National College of Ireland

Classification of Various Diseases in the Mango Crop Using Machine Learning

MSc Research Project
Data Analytics

Kavyasree Panuganti
Student ID: x23219360

School of Computing
National College of Ireland

Supervisor: Hamilton Niculescu

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name**:**	Kavyasree Panuganti		
Student ID:	x23219360		
Programme:	Data Analytics	**Year:**	2024-2025
Module:	MSc Research Project		
Supervisor:	Hamilton Niculescu		
Submission Due Date**:**	12/12/2024		
Project Title:	Classification of Various Diseases in the Mango Crop Using Machine Learning		
Word Count:	**6400**	Page Count	**20**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Kavyasree Panuganti

Signature: ……

12/12/2024

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Classification of Various Diseases in the Mango Crop Using Machine Learning

Kavyasree Panuganti
x23219360

Abstract

Mangoes are one among the most popular fruits which are widely popular for its taste and the nutritional values. Because of this huge demand mangoes are cultivated in vast areas across the world however these plants are affected by diseases due to pests and insects which will affect the quantity and the quality of the mangoes produced. Traditionally they are detected by checking them manually, which takes lot of time due to the manual labour work . With the rise in global demand for mangoes, the need for a fast, efficient, and automatic disease detection system has become critical. To achieve this the research is done to develop a robust hybrid framework leveraging deep learning techniques, such as Convolutional Neural Networks (CNN) and transfer learning, combined with traditional machine learning classifiers like Random Forest, Support Vector Machines (SVM), and XGBoost. The proposed approach in this research uses the pre-trained CNN architectures, including VGG16(Visual Geometry Group with 16 layers) and VGG19(Visual Geometry Group with 19 layers) for feature extraction. K-means clustering is employed for doing the image segmentation to segment the mango leaf images into different regions based on pixel intensities. After extracting the features from CNN these are passed to traditional classifiers for disease classification which are evaluated on a mango leaf dataset containing 4000 images. The hybrid approach has performed well when compared to standalone CNN models. Among all the models Random Forest using VGG16 features has delivered the highest validation accuracy of 97.75%, while SVM and XGBoost models gave the competitive results with accuracies of 96.5% and 95.75%.

1 Introduction

Mango is one of the most important commercial crops in the world as they contribute to both food security and economic growth. Mangos are produced by a tropical tree known as Mangifera indica which is originated from Myanmar, Bangladesh and Northeastern India, mangoes have been cultivated for over 4000 years making them an integral part of cultural and agricultural history. Over time, the cultivation of mangoes has been expanded throughout Asia, including Philippines, China and Indonesia and subsequently to other parts of the world (Saúco, V. 2004). Because of the huge popularity farmers have begun to cultivate the mangoes as commercial acreage. But now a days farmers are facing a lot of issues in maintaining the good quality and high yield of the crop because of the uncontrollable pests and the various diseases occurred to the plants (Misra, A.K. 1992). These diseases can be classified into those caused by fungi, bacteria, nematodes and different insect pests like mealy bugs, fruit flies and hoppers and diseases like sooty mould, powdery mildew, and anthracnose etc as shown in the table1, which reduces the yield and quality of fruits. These crop diseases pose a significant challenge to the agriculture industry, particularly in the regions where the mango cultivation plays an important role in the economy.

Table 1: Various diseases in Mango leaves

Disease Name	Symptoms
Anthracnose	Black spots and patches on leaves.
Powdery Mildew	Powdery fungal coating on leaves.
Red Rust	Orange-red spots on leaves
Sooty Mold	Black fungal growth on leaf surfaces.
Leaf Gall	Abnormal leaf growth or swellings.
Bacterial Leaf Spot	Small, water-soaked marks turning into black spots.
Black Rot	Dark, necrotic patches on leaves and stems.
Scab	Grayish lesions on leaves causing deformities.
Algal Spot	Greenish to reddish-brown patches on leaves.
Verticillium Wilt	Yellowing and wilting of leaves.

Traditionally farmers used to depend on the advice of the experts in detecting the mango diseases, but this process takes a lot of time and labour work for the large farms and checking each plant manually may led to many human errors. Also, as these experts are not available all the time farmers started depending on the pesticide shops for their help, but these people may sell higher margin pesticides regardless of the specific disease, causing the crop loss. Farmers need to detect the disease in the early stage to reduce agricultural losses, improving the crop yield and ensuring food security (Yahia, 2011). In the study done by De Luna (2019) has created an automated system that uses computer vision to detect and classify plant diseases. The setup includes sensors inside a motorized image-capturing box that tracks plant growth and helps in spotting the diseases. However, this system has various drawbacks, such as being costly and needing a trained expert to operate the box. To address these challenges this research uses machine learning techniques like Convolutional Neural Networks (CNNs), transfer learning, and image segmentation methods like K-Means clustering. It also evaluates models like Support Vector Machines (SVM), XGBoost, and Random Forest to check if the solution is both efficient and reliable.

Convolutional Neural Networks (CNNs) are type of deep learning algorithms that are very effective in analyzing visual data which are suitable for mainly image classification problem since they can automatically extract spatial hierarchies of features from input images. These are made up of various layers such as convolutional layers, fully linked layers, pooling layers. By using these CNNs the model can increase its ability to learn difficult patterns and features from the images of mango leaves, so that it can accurately detect the diseases and classify them. Along with CNNs architectures like VGG19, VGG16, and ResNet are used for feature extraction, and leverage transfer learning to fine-tune pre-trained models on the dataset consisting of mango leaves. To change appearance of image and make it more meaningful, segmentation techniques such as K-means clustering are used to divide a single image into several segments.

The research question posed in this study is:

Can advanced machine learning techniques, including CNN architectures and image segmentation methods, improve the accuracy and efficiency of mango crop disease classification?

To address this question, the following specific objectives were defined:

1. Investigate the current state-of-the-art methods for crop disease detection and classification.
2. Design a classification framework utilizing CNN architectures (e.g., VGG16, VGG19, ResNet) and advanced segmentation techniques.
3. Implement and compare the performance of machine learning models (SVM, XGBoost, Random Forest) with CNN-based approaches.
4. Evaluate the proposed methods using metrics such as F1 score, precision, accuracy, and recall.

This document is structures as:

Section 1: Introduction
This section introduces the study, discussing the significance of mango cultivation, the challenges posed by diseases and pests, and the aim to utilize machine learning for disease detection.

Section 2: Related Works
This section reviews previous research on plant disease diagnosis, focusing on both traditional and machine learning-based approaches in agriculture.

Section 3: Methodology
This section outlines the machine learning models, and image segmentation techniques used in the study, including CNNs, K-Means clustering.

Section 4: Design
This section describes the design of the system, including the architecture of the models and dataset used.

Section 5: Implementation
This section explains the implementation process, covering data preprocessing, model training, and integration of techniques like transfer learning.

Section 6: Evaluation and Results
The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score. Results are compared across different models.

Section 7: Conclusion and Future Work
This section summarizes the findings, discusses the study's implications, and suggests future research directions.

2 Related Work

For conducting any research, it is very important to highlight the earlier studies that are conducted. Recent advancements in machine learning (ML) and deep learning (DL) have provided the path for automated, efficient, and accurate methods of plant disease detection. This literature review explores key methodologies, focusing on the use of DL, hybrid approaches, image processing techniques, and transfer learning in agricultural applications.

