An Approach to Identify and Classify Banana leaf pests using Machine Learning and Deep Learning Neural Networks

MSc Research Project
Data Analytics

Yogesh Ravindra Rokade
Student ID: x19214057

School of Computing
National College of Ireland

Supervisor: Abubakr Siddig

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Yogesh Ravindra Rokade
Student ID:	x19214057
Programme:	MSc Data Analytics **Year:** 2021-2022
Module:	Research Project
Supervisor:	Abubakr Siddig
Submission Due Date:	19th September 2022
Project Title:	An Approach to Identify and Classify Banana leaf pests using Machine Learning and Deep Learning Neural Networks

Word Count: **12066** **Page Count** **36**

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.

Signature: Yogesh Ravindra Rokade

Date: 19th September 2022

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):	

An Approach to Identify and Classify Banana leaf pests using Machine Learning and Deep Learning Neural Networks

Yogesh Ravindra Rokade
x19214057

Abstract

Leaf pests' infection are the most common phenomena which can be seen across the world. In recent years, agriculture has significantly contributed to Gross Domestic Product (GDP) across several countries. Most Asian nations are agriculture-based, and their economies are heavily reliant on the export of agricultural products and make huge profit from production of crops. After China, India is the world's second largest fruit production, as well as India ranks first in banana production. As most of Asian countries have extreme weather, there has been a huge increase in bacterial and fungal diseases, thus degrading the quality and productivity of plants. Not detecting these diseases at early stages may result in low production of fruit. Many farmers are uncertain about disease type that has infected their crops, hence leading to inappropriate usage of pesticides on plant affecting their growth. Also obtaining expert guidance in such circumstances is time consuming and costly for farmers. Many survey states that due to crop losses huge number of farmer's attempt suicide. To overcome these severe issues, identifying and classifying the banana leaf pests at the earliest is important. The aim of this study is to build a machine learning model that can help farmers to classify the pests, reduce plant loss and eventually help increase GDP. For classification, six machine and deep learning models are evaluated with different performance metrics. The models VGG19 and DenseNet201 outperforms other models with an accuracy of 95%.

Index terms: *Banana leaf, gross domestic product, SVM, Random Forest, VGG19, DenseNet201, machine learning.*

Contents

List of Figures

1 Introduction

1.1 Research Background

Plant disease classification has always been an issue since correct classification of real time images and classification of different type of diseases on plant is difficult. According to report generated by Food and Agricultural Organizations of the United Nations (FAO)[1], more than 65% of world's population carry out agriculture. As different countries have different climatic conditions like global warming, sudden climatic change, it has been recently observed that, there has been decline in the production of food and one of the main reasons of decline in production of food is the attack of pests on plants which degrades the quality of crop. Also, from the study conducted by Simon (2011) in St. Vincent, the main factor leading to the degradation of banana production and quality was the presence of various pests and infections. Plant disease has a severe influence not only on the agriculture production, but also indirectly effects the environment through pollution. Pest attack on plant account for 15 to 20% of total agriculture loss. During the worst-case, Farmers would lose up to 50 to 55% of their harvest, which would be a substantial loss for farmers and the country's economy (Tian et al., 2021). Various preventive measures have been taken by FAO to reduce the loss which include educate the farmers about the pest, include expert advice etc, but the results were not that effective.

India is the world's largest agricultural product marketplace, accounting for around 19% of world trade [2]. After China, India ranks second in the production of fruits and first in the production and exporter of banana around the world. Figure 1 shows the total production of banana in India from 2015 to 2022 (reported in million metric tons). The bar graph clearly shows that the banana production in India has expanded significantly throughout the years. Figure 2 shows the GDP of India through agriculture sector which depicts the effect of GDP because of agriculture. Most of all the underdeveloped and developing countries rely heavily on the agriculture, it is necessary to preserve the quality and quantity of crop in heathy condition. Pests attack destroys the banana crop and reduces the overall quality of the crop, causing losses to around 85 percent of the world's farmers (Nagayets, 2005). Furthermore, Joshi and Jadhav (2016) notified the negative consequences which will occur due to low agricultural quantity. As the GDP significantly depends on the export of bananas, it becomes important to detect the pest at the stage.

