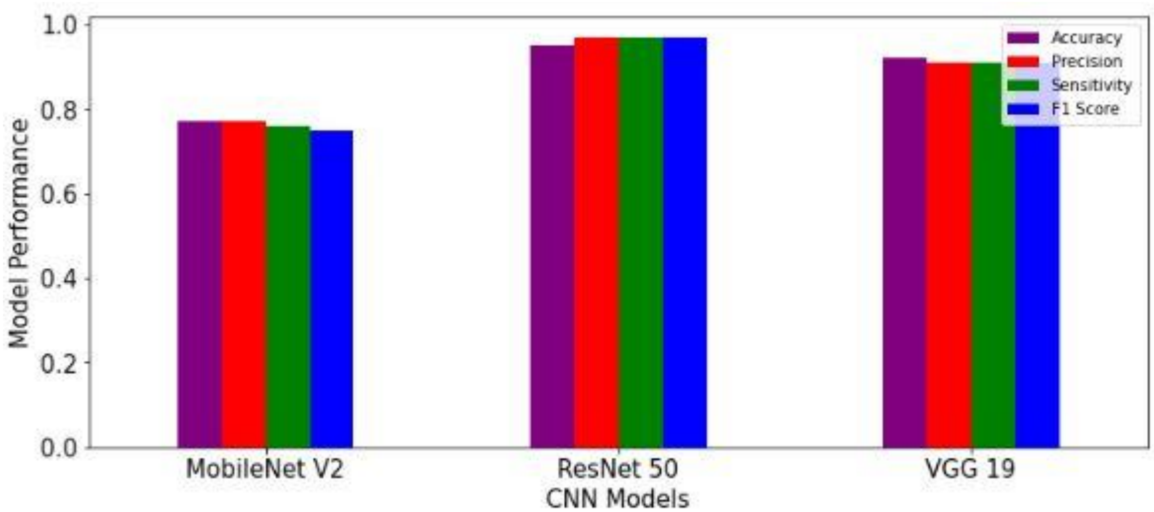


Figure 4.23: Comparison chart of accuracy and sensitivity of different ML techniques

Target values and predicted values for ResNet 50 model is shows in the table 4.11. Here, 20 target prediction against each sample target is considered to understanding the procedure.

Table 4.13: Target and Predicted value of new samples of ResNet 50 technique

Sample No	Target	Predicted	Remarks
1	0	0	Right
2	0	0	Right
3	0	0	Right
4	0	0	Right
5	0	0	Right
6	0	3	Wrong
7	1	1	Right
8	1	1	Right
9	1	1	Right
10	1	1	Right
11	2	2	Right
12	2	2	Right
13	2	2	Right
14	2	2	Right
15	2	2	Right
16	3	3	Right
17	3	3	Right
18	3	3	Right
19	3	3	Right
20	3	0	Wrong

4.4.3 Comparative Analysis of Different Machine Learning Techniques and CNN models

Table 4.13 shows the final accuracy, sensitivity, F1-score of all the machine learning and convolutional neural networks.

From the table the CNN model classifier ResNet 50 shows the maximum accuracy, precision, sensitivity and f1 score among all classifiers of machine learning and CNN classifiers. From machine learning classifiers xgboost maximum precision, sensitivity and f1 score. Gradient boosting shows maximum accuracy. But both are lower than the ResNet 50 CNN classifier. After discussing all aspects, ResNet 50 is the best classifier for predicting sunflower disease. Figure 4.22 shows the accuracy and sensitivity comparison of ResNet 50 and XGBoost.

Table 4.14: Machine Learning classifiers and CNN models performances

Classifier Name	Accuracy	Precision	Sensitivity	F1-score
Logistic Regression	83.47%	0.86	0.80	0.81
Random Forest	91.73%	0.91	0.90	0.91
Gradient Boosting	93.38%	0.94	0.91	0.92
Support Vector Machine	81.81%	0.88	0.77	0.77
K-Nearest Neighbors	90.00%	0.93	0.89	0.90
XGB Classifier	92.56%	0.94	0.92	0.93
ResNet 50	95.09%	0.96	0.96	0.96
VGG 19	92.91%	0.93	0.92	0.93
Mobile Net V2	76.76%	0.76	0.77	0.76