2.1 Deep Learning frame works such as CNN and transfer learning Approaches for Disease Detection

CNN has become the powerful tools for recognising the images and extracting the important features which helps in the detection of diseases in the plant in a more automatic manner. In the work done by Fuentes et al. (2017) shows that the models developed by these deep learning algorithms have achieved the remarkable performance on ImageNet and other datasets. Hughes et al. (2016) has used the pretrained CNN in his work where the focus was training the last few layers only. They have applied these to the coloured datasets which gave the good accuracy of 99% but the point to be noted in his research is the model was not performing well on the other datasets and Ferentinos (2018) has overcome this situation by merging the actual world images with already present diseases crop images to enhance the practicability that was missing in the case of Hughes. Gulavnai and Patil (2019) proposed Resnet-CNNs coupled with Transfer learning techniques for dectecting the disease automatically and identified 4 mango leaf diseases namely red rust and golmich, anthracnose, powdery mildew. The results showed that the ResNet50 gave better performance than the others ResNet-CNNs such as (ResNet18, ResNet34 and ResNet50) with 91.50 percent of accuracy. Inspired by VGG-16, VGG-19, and AlexNet, the authors Arivazhagan and Ligi (2018) have developed a CNN model with three hidden layers that could recognise five distinct mango leaf illnesses such as leaf burn, leaf gall, leaf webber, Alternaria leaf spots, and anthracnose. The suggested model classified diseased photos with an accuracy of 96.67%. However, by adding more photos to the dataset and adjusting the CNN model's parameters, the classification accuracy can be raised even higher. Wongsila et al. (2021) has suggested a CNN model built on the AlexNet architecture to identify the anthracnose leaf disease. The Tensor Flow framework and a dataset of real-world mango photos taken with a CDD camera were used to construct the system. More than 70% of the diseased mango leaves could be identified using the established technique.

Table 1: Summarizing the contribution done for classifying the diseases using deep learning techniques

Author(s)	Year	Approach	Key Contributions
Fuentes et al.	2017	Deep Learning Models on ImageNet and other datasets	Demonstrated deep learning models high performance on benchmark datasets like ImageNet.
Hughes	2016	Pretrained CNN focusing on training only the last few layers	Highlighted performance drop when tested on unseen datasets.
Ferentinos	2018	Merging real-world images with existing diseased crop images	Improved model applicability by incorporating real-world datasets.
Gulavnai and Patil	2019	ResNet-CNNs coupled with transfer learning	Identified 4 mango diseases; ResNet50 performed better than ResNet18 and ResNet34.
Arivazhagan and Ligi	2018	CNN model inspired by VGG-16, VGG-19, and AlexNet	Achieved high accuracy highlighted the need for more data and parameter tuning to improve performance.
Wongsila et al.	2021	AlexNet-based CNN using TensorFlow framework	Used real-world mango images captured by a CDD camera for disease detection.

2.2 Hybrid Approaches

In the hybrid approach the CNN based feature extraction is done and then these extracted features are used with the traditional classifiers such as Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost) and Random forest in order to efficiently detect the disease in the mango plants. A system for classifying mango diseases, specifically, anthracnose, bacterial black spot and sooty mold was presented by Prabu and Chelliah (2022).For improving the extraction of featues they have used a CNN with crossover-based levy flight distribution, and they have used MobileNetV2 model in the learning stage and SVM for diseases classification. The experimental results of their research show that the classification performances over other state-of-art methods. In the work by Sutrodhor et al. (2018), have constructed a MLAD (mango leaf ailment detection) based on Neural Network and Support Vector Machine (SVM) classifiers. MLAD can detect and atomically classify four diseases, namely scab, anthracnose, Red Rust and Sooty Mold, with an average accuracy of 80%. Authors Ramcharan et al. (2017) have incorporated the pretrained CNN models by training only the few of the last layers by using their dataset and then these features fed to the SVM for the classification. This combination has provided the accuracy of 92 percent.

For instance, Banerjee and Pamula (2020) evaluated multiple approaches for plant classification by using the PlantCLEF-2019 dataset. In their study they have compared the standalone CNN approach, CNN combined with AdaBoost, SVM, and their proposed hybrid model CNN combined with Random Forest. The standalone CNN model achieved an accuracy of 56%, while the inclusion of AdaBoost improved the accuracy to 59.6%. The SVM-based approach performed slightly better than the standalone CNN, achieving 58% accuracy. However, their hybrid CNN + Random Forest model outperformed all other methods, achieving a notable accuracy of 67%. In a study by Bashir et al. (2024) introduced a hybrid CNN-XGBoost model to improve the accuracy of DC power forecasting for photovoltaic (PV) systems. Validated on a 2.88 kW grid-connected PV system, the model achieved superior performance with an RMSE of 44.18 and R^2 of 0.996, halving the RMSE compared to traditional models

Table 3: Summarizing the contribution done for classifying the diseases using hybrid approaches

Author(s)	Year	Method	Key Findings
Prabu and Chelliah	2022	CNN with crossover-based levy flight distribution, MobileNetV2, and SVM	Classified mango diseases (Bacterial black spot, Anthracnose, Sooty mold) with improved performance over state-of-the-art methods.
Sutrodhor et al.	2018	MLAD framework using Neural Network and SVM	Detected four diseases with an average accuracy of 80%.
Ramcharan et al.	2017	Pretrained CNN models with features fed into SVM	Achieved a classification accuracy of 92%.
Banerjee and Pamula	2020	Standalone CNN, CNN + AdaBoost, SVM, CNN + Random Forest (hybrid model)	CNN + Random Forest achieved the best accuracy of 67%, outperforming other methods like SVM (58%) and standalone CNN (56%).
Bashir et al.	2024	Hybrid CNN-XGBoost model	Improved DC power forecasting with RMSE of 44.18 and R^2 of 0.996, halving RMSE compared to traditional models.

2.3 Image Processing Techniques

It is very important to identify the diseased leaf to save plant production. For the proper identification the image processing methods are applied to identify the unhealthy plant leaf
In the work done by Srunitha and Bharathi (2018) they have developed a method for automatic mango leaf diseases recognition and classification in which they have used k-means clustering for image segmentation, GLCM (gray level color co-occurrence metrics) for feature extraction and SVM for diseases classification. SVM classifier gets accuracy up to 96% but one of the model in their research made difficulty in the image segmentation as the leaf had several diseases in the same region. In the research done by Singh and Misra (2017) they have used

image processing and various segmentation methods to study the plants health. The segmentation approach is based on a number of features identified in the image, suggesting that the data may be based on their structure or borders, surface deformities, or color discrepancies. Abdu et al. (2019) employed machine learning methods to identify plant diseases by analyzing photos taken by long-range cameras, such as drones or even satellite imagery. By focusing on the leaf's area of interest and eliminating the surrounding area, segmentation of photos is accomplished, making it simple to use for improved outcomes. The SVM technique is used in this study as a classifier to identify plant leaf diseases.