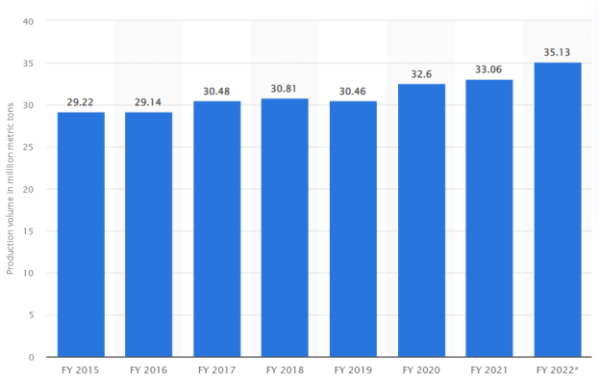

Figure 1: Production of bananas in India from 2015 to 2022 [3]

[1] https://www.fao.org/home/en/

[2]https://economictimes.indiatimes.com/news/economy/agriculture/indias-agricultural-export-grows-economic-survey/articleshow/80585995.cms

[3] https://www.statista.com/statistics/1038905/india-production-of-banana/

Figure 2: GDP of India due to agriculture from 1999 to 2021 [4]

Furthermore, according to report generated by National Crime Records Bureau India, around 17 percent of farmers committed suicide in 2014 because of agricultural damage caused by pests [5]. AT an early stage, the banana leaves are normal, and it is difficult to detect the disease with the visual inspection; expert advice is required by famers to understand the state of the plant, which is costly; additionally, some rural areas are difficult to reach for expertise, making it time consuming and complicated for farmers. This results in improper usage of pesticide by farmers which further damages the crop growth (Pantazi et al., 2019). As a result, in order to make it an effective approach and lower the overall cost of crop production for farmers, crop pest detection and locating the disease severity must be automated (Dhingra et al., 2018). Due to the slow processing speed of system, Bandi et al., (2013) first employed Nave Bayes (NB), Support Vector Machine and Random Forest to classify crop disease.

As the processing speed increases, many study has been done with more sophisticated models. The most prevalent banana leaf diseases are the Panama Wilt, Sigatoka, Mosaic which are types of fungal infection. Various studies have been conducted to determine if the banana leaf is heathy or diseases, as well as to identify prevalent pests. The Convolution Neural Network was used to differentiate the two varieties of banana pests: Sigatoka and Speckled Banana (Amara et al., 2017). Black sigatoka pest is widely found in the lowlands, were the infection causes brown streaks on the banana leaves, which eventually turn the leaf black (Raut and Ranade 2004). There are several organizations that assist farmers in manging crop development, and as a result, various plant disease control technologies have been developed to boost crop yield and successfully control pest infection ((Vipinadas and Thamizharasi 2016). As all these organizations are only operational in certain areas, they are not valuable to all farmers.

1.2 Research Question

"To what level the deep learning models (Efficientnetb1, VGG-19, DenseNet201) can give better sensitivity and specificity than simple machine learning models in the identification and classification of pests on banana leaves?"

[4]https://data.worldbank.org/indicator/NV.AGR.TOTL.ZS?contextual=default&end=2021&locations=IN&start=1999&view=chart

[5] https://ncrb.gov.in/sites/default/files/chapter-2A%20farmer%20suicides.pdf

1.3 Research Objective

The aim of research is to help farmers by making use of machine learning models with low computational power and high-level performance. This study eventually will also help increase GDP of countries which are highly reliable on agriculture by detecting the disease at early stage. Thus, deep learning models (VGG19, Efficientnetb1, DenseNet201) are used which does not require any feature extraction of images is employed (Mohanty et al., 2018). Furthermore, the Grey Level Co-occurrence Matrix (GLCM) approach will be employed to extract texture features for multilabel supervised machine learning models (Indriani et al., 2017). Based on performance metrics, both the techniques will be evaluated and used to detect and classify banana leaf pests. In this study comparison of method using machine learning (Haralick's features) and deep learning will be done and the method with high sensitivity and specificity will be used to detect the disease on the banana leaves.

1.4 Research Contribution

Agriculture is crucial to the economies of underdeveloped and developing countries. The limited export of agricultural goods has a substantial influence on developing countries' GDP. Since India is the largest exporter of banana to the world, maintaining the quality of banana is important which can be done by early detecting the pest on banana. Furthermore, the prior way of identifying the disease was monitoring with naked eyes by expertise to identify the pest and detect how severely the crop is infected with that pest, this method was time consuming, expensive, and occasionally erroneous for farmers. By making use of machine learning, automatic detection and classification of plant disease is possible with high accuracy and minimum labour.

1.5 Format of the paper

The paper's format has been organised accordingly to the needed criteria and described further for better readability. Section 2 – 'Related work', which shows previous related research. Section 3 – 'Methodology', which explains the method used to identify and classify the banana leaf pests. Section 4 – 'Design Specification', which gives the information of all the stages of the research. Section 5 – 'Implementation', which shows the implementation of machine learning models. Section 6 – 'Evaluation' & results', Section 7 – 'Discussion and findings of the research is discussed', and Section 8 – 'Conclusion and Future Work'

2 Related Work

Machine Learning has widely been used to solve various agricultural related problems. It is crucial for farmers to grow and harvest crops on a constant basis in order to feed the ever-growing population of the world. It is necessary to enhance food production while also maintaining the food quality by killing the disease which hampers their grow and degrades the quality, this can be done by making use of effective pesticide at the early stage of disease. Because of this infection, the soil quality degrades which in turn results in economic loss to farmers and thus they sometimes commit suicide. It is feasible to diagnose damaging pests at an early stage using appropriate technologies or methodologies, and respectively good disease control strategy may be implemented and followed.