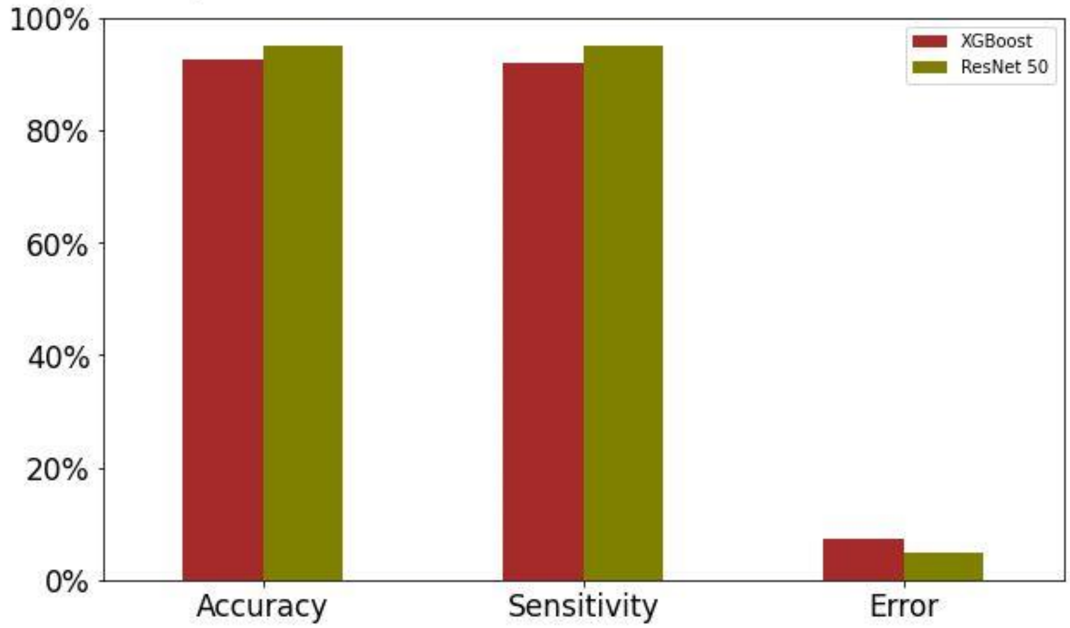


Figure 4.22: Comparison chart of XGBoost and ResNet 50 techniques

4.5 Extracted Feature Selection Result

Figure 4.23 illustrate the extracted features ranking graph using anova features selection method. From this figure image texture feature energy is highly responsible for predicting sunflower disease it has highest f value and kurtosis and skewness are moderately responsible for disease prediction. And among extracted 10 features image correlation feature shows the minimum f value that’s why correlation has lowest responsibility for detecting sunflower disease.

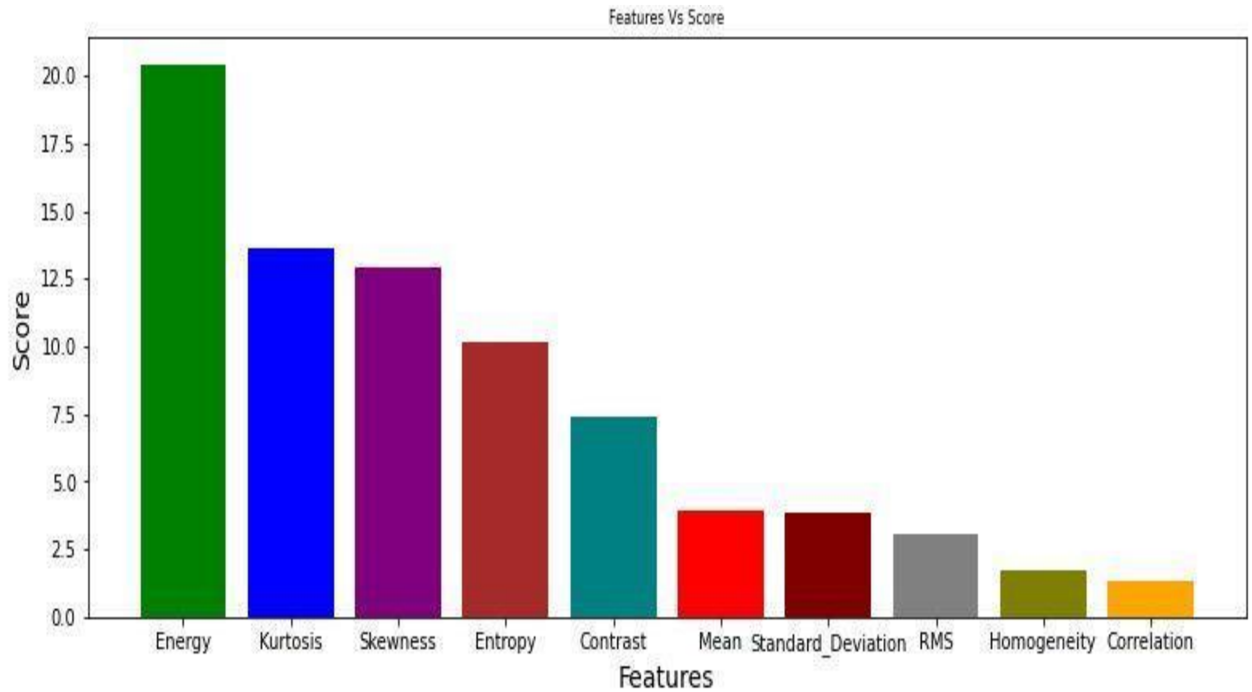


Figure 4.23: Image extracted features ranking according to the F value

4.5 Summary

In this chapter, analyze the parameters, results of all algorithms are discussed in details and select the final model among neural networks and decision tree algorithm which is considered for this work. And show the comparison between all of the techniques.