Table 4: Demonstration of work done in various image segmentation techniques

Author(s)	Year	Methodology	Results/Findings	Challenges
Srunitha & Bharathi	2018	K-means clustering for image segmentation. GLCM for feature extraction. SVM for classification.	Achieved up to 96% accuracy using the SVM classifier.	Difficulty in segmenting images when multiple diseases are present in the same region of the leaf.
Singh & Misra	2017	Image segmentation based on structural features, surface deformities, and color discrepancies.	Identified multiple features in images to assess plant health effectively.	NA
Abdu et al.	2019	Image segmentation focused on the leaf's area of interest by removing surrounding areas.	Improved outcomes by isolating the leaf's area of interest, making the process effective and straightforward.	Challenges in segmentation for diverse leaf complexities in real-world scenarios.

The specific problem identified in the literature is the lack of a reliable and accurate system for detecting and classifying the mango diseases. Where most of the models used standalone CNN models to classify the disease which performed on specific datasets but was unable to perform on the new datasets and few other studies have performed on traditional machine learning methods like SVM or Random forest alone in detecting the disease in which they have issues in detecting the diseases when there are multiple diseases in single leaf. To overcome these issues in this project I have used the Hybrid model which uses pretrained CNN models like VGG16 and VGG19 in extracting the features and integrated these with the traditional machine learning methods like SVM, RandomForest and Xgboost and compared those results providing a detailed analysis of their strengths and weaknesses. The data was handled in a proper way by doing the preprocessing and augmentation techniques and the Image segmentation is done by using the K-Means clustering to group the image pixels into clusters which can reduce the segmentation challenges.

3 Research Methodology

This research aims to classify the mango crop diseases using the advanced machine learning techniques such as Convolutional Neural Network (CNN) method along with transfer learning techniques. Kdd (Knowledge Discovery in Databases) approach is used in this research as it involves the series of systematic steps to conduct effective research as shown in figure.

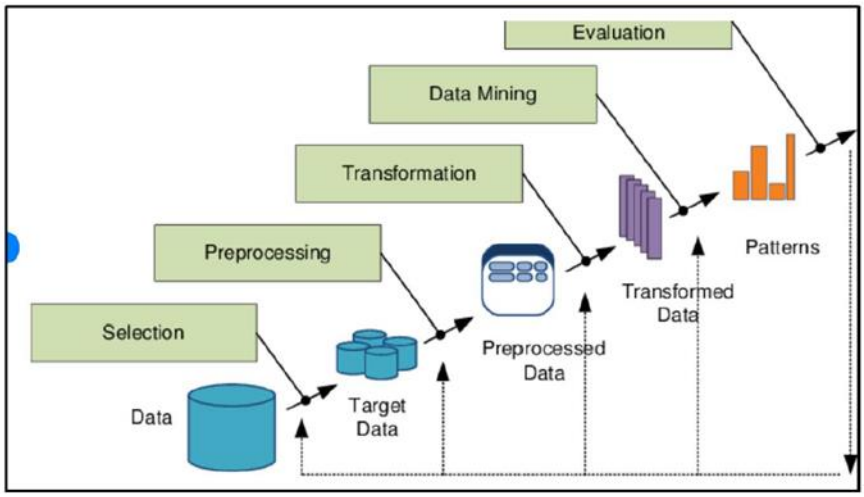


Figure 1: KDD Approach

3.1 Selection of the data

The dataset used in this dataset is collected from the public repository. The dataset includes:

- **Classes:** Dataset has the images of both healthy leaves and those affected by various diseases such as powdery mildew, anthracnose, sooty mould, Bacterial Canker, Cutting Weevil, Die Back, gall midge as shown in the figure 2.
- **Number of Images:** Dataset contains around 4000 images.
- **Categorization:** The data is divided and stored into various folders, each representing a different disease.

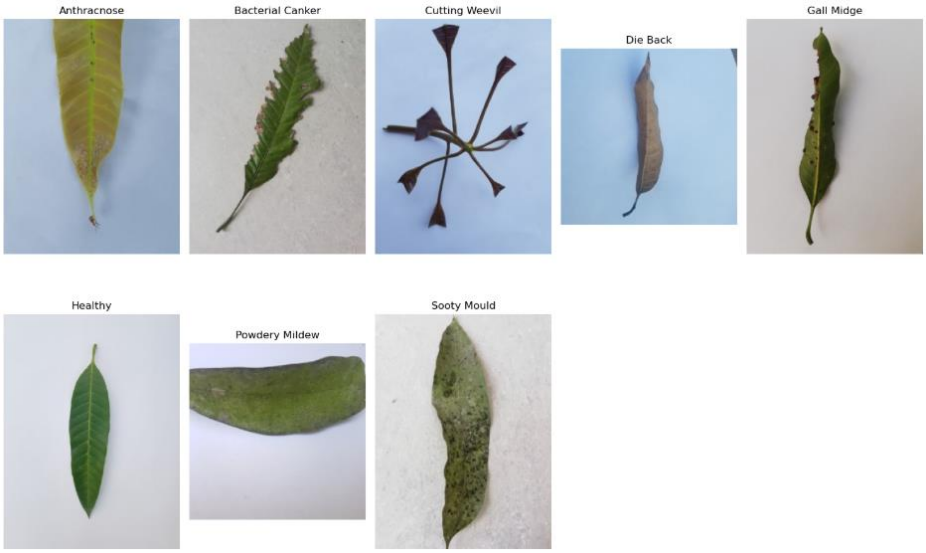


Figure 2: Various images present in the Dataset

3.2 Preprocessing of the data

Preprocessing of data is an essential step to prepare the original images so that they can be used for the machine learning models. The following preprocessing steps are employed:

1. **Image Resizing:** All images in the dataset are resized to 224x224 pixels. This will make sure that they are consistent and compatible to the model.

2. **Normalization:** Each pixel value of the images is normalized to a range between [0, 1] by dividing by 255. Normalization helps in stabilizing and speeding up the training process by ensuring all input values have the same scale.

3. **K-Means Clustering for Image Segmentation:**
   K-Means clustering is used to divide the leaf images to separate the diseased areas from the background which make the process easier to extract important details by highlighting the patterns in the diseased areas.
   - The images are converted into a format where each pixel was treated as a separate data point and K-Means is applied to group the pixels into clusters, such as healthy areas and diseased regions. Later the background is removed to focus on the features of the leaf.

4. **Label Encoding:** It is used to convert disease classes into numeric values, to make them understand to machine learning algorithms

5. **Data Augmentation Techniques:**
   To increase the dataset's size and variability data augmentation is applied to the existing data. Below are the specific techniques applied:
   - **Rotation:** Images were randomly rotated up to 20 degrees to mimic different angles when photos are taken,
   - **Width and Height Shifts:** Images were shifted horizontally or vertically by up to 20% helping the model to recognize leaves even if they're not perfectly centered.

- **Zoom:** A zoom range of 20% is applied to simulate images taken at different distances
- **Horizontal Flip:** Images are randomly flipped horizontally to create mirror versions of the existing data
- **Fill Mode:** Empty spaces that are created during the transformations are filled with nearby pixel values.

3.3 Transformation of Data

In the transformation phase, the pre-processed images were prepared for modelling by:

1. **Conversion to NumPy Arrays:** Each image is converted into a NumPy array, enabling it to be compatible with the deep learning frameworks like TensorFlow and Keras that are used. NumPy arrays also facilitate easy manipulation and efficient data storage.