2.1 Statistical Techniques

Various statistical techniques have been used for classification of various disease. Eye disease classification, kidney disease detection is some of the state-of-art with this methodology has been applied and have shown better results. Kaul, Pandey and Goel (2018), detected the disease on soyabean leaves which is caused due to fungus and changing climate. The images of the soyabean leaves were collected from the PlantVillage dataset, which is utilized for plant research on a variety of concerns. A data collected had 4775 images in which 1079 images were healthy, and the remaining images were diseased images of the soyabean leaves. The research took both colour and textural aspects of image to detect the disease and both these factors were considered to extract the features from the soyabean leaf images. The data was split into 3 parts of training and testing and thus 30%, 40% and 50% was used as testing data. When model was applied on the textural features, the classification accuracy was 69% but when the colour features were considered, the model showed slightly higher classification accuracy of 81%. The author when considered both the features, the model showed higher accuracy when compared with other features individually. Different features may be evaluated to improve the model performance, and images from other dataset or real time field images can be used to train the model.

Similarly, Shrivastava and Hooda (2014) used the soyabean leaves to classify the two types of soyabean pests. The images for study were captured using mobile model and the size of each image was around 125KB. The images of the soyabean were segmented before extracting the features from the image. The author emphasized the importance of this step as this eliminate the unwanted noise in the image. The segmentation is carried out by means of morphological operations. Textural and shape features were used as the feature extraction method. Based on this K-NN classification model was evaluated. The author noted accuracy of around 70%.

Mokhtar et al., (2015) made use of tomato leaves to carry out the research of classification of disease. The author segmented the tomato leaves to get the ROI and afterwards derived Geometric and Histogram features to classify two type of tomato pests. The data was procured from internet and was resized to get the desired images of the tomato leaves for research. For segmentation, K-means clustering was utilized, and the resultant tomato leave was extracted. Area under infected part, total area, image length was some of the geometric features which were utilized. Further classification of the tomato leaves was done using SVM with different kernel parameter to get higher accurate results. The author got accuracy of 91% when considering both geometric and histogram features.

2.2 Image Pre-Processing Technique

Image pre-processing is a wide concept which includes various images processing techniques like cleaning the image, image segmentation, image feature extraction, feature selection. The various image cleaning methods are noise reduction, image resizing, image enhancement etc. For segmentation of image, the most widely used techniques are Edge detection and thresholding. Before applying any machine learning model, the feature extraction step is important to get some information from the image. Feature extraction technique is a method to extract the features from the images, which may be of shape, colour, textural etc.

Iqbal et al., (2018) detected the infection on citrus plant leaves using image cleaning, image segmentation and feature extraction method. The author compared the results of each

technique based on the performance, their advantages, and disadvantages. Various methods were used to clean the image which include image enhancement, image median filtering, image scaling and noise reduction. For image segmentation, the author evaluated and compared the various methods like K-means clustering, thresholding, Edge detection, and region segmentation. After segmentation of the image, various feature extraction technique like texture, colour, SURF, SHIFT and Histogram of Gradient were used to extract the features and to train the machine learning models based on this feature. K-means clustering was shown as the best segmentation approach, while textural feature extraction approach was found be the prominent and best for the classification problem. Grey level cooccurrence matrix is another feature extraction method which is used to extract various features such as energy, entropy, correlation, homogeneity etc. Similarly, Bhimte and Thool (2018) used the K-mean clustering for segmentation and grey level occurrence matrix to extract the textural features from the image to classify cotton leaf pests The results from segmentation and feature extraction method were passed to the SVM classifier for classification. The results obtained by combination of K-mean Clustering and GLCM as feature extraction was higher than other method.

Feature extraction has been the critical part for any classification problem to attain high performance but there are also other feature extraction techniques which increases the computing time, the memory storage, and makes classification ever more difficult to process. Priyanka and Kumar (2020) in their research of classification of ultrasonic kidney images using ANN described this problem and suggested using hybrid method of feature extraction to overcome this limitation. To solve this problem, Feature selection was done on the extracted features so that the features which are more valuable to the model will be considered for evaluation. The dimensions were reduced using the principal component analysis and thus this improves the computational time and memory storage problem.