CHAPTER 5

CONCLUSION

5.1 Introduction

This Chapter Describes The Conclusion And Future Research Scope Of This Thesis Work. First The Overall Summary Of This Thesis Work Is Presented In The Conclusion Section. Then The Potential Future Research Directions Are Provided In Future Research Scope Section

5.2 Conclusion

This research focuses on the use of different supervised classification techniques to enhance the early detection of sunflower disease. Now a days sunflower disease is a fatal disease for the farmers who are cultivating the sunflower. They also faced economical loss. Here used real time sunflower image dataset, which are collected from field level. Here, 4 sunflower fungal disease gray mold, downy mildew, leaf Scars, and leaf Rust is used for predicting. Disease free sunflower is also used for detection. Sunflower fungal disease is more serious than other diseases. Different machine learning and deep learning models are used for predicting these diseases. The 11 features are extracted from the 602 image data for learning the machine learning models which are done by GLCM technique. And 1500 augmented images of gray mold, downy mildew, leaf Scars and disease free are for CNN model prediction.

The machine learning techniques such as logistic regression, random forest, gradient boosting, support vector machine, k-nearest neighbor and xgboost shows the prediction accuracy respectively 83.47%, 91.73%, 93.38%, 81.81%, 90.00% and 92.56%. And convolutional neural network resnet 50 shows 96.19%, vgg 92.00% and mobilenet v2 76.76% accuracy to detection.

By the evaluation the experimental results of all algorithms, this work concludes that, the resnet 50 cnn model can be considered as a best algorithm because of its highest prediction accuracy 96.19%. As resnet 50 gives better accuracy it can be used for early sunflower disease prediction.

5.3 Future Research Scope

There is a possibility to generate a new algorithm or a hybrid one which will provide better accuracy than the existing algorithms. Data from several districts of Bangladesh can be used for wide research in this topic in future.

REFERENCES

[1] The Top Sunflower Seed Producing Countries In The World: https://www.worldatlas.com/articles/the-top-sunflower-seed-producing-countries-in-the-world.html

[2] National Sunflower association: https://www.sunflowernsa.com/growers/Diseases/

[3] V.Singh, "Sunflower leaf diseases detection using image segmentation based on particle swarm optimization," Artificial Intelligence in Agri-culture, vol. 3, pp. 62-68, 2019.

[4] J. Liu, F. Lv, and P. Di, "Identification of sunflower leaf diseases based on random forest algorithm," In 2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS), pp. 459-463, IEEE, 2019.

[5] Sinan Uguz and Nese Uysal, "Classification of olive leaf diseases using deep convolutional neural networks," Neural Computing and Applications, vol. 33, pp. 4133-4149, 2021.

[6] S. V. Militante, B. D. Gerardo, and R. P. Medina, "Sugarcane Disease Recognition using Deep Learning," In 2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), pp. 575-578, IEEE, 2019.

[7] Zhaohua Huang et al., "Grape Leaf Disease Detection and Classification Using Machine Learning," 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics) , IEEE, November 2020.

[8] Surampalli Ashok et al., "Tomato Leaf Disease Detection Using Deep Learning Techniques," 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 979-983, 2020.

[9] S. Gayathri et.al., "Image Analysis and Detection of Tea Leaf Disease using Deep Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 398-403, 2020.

[10] T. Rumpf et.al., “Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance,” Computers and Electronics in Agriculture, vol. 74, pp. 91-99, 2010

[11] Edna Chebet Too et al., “A comparative study of fine-tuning deep learning models for plant disease identification,” Computers and Electronics in Agriculture, vol. 161, pp. march 2018.

[12] Aditya Karlekar and Ayan Seal, “SoyNet: Soybean leaf diseases classification,” Computers and Electronics in Agriculture, vol. 172, may 2020.

[13] Geetharamani G and Arun Pandian J, “Identification of plant leaf diseases using a nine-layer deep convolutional neural network,” Computers and Electrical Engineering, vol. 76, pp. 323-338, june 2019.

[14] Miaomiao Ji et al., “Automatic Grape Leaf Diseases Identification via United Model Based on Multiple Convolutional Neural Networks,” Information Processing in Agriculture, vol. 7, October 2019.