2. **Scaling for Model Input:** Images are resized to 224x224 pixels with 3 color channels (RGB) to match the input requirements of pre-trained CNN models, ensuring consistent input structure and effective use of pre-trained weights.

3.4 Integration of suitable Data Mining techniques

To determine the most effective model, multiple models are developed and evaluated those models based on their performance and accuracy. The first approach of the research is to build deep learning models based on the Convolutional neural networks (CNNS) such as VGG16, VGG19, and ResNet50 which are implemented using transfer learning. These pretrained models are originally trained on the ImageNet dataset which are used for their strong feature extraction capabilities. The convolution layers of these models are frozen and new fully connected layers are trained specifically for classifying the diseases in the mango leaves. By retaining the learned features of the based model and retraining only the final few layers computational demands are reduced.

The second approach involves the traditional machine learning techniques such as Support Vector Machine (SVM), Random Forest, and eXtreme Gradient Boosting (XGBoost) trained on deep features extracted from the CNN models. The goal is to verify if these traditional models, when used with the features derived from the pre-trained CNNs could match or exceed the performance of CNNs itself in turns of disease classification accuracy.

3.5 Interpretation or evaluation of the results

For evaluating the model, metrics such as Accuracy, precision, Recall and F1-score is used. In addition to these metrics' confusion metrics are also used.

Accuracy: Accuracy is defined as the proportion of the correctly predicted instances to the total number of instances in the dataset.

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN}$$

Prediction: This is also known as positive prediction value which is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$PRECISION = \frac{TP}{TP + FP}$$

Recall: Recall is the ratio of correctly predicted positive observations (True positives) to that of all positive observations.

$$RECALL = \frac{TP}{TP + FN}$$

F1-Score: It is the harmonic mean of precision and recall.

$$F1\ SCORE = \frac{1}{\frac{1}{PRECISION} + \frac{1}{RECALL}}$$

Confusion matrix: Confusion matrix is a table which evaluates true values versus the predicted values in a matrix format as shown in figure 2.

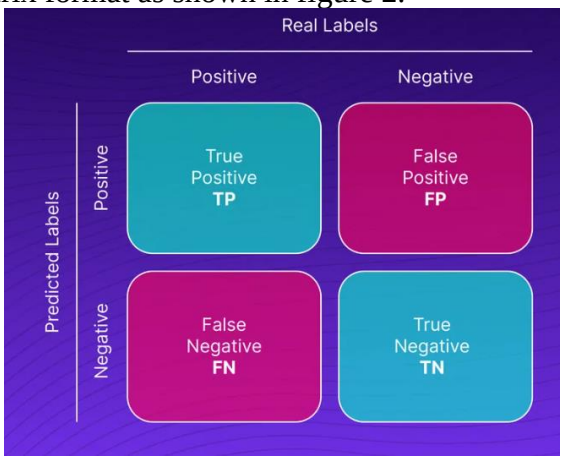


Figure 2: Confusion matrix

4 Design Specification

In this section, the model designing and the architecture of all the components that are used to detect diseases is explained. The architecture follows a 2-tier layout. To develop the models Python 3.9.13 is used as the programming language because of its wide range of libraries such as Keras and TensorFlow. Also to analyze and visualize the results matplotlib library is used. The entire programming such as data preprocessing, transformation, and model implementation are conducted using Jupiter Notebook.
The models that are implemented in this study are as follows:

4.1 Convolutional Neural Network (CNN)

The first model that is implemented is a basic convolutional neural network (CNN), which is a deep learning algorithm that automatically learn the features from the provided input images. The CNN model in this study is used to classify mango leaf diseases by assigning weights and biases to different regions of the images. The CNN has the several layers like the convolutional layer, pooling layer, fully connected layer, and activation functions that are used in this research for classifying the diseases in mango leaves. The function of each layer is explained detailly in this study below:

- **Convolution Layer:** The convolution layer is the main building block of the CNN. This layer takes the input image into RGB channels and then applies the filters to produce a feature map. This feature map helps in finding the basic features such as edges and curves in the image and the output of this layer is given to the next layer as input.

- **Activation Function:** After each convolution, a non-linear activation function is applied to increase the network's ability to capture complex features. In this research, the Rectified Linear Unit (ReLU) is used, which changes all negative values to zero, adding non-linearity to the model and helping it learn complex patterns in the data.
- **Pooling Layer:** By using this layer, the spatial size of the feature maps can be reduced which in turn reduces the computational requirements and reduces the problem of overfitting.
  In this research Max pooling is used to retain the maximum value of the specified region without removing the important features.
- **Fully Connected Layer:** The Fully Connected Layer (FC) will make the final classification by finding the high-level features that are correlated with specific classes of mango diseases and the output is then converted in to single dimensional vector using a flattening process. After the conversion the output is passed through dense layers.
- **Dropout Function**:
  Dropout function is used to avoid the overfitting.
- **Softmax Function**: Finally, the last layer we use is a **Softmax layer** for the classification of different diseases .This layer helps in the probability distribution across all the classes, ensuring that the model is trained to perform multi-class classification for mango disease identification.

4.2 Transfer Learning

Transfer Learning is employed in this study to leverage the power of pre-trained deep learning models. This technique will use an existing model, which is pretrained on the large datasets and using it for the current problem of mango disease classification.by using the transfer learning the computational burden and training time can be reduced since the feature extraction layers are already well trained to recognize the common factors.

- **VGG16 and VGG19**: VGG16 and VGG19 which are applied in the study is introduced by the Simonyan and Zisserman (2015), this model will use the stacked 3x3 convolutional layers and max pooling to extract the meaningful features from the images. For the classification of disease in the mango leaves only the last few fully connected layers are retrained while the rest of the model remains unchanged.in VGG16 and VGG19 the numbers 16 and 19 indicate the total weight of the layers in those respective models.Thus in this research these architectures are used for the proper image recognition in disease detection of the plant leaf.
- **ResNet**: ResNet is a deep convolutional neural network that addresses the vanishing gradient problem through the use of residual learning.In this model ResNet50 (with 50 layers) is employed to extract robust features for mango leaf disease classification. Only the fully connected layers are retrained while the rest of the model remains frozen.

4.3 Support Vector Machines (SVM)

SVM is one of the supervised learning algorithms which is widely used in the classification problem. In this research SVM algorithm is used to classify the diseases in the mango leaf based on the deep features that are extracted using the CNN models in the above stage. Each mango leaf image is represented as a point in n-dimensional space where n is the

number of features, and SVM attempts to find the optimal hyperplane that separates different classes.

4.4 Random Forest

Random Forest is an ensemble learning technique that works by creating multiple decision trees. In this research random forest uses the features extracted from the above CNN models to classify the diseases tin the mango leaves.By using the Random forest overfitting can be reduced by averaging the outputs of many trees.

4.5 eXtreme Gradient Boosting (XGBoost)

XGBoost is an open-source library that provides high-performance implementation of gradient-boosted decision trees. This method builds an ensemble of trees in an iterative fashion, where each new tree aims to correct the errors of the previous one. XGBoost has been applied to the extracted features from mango leaf images, leveraging the boosting approach to effectively reduce errors and provide better classification accuracy. This approach works by training successive models until improvements plateau, thereby producing an optimal model for classification.