Further, Vipinadas and Thamiizharasi (2016) in their paper illustrated the usage of colour, textural and image region feature extraction method to classify whether the banana leaf is infected or healthy. In their study, the RGB images was first converted to YCbCr colour format image with the help of image processing. The segmentation of the diseased images of the banana leaves is done with the help of thresholding and Adaptive contrast map which converts the image in the format 0’s and 1’s. Thus, the diseased part of the image is displayed in white and rest unwanted background part is displayed black pixels. With the help of these two-image segmentation technique, the diseased part of the banana leaf is segmented. This approach was useful for author in precisely classifying the banana leaves as healthy or infected. Similar approach has been conducted in this research to make use of Thresholding for image segmentation and later using GLCM as feature extraction technique to classify various pests on the banana leaves.

Sun, Jia, and Geng (2018) developed a plant disease detection model in their research work that employs image pre-processing techniques for classification of the disease. The research was carried out in MATLAB processing tool. The research was varied out in four steps which include image edge detection to get the target leaf removing the unwanted background image. This is also done to remove the noise from the image. Next Image segmentation was carried out with the help of histogram segmentation to get the infected part from the healthy green leaf. Further feature extraction was carried out on the extracted lesion part of the leaves and features such as shape, texture and colour were used to classify the type of the disease. Multiple linear regression model was used as the classification model for the study.

Similarly, Poornam and Francis (2021), used image processing techniques and emphasized the importance of image pre-processing method. The author stated that image processing technique helps the machine learning model to train the images in less time, hence computing resources is reduced due to this. Image resizing was performed on the plant leaves data which was gathered from Image Net dataset. CNN model was used for classification of various plant disease where the author got an accuracy of 91%. Hu et al., (2014) used K-means clustering to segment the image of the banana fruit. The image of the banana fruit was subjected to convert into binary image by colour inversion and later image enhancement was done on the threshold image. The noise from the image was removed by using the median filtering technique. Thus, the segmented image of the banana was obtained from enhanced threshold image for further research. Similarly for this research, to segment the images of banana leaf pest, binary images (thresholding) will be used as the image segmentation method.

2.3 Machine Learning in classification of plant disease

Machine Learning method which was first applied in the field of health and agriculture was in the 1970s, and since then, machine learning has demonstrated its capability in helping, detecting, and evaluating a variety of diseases. With the advent of big data technologies and powerful computers, machine learning has developed to provide new potential for data science in the field of agricultural domain. Jordan, M.I and Mitchell (2015) describes machine learning classifiers as an automated trained system with unique capacity to grasp the relationship between various factors supplied as input and present the accurate diagnostic on some parameters. Agriculture management systems are now turning into intelligence applications that assist farmers in making correct decisions (Liakos et al., 2018). Traditionally, laboratory procedures such as polymerase chain reaction, thermography and hyper spectral techniques and gas chromatography were performed to identify disease on plant, however, these procedures were discovered to be taking more time and was less efficient (Ramesh et al., 2018). They also emphasized the importance of classifying the disease at early stage since the pathogens if not detected early, results in low food production which in turn leads to food insecurity. The author applied Random Forest machine learning algorithm to classify whether the plant is healthy or diseased. For the research, three feature descriptors were used: Hu moments, Haralick's texture feature and Colour Histogram with HoG as feature extraction method. First the images of the plant leaves were converted into Grayscale and then Haralick's features were extracted. For colour histogram, the image was converted into HSV colour space and then model was applied on these features. The Random Forest showed an accuracy of 70.14%.

Similarly, Shruthi, Nagaveni, and Raghvendra (2019) highlighted the need of farmers considering economic factors, analysing soil, and then selecting a suitable crop for farming. Farmers' traditional way of detecting pest growth on plant was by visualizing the plant with eyes, which was time consuming and within this the growth of pest may be multiplied. As a result, the author suggested that machine learning approaches be used to make agriculture sector effective. For the research, the image was acquired from cameras and data annotation was done to label the type of diseases. After than image processing on the plant leaves was performed which include image segmentation to remove the unwanted background followed by feature extraction. Colour features were used to extract the features and then classification models were evaluated on plant leaves. The author also made use of CNN to detect the plant disease type. The CNN model performed well than the other machine learning models. Similarly, for classification of five different banana leaf disease machine learning models will

be performed along with deep leaning models to compare the sensitivity and specificity of both techniques.

Panigrahi et al., (2020), described the importance of agriculture for a countries economy and it is the main source of income for most of the people. Various efforts are being taken to increase the production of food by researchers, analysts, and government. To address this issue, the author used machine learning technique to help reduce the growth of pests on maize plant. In this research, Naïve Bayes, KNN, SVM, Decision tree and Random Forest was used for classification. The image was procured from Plant village website. These images were subjected to grey scale to remove the unwanted noise and image segmentation was performed using the edge detection method. The author achieved an accuracy of around 70% with random forest classifier. Likewise, Hatumal, Shakya and Joshi (2020), presented an idea to overcome agricultural economic loss in developing countries by introducing a classification model which classifies plant diseases. Various Haralick's features like Entropy, Correlation, Contrast, were extracted and they passed these features of plant to various machine learning models like SVM, CNN, KNN and Random Forest. The CNN model showed higher accuracy of 97.89 percent for classification in their research.