[15] Akash Sirohi and Dr. Arun Malik., “A Hybrid Model for the Classification of Sunflower Diseases Using Deep Learning,” 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), pp. 58-62, june 2021.

[16] Mercelin Francis and C. Deisy, “Disease Detection and Classification in Agricultural Plants Using Convolutional Neural Networks – A Visual Understanding,” 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 1063-`1068, march 2019.

[17] D. Penghui, L. Fang, and W. Xue,” The Research on the Feature Extraction of Sunflower Leaf Rust Characteristics Based on Color and Texture Feature,” In 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 457-460, IEEE, 2015.

[18] A. Rajbongshi, T. Sarker, M. M. Ahamad, and M. M. Rahman, ”Rose Diseases Recognition using MobileNet,” In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-7, IEEE, 2020.

[19] I. Gogul, and V. S. Kumar, “Flower species recognition system using convolution neural networks and transfer learning,” In 2017 fourth international conference on signal processing, communication and net-working (ICSCN), pp. 1-6, IEEE, 2017.

[20] S.Ramesh and D.vydeki, “Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm,” INFORMATION PROCESSING IN AGRICULTURE, pp. 249-260, 2020.

[21] Yang Lu et al., “Identification of Rice Diseases using Deep Convolutional Neural Networks,” Neurocomputing, 2017

[22] Mr. Pramod S. landge et al., “Automatic Detection and Classification of Plant Disease through Image Processing ,” vol. 3, july 2013.

[23] Huiqun Hong et al,. “Tomato Disease Detection and Classification by Deep Learning,” 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), june 2020.

[24] Gunasekaran, S., & Ding, K. Using computer vision for food quality evaluation. Food Technology, 6, 151-154, 1994

[25] Shishi Dong, Zhexue Huang. A Brief Theoretical Overview ofRandom Forests[J]. Integrated technology, pp.1-7, 2013.

[26] A. Krizhevsky, I. Sutskever, and G. H. E. Hinton, “ ImageNet Classification with Deep Convolutional Neural Networks”, in advances in Neural Information Processing Systems, 2012.

[27] J. Amara, B. Bouaziz, and A. Algergawy, “A Deep Learning-based Approach for Banana Leaf Diseases Classification,” BTW, pp.79–88, 2017.

[28] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, “Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification,” Computer Intelligence Neuroscience, vol. 2016, 2016.

[29] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 770-778, 2016.

[30] Liu, X.; Jia, Z.; Hou, X.; Fu, M. “Real-time marine animal images classification by embedded system based on mobilenet and transfer learning,” In Proceedings of the OCEANS 2019, Marseille, France, pp. 1–5, 17–20 June 2019.

[31] Michele, A.; Colin, V.; Santika, D. MobileNet Convolutional Neural Networks and Support Vector Machines for Palmprint Recognition. Procedia Comput. Sci., vol. 157, pp. 110–117, 2019.

[32] Gavai, N.R.; Jakhade, Y.A.; Tribhuvan, S.A.; Bhattad, R. “MobileNets for flower classification using TensorFlow,” In Proceedings of the 2017 International Conference on Big Data, Iot and Data Science, Pune, India, 20–22 December 2017; pp. 154–158.

[33] Srinivasu, P.N.; SivaSai, J.G.; Ijaz, M.F.; Bhoi, A.K.; Kim, W.; Kang, J.J. Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. Sensors 2021.

[34] S. K. Maria, S. S. Taki., M. J. Mia, A. A. Biswas., A. Majumder, F Hasan, “Cauliflower Disease Recognition Using Machine Learning and Transfer Learning,” In: Somani A.K., Mundra A., Doss R., Bhattacharya S. (eds) Smart Systems: Innovations in Computing. Smart Innovation, Systems and Technologies, vol 235. Springer, Singapore.

[35] Jovic, A.; Brkic, K.; Bogunovic, N. A Review of Feature Selection Methods with Applications. In Proceedings of the 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 25–29 May 2015; pp. 1200–1205.

[36] Ding, H.; Feng, P.-M.; Chen, W.; Lin, H. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol. BioSyst. 2014, 10, 2229–2235.

[37] Bakar, Z.A.; Ispawi, D.I.; Ibrahim, N.F.; Tahir, N.M. Classification of Parkinson’s Disease based on Multilayer Perceptrons (MLPs) Neural Network and ANOVA as a Feature Extraction. In Proceedings of the 2012 IEEE 8th International Colloquium on Signal Processing and its Applications, Melaka, Malaysia, 23–25 March 2015; pp. 63–67.

[38] Kim, H.-Y. Analysis of variance (ANOVA) comparing means of more than two groups. Restor. Dent. Endod. 2014, 39, 74–77.