5 Implementation

5.1 Environment:

The models in this research is done by using python 3.7 in the jupyter Notebook IDE where the deep learning models are implemented using TensorFlow 2.x and Keras. Keras provide a user-friendly interface which makes it easy to use the tensor flow functions and model building process. For image processing tasks such as resizing and normalizing input images OpenCV is used, while Matplotlib is used for visualization purposes.

```
TensorFlow version: 2.17.0
Num GPUs Available: 0
OpenCV version: 4.10.0
```

Figure 3: Specifications

5.2 Handling of Data:

As mentioned in the subsection 3.1 the dataset used in this model is comprised of 4000 images of both healthy and diseased. In order to prepare the data Image processing was the first step mentioned in the subsection 3.2 where all images were resized to a consistent dimension of 224x224 pixels. As a next step normalization was applied to scale the pixel values to a range between 0 and 1 by dividing each pixel value by 255.Then the disease classes were converted to numeric labels as shown in the table 5. These numeric labels were assigned during the label encoding process to dataset

Table 4: Shows the label encoding for the diseases

Disease Class	Numeric Label
Powdery Mildew	0
Anthracnose	1
Sooty Mould	2
Bacterial Canker	3
Cutting Weevil	4
Die Back	5
Gall Midge	6
Healthy	7

To improve the datasets variability and to avoid the risk of overfitting data augmentation techniques are applied as mentioned in the subsection 3.2.as a most important step in data handling involved K-Means clustering for image segmentation, aimed at isolating the disease-affected regions from the healthy portions of the leaves. The K-Means clustering algorithm was applied to group the image pixels into distinct clusters based on their intensity levels, representing diseased areas, healthy leaf regions, and background noise. The segmentation is configured into 3 clusters corresponding to diseased regions, healthy leaf areas, and background noise. The segmentation process used the Euclidean distance to measure how similar the pixel colours were. From the below figure we can see the segmented images that clearly highlighted disease-specific regions on the mango leaves.

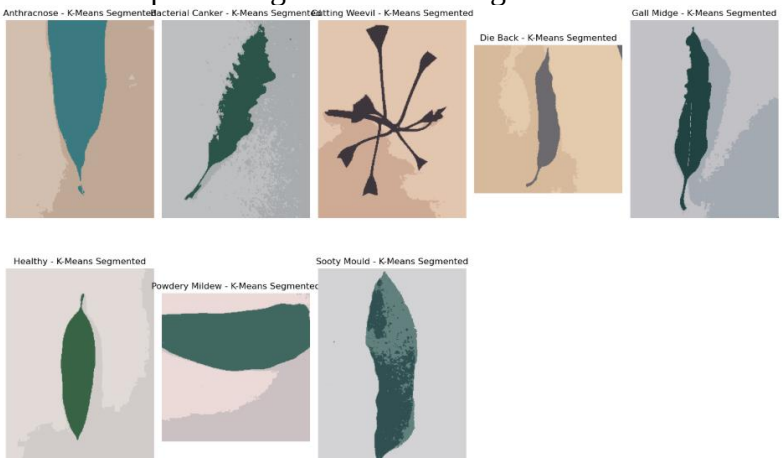


Figure 4: Segmented Images

5.3 Architecture:

As mentioned in subsection 4.1, the architectures of all the models built for classifying mango crop diseases are described below. The study implements a variety of models, including a basic Convolutional Neural Network (CNN), transfer learning with pre-trained models (VGG16, VGG19, ResNet50), and traditional machine learning classifiers (Random Forest, SVM, and XGBoost). These models are tailored to efficiently detect and classify mango leaf diseases from input images.

5.3.1 Convolutional Neural Network (CNN)

The first model implemented is a basic CNN which automatically learns features from input images by assigning weights and biases to different areas of the image. To implement the model packages such as Conv2D, MaxPooling2D, Activation, Flatten, Dropout, Dense, and others are imported from the TensorFlow library using Keras.The CNN implementation in this project used pre-trained architecture (VGG16, VGG19, or ResNet50) as the base for feature extraction, leveraging their pre-learned convolutional layers. Custom layers, including a Flatten layer, a Dense layer with ReLU activation, a Dropout layer, and a final Dense layer with softmax activation, are added for classifying mango leaf diseases and the pre-trained layers are frozen, focusing training only on the new layers to reduce computational costs and improve efficiency.The CNN model is trained with the following hyperparameters:
Batch size: 32, Learning rate: 0.001, Epochs: 10, Optimizers: Adam.

5.3.2 Transfer Learning

As mentioned in subsection 4.2, transfer learning leverages pre-trained neural network models to reduce computational costs and training time. In this study, transfer learning is implemented using VGG16, VGG19, and ResNet50, which are pre-trained on the ImageNet dataset containing millions of images across 1000 classes.
These models are adapted for mango leaf disease classification by:

1. Freezing the convolutional base layers to retain learned features.
2. Adding custom dense layers to classify diseases into multiple classes.
3. Using softmax activation as the classifier for multi-class classification.

Parameters for transfer learning models:

- **Input shape:** 224x224x3
- **Dense layers:** One fully connected layer with 128 neurons, ReLU activation, and a dropout layer to prevent overfitting.
- **Final layer:** Softmax activation for disease classification.
- **Optimizers:** Adam with a learning rate of 0.001.

5.3.3 Random Forest

As mentioned in subsection 4.4, Random Forest is an ensemble learning method used for performing classification. In this study, the RandomForestClassifier from the sklearn.ensemble library is used and the features that are extracted from the pretrained models is used in the training.

5.3.4 Support Vector Machine (SVM)

As mentioned in subsection 4.3, SVM is a supervised learning algorithm used for classification tasks. In this study, the svm.SVC module from the sklearn library is used and the features that are extracted from the pretrained models is used in the training.In this model Linear kernel is used for improved performance and the model is evaluated using an 80:20 train-test split.

5.3.5 eXtreme Gradient Boosting (XGBoost)

As mentioned in subsection 4.5, XGBoost is a gradient-boosted decision tree algorithm that improves model accuracy by iteratively correcting classification errors. In this study, the XGBClassifier from the xgboost library is used and the features that are extracted from the pretrained models is used in the training. In this research this model is evaluated against other models for its ability to classify mango leaf diseases accurately.

6 Evaluation

In this section the performance of the models such as Random Forest, XGBoost and SVM are evaluated on classifying and detecting diseases in the mango plants. These models are trained on the feature representations that are derived from the best performing deep learning models such as deep learning models such as VGG16, VGG19, ResNet50.To evaluate these models' various metrics like accuracy of the classification, precision and recall metrics are considered.

6.1 Performance of Deep Learning Models

1. **VGG16 Model**

   The VGG16 model performed very well with the steady decrease in the loss from 1.8264 to 0.2730 in the last epoch as shown in the figure 7 achieved a training accuracy

of 89.95% and a validation accuracy of 94.50% after 10 epochs. This progression in the model reflects the model's convergence and ability to generalize well on the validation data. Also, from the confusion matrix that's developed we can observe that diseases like Anthracnose and Powdery Mildew, are classified wrongly for some samples. For Anthracnose, model was able to identify the 37 samples correctly out of 50 , and 5 samples are wrongly identified as Die Back and 2 as Bacterial Canker.