Rahamathunnisa (2020) emphasizes the importance of agricultural products in India and describes how its is crucial to detect the vegetable disease in time. In their research, they made use of K-means clustering to segment the images of infected vegetables from the rest part and then extract features from the segmented part of vegetables. They also stated the importance of features extraction as it helps to reduce the processing time by converting the images to some statistical value and applying machine learning models. The results from SVM showed higher performance. Chaudhari and Patil (2020) collected four different diseased images of banana leaves from various farms and classified the disease using Support Vector machine. For this, the images were firstly converted into L*a*b color space to find the dissimilarity in the image and then image enhancement was done using Adaptive Histogram Equalization Contrast to increase the image quality. After this, image segmentation was done using K-mean clustering to remove noise and get the Region of interest (ROI). Features were extracted and SVM model was used to classify various diseases. The accuracy obtained was around 80% but the biggest drawback of research was as the images were collected from farm, correct labelling of the disease must be done or else the model will show improper label for that assigned disease.

Similarly, Agarwal, Sarkar, and Dubey (2019), detected three main disease which hampers the growth of Apple fruit or leaf, namely apple scab, core rot and black rot canker. They began with gathering the required image for research and performed image preprocessing to remove errors and unwanted noise from the image. Later k-means clustering was used to obtain the defective part of the fruit as it was compatible with the classifier. The feature extraction was done using the GLCM method, and finally the apple fruit images were classified using the SVM classifier with an accuracy of 98 percent. Tumang (2019) followed the similar approach to detect the pests and disease on Mango plant. The research was carried out in MATLAB platform. In the research, three main type of mango disease was classified using SVM classifier. Image enhancement was done followed by image segmentation using K-mean clustering and then features were extracted using GLCM with constract, variance, smoothness, kurtosis and skewness as input to the classifier. The SVM showed an accuracy of around 45% using SVM classifier. Similarly, image processing with threshold image segmentation and GLCM feature extraction is used for classification and identification of banana leaves.

2.4 Deep Learning

Nowadays, technological advancement in the field of agriculture have successfully met the need to produce enough food to feed the entire world's population. Even with the developments, there are various other factors influencing the growth of crops like, climatic changes, infection of pesticides on crops, and also decrease in count of pollinators etc. Most of the farmers in the developing countries solely rely on the production of agricultural products for their livelihoods, thus making it important to prevent crop loss. Several subject experts by contributing their knowledge made tremendous progress in preventing loss of crops with disease infection. Machine learning and deep learning approach plays a significant role in several domain areas like banking and finance, entertainment, agricultural, health and medical, and so on which solves the complex problems at the initial stages helping humans. Rigorous implementation of deep learning techniques on real world problems have been made. For e.g., identification of diseases, it can be detection of diseases in humans, or it can be diseases detection in crops, etc. To successfully tackle challenges and to solve the problem of diagnosing the plant leaf disease it is important to train the deep learning architectures with both unhealthy (diseased leaf) and healthy leaves. State of art architectures implemented in this research project are VGG19, DenseNet201, and EfficientNet.

Many research in healthcare domain have implemented transfer learning for early detection and analysis of the illness. Researchers in their article developed a deep learning model which can successfully be able to detect the abnormalities of the lung with the help of radiograph images. Several other deep learning architectures such as VGG19, AlexNet, VGG16, ResNet50 with SoftMax classifier were implemented at the begin and later based on the evaluation of all models it was concluded that VGG19 had the better accuracy. The research further developed a customised VGG19 architecture which had distinct classifiers such as random forest, svm, knn, and decision tree. VGG19 combined with random forest classifier gave outstanding accuracy making the model useful to diagnosis clinical radiographs images (Dey et al., 2021). Deep learning has proven its ability of efficiently training on large number of datasets. Not only deep learning models can effectively diagnosis pneumonia from radiograph images, in recent pandemic many deep learning architectures have been used to detect face mask on individual, or diagnosis of covid19, for also detection of heart disease and so on. Jaiswal et al. (2021) in the research article planned on developing a system which when implemented can detect covid19 in chest area. DenseNet201, a pre-trained deep transfer learning model is used to perform classification between infected covid19 patient and health patient. Further a comparison on DenseNet201 and other pre trained architectures like VGG16, ResNet was carried out. Similarly, Yu et al. (2019) proposed DenseNet201 to detect the breast cancer illness at the initial stages and help the radiologists to improve work efficiency. The images utilized to develop semi-automatic system are the mammogram images, this architecture have higher specificity and sensitivity as well.