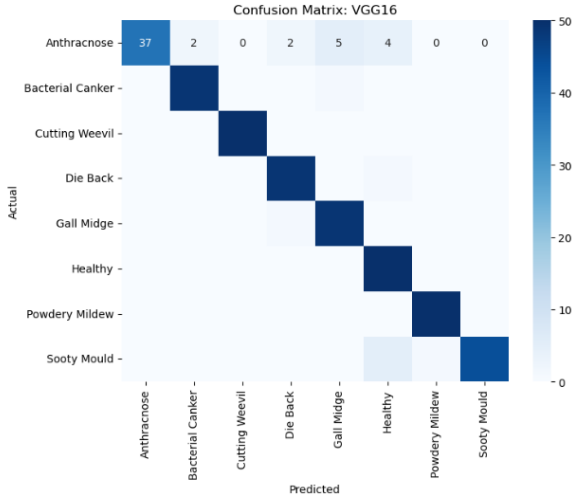


Figure 5: Confusion Matrix for the VGG16 Model

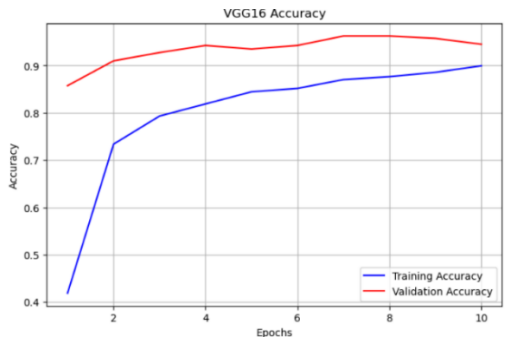


Figure 6: Plots for VGG16 Model Accuracy

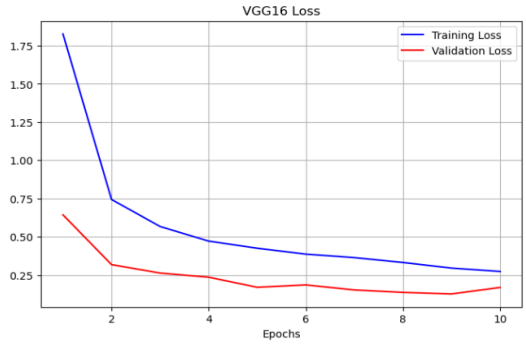


Figure 7: Plots for VGG16 Model Loss

2. **VGG19 Model**:

The VGG19 model has performed well outperforming the VGG16 with validation accuracy of **96.75%**, outperforming VGG16 but the training accuracy was slightly lower at 73.54% which indicates that the model is slightly overfitting. From the figures 9, 10 we can observe that that the loss is gradually decreased and the accuracy is increased over the 10 epochs similar to VGG16.The confusion matrix for VGG19 shows the improved classification performance particularly to Anthracnose disease classifying 44 out 50 samples correctly. Yet 2 samples are misclassified as **Die Back** and 1 each as **Gall Midge** and **Powdery Mildew**.

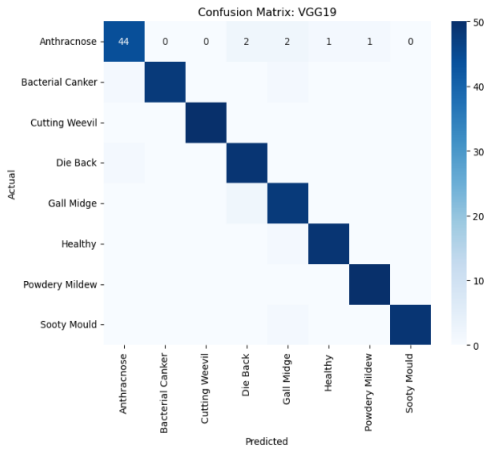


Figure 8: Confusion Matrix for the VGG19 Model

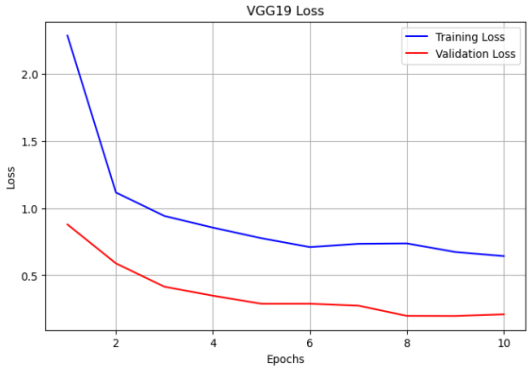


Figure 9: Plots for VGG19 Model Accuracy

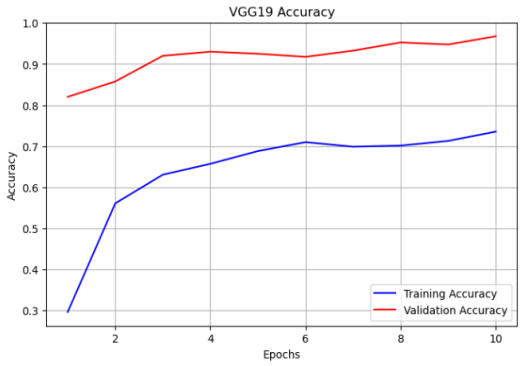


Figure 10: Plots for VGG16 Model Loss

3. ResNet50 Model

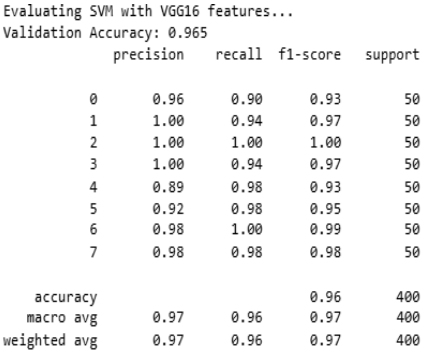
The ResNet model did not perform well with least validation accuracy of 12.5% which is very less when compared to VGG16 and VGG19. Therefore, features produced from the ResNet model are not used in further training process for traditional models.

```
VGG19 training completed and model saved!
Training ResNet50...
Epoch 1/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 423s 4s/step - accuracy: 0.1353 - loss: 4.1053 - v
al_accuracy: 0.1250 - val_loss: 2.0795
Epoch 2/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 402s 4s/step - accuracy: 0.1199 - loss: 2.0987 - v
al_accuracy: 0.1250 - val_loss: 2.0795
Epoch 3/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 381s 4s/step - accuracy: 0.1171 - loss: 2.0797 - v
al_accuracy: 0.1250 - val_loss: 2.0795
Epoch 4/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 390s 4s/step - accuracy: 0.1174 - loss: 2.0797 - v
al_accuracy: 0.1250 - val_loss: 2.0795
Epoch 5/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 381s 4s/step - accuracy: 0.1174 - loss: 2.0796 - v
al_accuracy: 0.1250 - val_loss: 2.0794
Epoch 6/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 384s 4s/step - accuracy: 0.1083 - loss: 2.0796 - v
al_accuracy: 0.1250 - val_loss: 2.0794
Epoch 7/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 383s 4s/step - accuracy: 0.1062 - loss: 2.0796 - v
al_accuracy: 0.1250 - val_loss: 2.0794
Epoch 8/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 375s 4s/step - accuracy: 0.1077 - loss: 2.0796 - v
al_accuracy: 0.1250 - val_loss: 2.0794
Epoch 9/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 376s 4s/step - accuracy: 0.1031 - loss: 2.0796 - v
al_accuracy: 0.1250 - val_loss: 2.0794
Epoch 10/10
100/100 ━━━━━━━━━━━━━━━━━━━━ 376s 4s/step - accuracy: 0.1045 - loss: 2.0796 - v
al_accuracy: 0.1250 - val_loss: 2.0794
```

Figure 11: Training Summary of ResNet model.

6.2 Performance of Hybrid Models

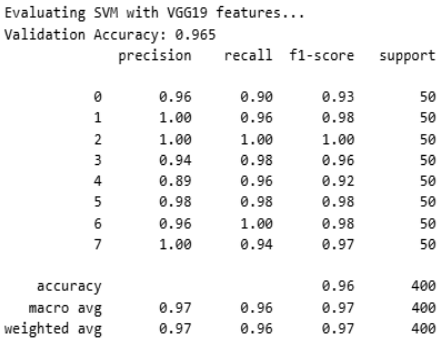
1. **SVM (VGG16 and VGG19 Features)**
   SVM model when combined with the features of VGG16 performed very well with the accuracy of 96.5% which shows that it is very capable in handling the complex classification tasks very effectively. The other metrics of this model such as precision and recall are 97% and 96.5% which showcases the well balanced performance of this model in minimizing false positives and false negatives.
   When SVM is used with VGG19 features it maintained the same accuracy of 96.5% and the precision was slightly reduced to 96.85% .This consistency across different feature shows the SVM's versatility as a classifier.



Evaluating SVM with VGG16 features...
Validation Accuracy: 0.965

	precision	recall	f1-score	support
0	0.96	0.90	0.93	50
1	1.00	0.94	0.97	50
2	1.00	1.00	1.00	50
3	1.00	0.94	0.97	50
4	0.89	0.98	0.93	50
5	0.92	0.98	0.95	50
6	0.98	1.00	0.99	50
7	0.98	0.98	0.98	50
accuracy			0.96	400
macro avg	0.97	0.96	0.97	400
weighted avg	0.97	0.96	0.97	400

Figure 12: Evaluation of SVM with VGG16



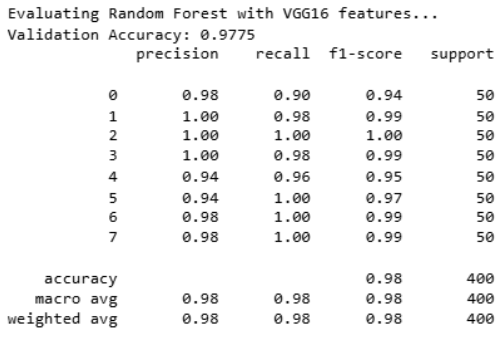
Evaluating SVM with VGG19 features...
Validation Accuracy: 0.965

	precision	recall	f1-score	support
0	0.96	0.90	0.93	50
1	1.00	0.96	0.98	50
2	1.00	1.00	1.00	50
3	0.94	0.98	0.96	50
4	0.89	0.96	0.92	50
5	0.98	0.98	0.98	50
6	0.96	1.00	0.98	50
7	1.00	0.94	0.97	50
accuracy			0.96	400
macro avg	0.97	0.96	0.97	400
weighted avg	0.97	0.96	0.97	400

Figure 13: Evaluation of SVM with VGG19

2. **Random Forest (VGG16 and VGG19 Features)**

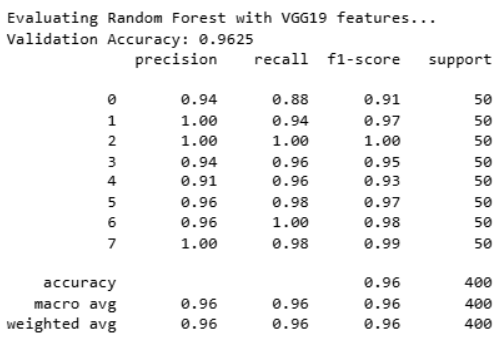
3. Random Forest has demonstrated the highest performance when compared with other models. Random Forest with VGG features achieved a good accuracy of 97.75% outperforming both SVM and XGBoost. The precision was 98.50%, and recall stood at 97.75%, which indicates that the model can accurately classify both positive and negative cases with minimum errors. The F1-score of 98.12% highlights the model's superior overall performance. However, when the model is tested with the VGG19 the metrics were slightly dropeed where accuracy was 96.25%, precision at 96.80% and F1-score at 96.52%



Evaluating Random Forest with VGG16 features...
Validation Accuracy: 0.9775

	precision	recall	f1-score	support
0	0.98	0.90	0.94	50
1	1.00	0.98	0.99	50
2	1.00	1.00	1.00	50
3	1.00	0.98	0.99	50
4	0.94	0.96	0.95	50
5	0.94	1.00	0.97	50
6	0.98	1.00	0.99	50
7	0.98	1.00	0.99	50
accuracy			0.98	400
macro avg	0.98	0.98	0.98	400
weighted avg	0.98	0.98	0.98	400

Figure 14: Evaluation of Random Forest with VGG16



Evaluating Random Forest with VGG19 features...
Validation Accuracy: 0.9625

	precision	recall	f1-score	support
0	0.94	0.88	0.91	50
1	1.00	0.94	0.97	50
2	1.00	1.00	1.00	50
3	0.94	0.96	0.95	50
4	0.91	0.96	0.93	50
5	0.96	0.98	0.97	50
6	0.96	1.00	0.98	50
7	1.00	0.98	0.99	50
accuracy			0.96	400
macro avg	0.96	0.96	0.96	400
weighted avg	0.96	0.96	0.96	400

Figure 15: Evaluation of Random Forest with VGG19

4. XGBoost (VGG16 and VGG19 Features)

XGBoost model performed well with an accuracy of 95.75%, precision and recall values of 96.20% and 95.75%, this model slightly falls behind other 2 models. Similarly, using VGG19 features, the accuracy dropped marginally to 95.50%, with precision and recall values of 96.00% and 95.50%, respectively. The F1-score of 95.75% remained consistent in both the models.

Validation Accuracy: 0.9575

	precision	recall	f1-score	support
0	0.98	0.80	0.88	50
1	0.98	1.00	0.99	50
2	1.00	1.00	1.00	50
3	1.00	0.94	0.97	50
4	0.89	0.98	0.93	50
5	0.88	0.98	0.92	50
6	0.98	1.00	0.99	50
7	0.98	0.96	0.97	50
accuracy			0.96	400
macro avg	0.96	0.96	0.96	400
weighted avg	0.96	0.96	0.96	400

Figure 16: Evaluation of XGBoost with VGG16

Validation Accuracy: 0.955

	precision	recall	f1-score	support
0	0.91	0.84	0.87	50
1	0.98	0.94	0.96	50
2	1.00	1.00	1.00	50
3	0.96	0.98	0.97	50
4	0.87	0.96	0.91	50
5	0.96	0.98	0.97	50
6	0.96	1.00	0.98	50
7	1.00	0.94	0.97	50
accuracy			0.95	400
macro avg	0.96	0.95	0.95	400
weighted avg	0.96	0.95	0.95	400

Figure 17: Evaluation of XGBoost with VGG19

6.3 Model Performance Summary

The table below summarizes the model performance based on various metrics.