Koonce (2021) states EfficientNet neural network to be state of art technique for image recognition which was developed surprising for mobile devices by combining several different techniques and once the base model is generated it is very convenient to combine the base model with other techniques like data augmentation in order to acquire better results. Researchers in the article proposed a system of classifying different kinds of fruits using EfficientNet and MixNet, further validating the results of the architecture on dataset named real fruit. Also, they put forth interesting insights on how important the input dataset is to obtain better prediction results, regardless of the architecture being more complex. Thus,

suggesting on performing quantitative and qualitative analysis on the different kinds of input data(images), architecture of the model and so on (Duong et al., 2020).

Table 1: Literature Review Summary

Author	Methods	Objective	Advantage	Limitations/Gaps
Shrivastava and Hooda (2014)	Machine Learning-KNN Classifier	Detection of disease on Soyabean plant	KNN classifier showed good accuracy	Only two diseases were classified.
Mokhtar et al (2016)	Support Vector Machine	Tomato leaves disease detection	Achieved study objective	Missing dataset description and feature extraction method is missing.
Vipinadas and Thamiizharasi (2016)	Machine Learning-SVM	Detection of disease on Banana leaves	Segmentation using Thresholding was done.	Classified only Banana leaves as healthy or diseased. No accuracy mentioned.
Ramesh et al. (2018)	HOF Feature Extraction Random Forest Naïve Bayes Support Vector Machine	Random Plant leaves	Haralicks features were used to classify leaf as healthy or diseased	Segmentation was not performed. Low Classification Accuracy
Poornam and Francis (2021)	CNN	All Plant leaves	Included all plant for study.	Few diseases were not classified properly.
Jaiswal et al (2021)	Deep Lesrning	Detecting Covid in patients	Accuracy is good.	Complex structure, High computational time. Performance was not compared.

3 Research Methodology

When it comes to data mining projects, the two most often used research approaches are (KDD) Knowledge Discovery in Database and (CRISP-DM) Cross Industry Standard Process for Data Mining. These approaches are effective for illustrating the life cycle of a certain project. Both the data mining techniques require many phases to be completed when completing a project. The research is carried out with CRISP-DM methodology to complete the entire project. CRISP-DM is simple descriptive phase methodology, where it is impossible to identify the relationship within the data at the start of the process. It is critical to recognize that the phases of the cycle may be linked, and which should be entirely based on goals or literature review (Chapman et al., 2000). Figure 3 shows the phases of this study which is similar to CRISP-DM methodology. The CRISP-DM like methodology is followed as it includes a business component that helps in complete understanding of the research

project. The banana leaf pest classification methodology has six main phases: Understanding the agricultural business, field image data collection, pre-processing the data, data model creation, evaluation of the model and finally deployment of the model.

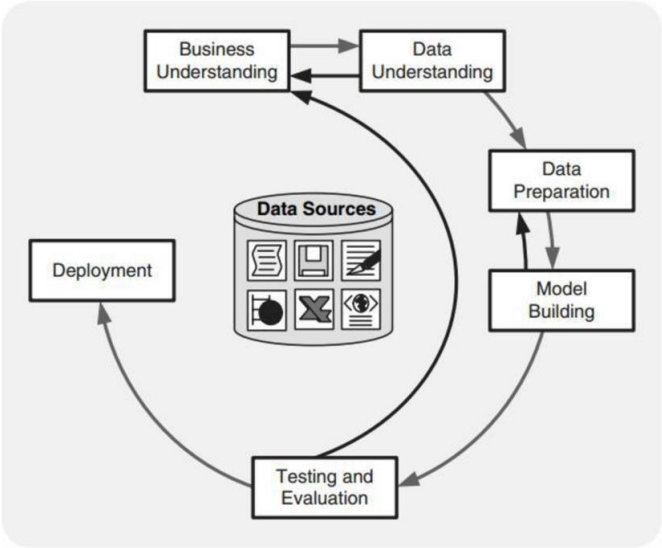

Figure 3: Banana Leaf Pest Identification Methodology Phases

3.1 Business Understanding

The first phase in banana leaf pest identification methodology is understanding the overall agricultural business where the major goal is to grasp the general needs of the issues in the study. During the business understanding phase, farmers must have the knowledge of pesticide and must know when to apply the pesticide on banana crop. This will surely assist in meeting our research goal of early identification of pest and usage of the proper pesticide to kill the disease.

3.2 Data Understanding

The next phase in pest identification methodology is the data understanding/data gathering where the image for the research is gathered and further processed to obtain significant insights from the data. The data was procured from Kaggle and Mendeley repository which is open-source repository for research purpose. The Kaggle dataset consist of images of various pests on banana leaves which include healthy as well as diseased banana images. The Diseased images of the banana leaf pests has three major banana pests' images which are cordana, pestalotiopsis and sigatoka. Real field images were taken using mobile camera. All the pests' images are stored in separate folders with respective disease label on it. The Mendeley dataset consist of field images of banana leaf pest which contains healthy, sigatoka and xamthomonas. The healthy and sigatoka images of Kaggle and Mendeley are merged into a single folder to increase the size and variation of images. The images are in .jpeg and jpg format. The total size of the dataset is 1174 images, and the images have different resolution as it is real time field images. Dataset links:

https://www.kaggle.com/datasets/kaiesalmahmud/banana-leaf-dataset
https://data.mendeley.com/datasets/rjykr62kdh.

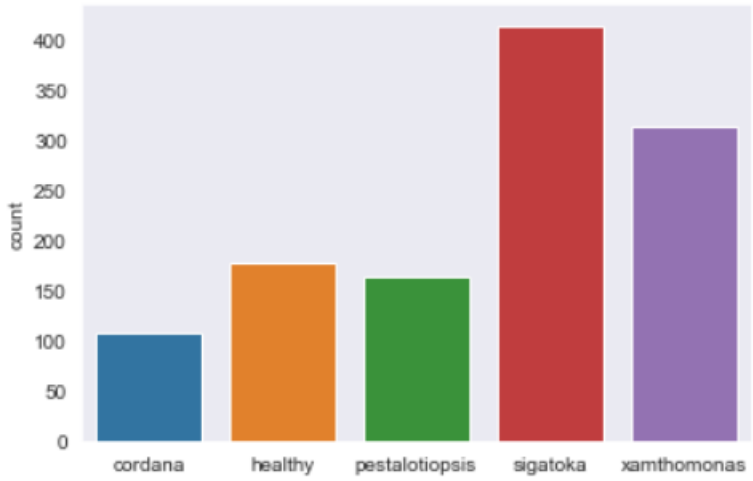

Figure 4: Total count of Banana pest images

3.3 Data Preparation

The next phase after data gathering is data processing. The image processing is crucial as all selection and transformation of images will be passed onto machine learning models; all the activities performed in this phase will have an impact on the output of the modelling approaches. As the images were taken from mobile camera, it had watermark of the cell phone, so to avoid machine learning models learn unwanted patterns, the watermarks were cropped out to focus on the banana leaf image. Also, some images were blurred which will make machine learning models to learn unwanted noise, all those images were removed from the dataset. As the dataset was taken from two different repositories, the name on the images were different, so to read the images and make further research all the images were renamed to their respective folders.

Processing an image which has large dimensions takes time which results in loss of computational power. To overcome this computational cost issue, we have removed the undesired Region of Interest (ROI) from our target image. To obtain the requisite ROI, we first convert all the images to Gray-level scale. Once all the images are converted to Grayscale, the images are converted to Binary images (threshold images). This modification of image is performed to make the image machine-readable for further processing. To get the desired ROI, the unwanted background is removed by means of thresholding and resizing the image. Hence, the ROI was extracted for further research.

3.3.1 Feature Extraction using GLCM

The feature extraction of the banana images is done with the help of Grey Level Cooccurence Matrix (GLCM). GLCM is a statistical method in which the spatial relationship of pixels is considered. The GLCM is a matrix that displays how often a number of distinct sets of pixels present in a grey-level image. The GLCM function evaluate the textural pattern of image by determining how frequently pairs of pixels with defined value and defined spatial relationship occur in the given image, thus forming a GLCM and then calculating the statistical measurements from the matrix. The GLCM matrix is calculated only for a limited set of distances and angles (Alazawi et al., 2019). Similarly, Hossain et al., (2019) made use of the GLCM method to get the textural features on the plant images and detect the disease on them.

Out of 14 features, Four of Haralick's feature: Contrast, Energy, Homogenity and correlation, are utilised in this research (Haralick et al., 1973).

- Contrast: The contrast measures the local differences in the gray-level co-occurrence matrix.

$$Contrast\ (d,\theta) = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} (i-j)^2\ P_d^{\theta}(i,j)$$

In the contrast equation, (i,j) represents the elements acquired in GLCM (unnormalized), N is the number of grey-levels and P_d^{Θ} is the pixel probability, d is the interval, Θ is the direction.

- Correlation: The correlation measures the combined probability relationship occurrence of the provided pixel pair.

$$Correlation = \sum_i^M \sum_j^N \frac{(i-\mu)(j-\mu)P[i,j]}{\sigma^2}$$

In the correlation equation, μ represents the mean and σ represents the standard deviation of the banana leaf image(i,j).

- Homogeneity: The Homogeneity represents the level to which the allocation of elements in the grey level matrix is near to the GLCM diagonal. The homogeneity will be larger when the grey-level pixel pair are determined to be identical.

$$Homogeneity = \sum_i^M \sum_j^N \frac{P[i,j]}{1+|i-j|}$$

In Homogeneity equation, M and N are the dimensions of the (i,j) image.