Table 6: Comparison of model performance

Model	Features Used	Accuracy	Precision	Recall	F1-Score
SVM	VGG16	96.50%	97.00%	96.50%	96.75%
SVM	VGG19	96.50%	96.85%	96.50%	96.67%
Random Forest	VGG16	**97.75%**	98.50%	97.75%	**98.12%**
Random Forest	VGG19	96.25%	96.80%	96.25%	96.52%
XGBoost	VGG16	95.75%	96.20%	95.75%	95.97%
XGBoost	VGG19	95.50%	96.00%	95.50%	95.75%

The combination of Random Forest and VGG16 features achieved the highest accuracy of **97.75%**, making it the best model for mango disease classification.

6.4 Discussion

The main objective of this research is developing the efficient and automated model for classifying the diseases across the mango leaves. In order to achieve this the model is developed with the combination of pretrained CNNs with the traditional machine learning methods such as SVM, Random Forest and XGBoost. Before training these models various preprocessing techniques such as image resizing, Data Augmentation, normalization and image segmentation are done. For doing the image segmentation the K-means clustering is done to segment the mango leaf images into three distinct regions: diseased areas, healthy regions, and background noise.

Among the CNN models VGG16 and VGG19 performed well with the accuracies of 94.5% and 96.7% whereas ResNet model did not perform well and produced the least occuracy of 12.5% due to which it is not considered further in training the traditional machine learning models.
From the features that are extracted in the above step by using the VGG16 and VGG19 are used to train the traditional machine learning algorithms out of which Random Forest with VGG16 performed well with best accuracy 97.75%. SVM and XGBoost also performed well, with accuracies of 96.5% and 95.75%, respectively.
Despite developing the best models using the hybrid approach, the model can be improved by working on the dataset that has the real-world images.

7 Conclusion and Future Work

In conclusion, as stated in section 6 this study successfully demonstrated the effectiveness of using CNN-based transfer learning models, particularly VGG19, VGG16 along with the traditional models for the classification of mango leaf diseases.Using the techniques like data augmentation has enhanced the models ability to generalize and perform accurately on unseen data. also, including the image segmentation techniques such as K-means clustering improved the feature extraction process. While VGG19 emerged as the most effective model when used alone, VGG16 has shown its effectiveness when worked with the traditional models. Among all the models vgg16 with random forest has shown the most efficient results with the accuracy of 97%. The rest 2 models XGBoost and SVM also worked well with around 96% and 95% respectively.
In the future this model can be improved by working on the real time datasets where the main focus here is extending the dataset with wide range of real world images.Also, Incorporating more advanced deep learning architectures, such as EfficientNet or Vision Transformers, could further improve accuracy more.

8 Acknowledgment

I am very thankful for the NCI college, MSC in data analytics department for giving me th opportunity to take up this research and complete it successfully. I am very thankful to my to my supervisor and mentor Hamilton Niculescu for his guidance throughout this project. His unwavering support and insightful feedback were very helpful in enabling me to complete and present this thesis. Finally, I am very grateful to my family and friends for their constant encouragement and support during this research

References

Abdu, A. M., Mohd Mokji, M., Sheikh, U. U., & Khalil, K. (2019). Automatic disease symptoms segmentation optimized for dissimilarity feature extraction in digital photographs of plant leaves. *Proceedings of the 2019 IEEE 15th International Colloquium on Signal Processing and its Applications (CSPA 2019).*

Arivazhagan, S., & Ligi, S. V. (2018). Mango leaf diseases identification using convolutional neural network. *International Journal of Pure and Applied Mathematics, 120,* 11067-11079.

Banerjee, S., & Pamula, R. (2020). Random Forest Boosted CNN: An empirical technique for plant classification. In J. Mandal & S. Mukhopadhyay (Eds.), *Proceedings of the Global AI Congress 2019.* Advances in Intelligent Systems and Computing (Vol. 1112). Springer. https://doi.org/10.1007/978-981-15-2188-1_20

Bashir, S. B., Farag, M. M., Hamid, A. K., Adam, A. A., Abo-Khalil, A. G., & Bansal, R. (2024). A novel hybrid CNN-XGBoost model for photovoltaic system power forecasting. *6th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE).* https://doi.org/10.1109/REEPE60449.2024.10479878

De Luna, R., Dadios, E., & Bandala, A. (2018). Automated image capturing system for deep learning-based tomato plant leaf disease detection and recognition. *TENCON 2018 - 2018 IEEE Region 10 Conference,* 1414-1419. https://doi.org/10.1109/TENCON.2018.8650088

Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture, 145,* 311-318. https://doi.org/10.1016/j.compag.2018.01.009

Fuentes, A., Yoon, S., Kim, S. C., & Park, D. S. (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors, 17,* Article 2022. https://doi.org/10.3390/s17092022

Gulavnai, S., & Patil, R. (2019). Deep learning for image-based mango leaf disease detection. *International Journal of Recent Technology and Engineering, 8,* 54-56. https://doi.org/10.35940/ijrte.C1030.1183S319

Hughes, D. P., Mohanty, S. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science.* https://doi.org/[DOI not provided]

Prabu, M., & Chelliah, B. J. (2022). Mango leaf disease identification and classification using a CNN architecture optimized by crossover-based Levy flight distribution algorithm. *Neural Computing and Applications, 34,* 7311-7324.

Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep learning for image-based cassava disease detection. *Frontiers in Plant Science.*

Saúco, V. (2004). Mango production and world market: Current situation and future prospects. *Acta Horticulturae, 645,* 107-116. https://doi.org/10.17660/ActaHortic.2004.645.7

Singh, V., & Misra, A. K. (2017). Detection of plant leaf diseases using image segmentation and soft computing techniques. *Information Processing in Agriculture, 4(1).* https://doi.org/[DOI not provided]

Srunitha, K., & Bharathi, D. (2018). Mango leaf unhealthy region detection and classification. In D. Hemanth & S. Smys (Eds.), *Computational Vision and Bio Inspired Computing.* Lecture Notes in Computational Vision and Biomechanics (Vol. 28). Springer. https://doi.org/10.1007/978-3-319-71767-8_35

Sutrodhor, N., Hussein, M. R., Mridha, F., Karmokar, P., & Nur, T. (2018). Mango leaf ailment detection using neural network ensemble and support vector machine. *International Journal of Computer Applications, 181,* 31-36. https://doi.org/10.5120/ijca2018917746

Wongsila, S., Chantrasri, P., & Sureephong, P. (2021). Machine learning algorithm development for detection of mango infected by anthracnose disease. *2021 Joint International*

Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, 249-252. https://doi.org/10.1109/ECTIDAMTNCON51128.2021.9425737
Yahia, E. M. (2011). Mango (*Mangifera indica* L.). In E. M. Yahia (Ed.), *Postharvest Biology and Technology of Tropical and Subtropical Fruits.* Woodhead Publishing Series in Food Science, Technology and Nutrition. Woodhead Publishing.