- Energy: The Energy is also known as the angular second moment. The energy returns the sum of squared items in GLCM. It counts the overall repeating pixel pairs, if the value of energy is high, it means that one pixel pair may recur several times.

$$Energy(Angular\ Second\ Moment) = \sum_i^M \sum_j^N P^2\ [i,j]$$

In Energy equation, M & N are the dimensions of the (i,j) image.

3.3.2 Feature Scaling

Feature Scaling is one of the important data preprocessing steps which is done on the independent features or variables. As the real time data may have high variation in magnitude, so it will output wrong prediction. Feature scaling if not performed, result in a significant imbalance as the values with the high range will be considered as greater significance in prediction and the value with lower range will be treated as lower significance for prediction. The two main techniques which are usually performed for feature scaling are Normalization and Standardization. Normalization is a technique when the data points are of higher value and we want the value in interval [0,1] or [-1,1]. Basically, higher values are

brought to same scale as of all the other datapoints between [0,1]. Standardization technique is used when we require mean =0 and variance of data =1. For the classification of pest on banana leaves, min-max scaler is used to scale the features and normalize the data in the range 0 to 1.

The equation of Min-Max scaler is as follows:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3.4 Modelling

3.4.1 Random Forest Classifier:

The Random Forest is a machine learning method based on the idea of numerous decision trees. It is a supervised learning technique that is based on many decision trees. The main reason of using many such decision trees is to create a robust and powerful decision which is not possible with a one decision tree. The classifier's decision tree is completely randomized; hence the tree is termed as 'Random'. The rules are automatically set out at each node of the tree making the tree to make more correct decisions. The random forest classifier is widely used to solve classification problems. Random forest performs effectively especially when there are large number of data and few number of features (Iqbal and Talukder 2020). For the study, the author used 450 infected and healthy images of potato and thresholding was performed to extract the desired Region of Interest, the random forest showed higher accuracy of 97 percent. The author stated that the major reason for achieving higher prediction rate was the usage of higher number of decision tree and combining a greater number of decision trees. All the trees in random forest operates in parallel, and no interaction occurs between them while they are being constructed. The Random forest's general structure is seen below:

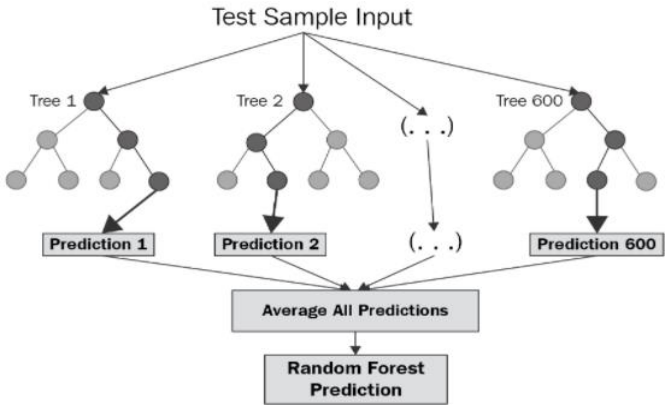

Figure 5: Random Forest Classifier Architecture

The Random Forest ensures that the output of a particular tree is uncorrelated to the output of remaining tree. This is accomplished by bootstrapping the input data. Various estimators are evaluated in the model, and optimization can be done to find the best set of parameters for the random tree (ur Rahman et al., 2017).

3.4.2 K-Nearest Neighbors Classifier (K-NN)

K-Nearest Neighbors is a supervised learning technique which is commonly used to solve carious classification problems. KNN gets all the sample data, and the next data point will be allocated based on the nearest points in the training data. The value for K is determined by calculating the Euclidean distance. Depending on the value of the Euclidean distance, the data point will be contained to a particular group if that point is closest to that group, thus it effectively creates the boundary to categorize the data. As a result, the value of K determines how KNN is executed, and thus high value of k will provide more accurate predictions as the variance in the data will be low. Hossain et al., (2019) applied a KNN classifier to identify and categorize plant leaf diseases since the system requires no learning phase and trains the data simultaneously along with making predictions. Because of this, KNN is quicker than most of the machine learning models.

3.4.3 Support Vector Machine (SVM)

Support Vector Machine is a type of supervised machine learning method that is used for classification and regression tasks. SVM is also used to find the anomalies (outlier) from the data . The SVM technique produces a decision border that divides the dimensional space into a class set, allowing us to input new data value for further reference. The Hyperplane is the line or decision boundary that splits the data points (also known as vector) and by making use of the data point points, it becomes easier to generate a hyperplane. All these extreme vectors which generate the optimal line or hyperplane are called as Support vectors, and the method together is known as support vector machine. If a straight line or a linear function cannot separate the data, a function can be used instead to move the data to high dimensionality space. After this, a kernel function is applied to the data to generate a hyperplane. A kernel function accepts the data points as input and returns the dot product of data points in the feature space as output. From the literature review, support vector machine has very well performed on various classification problems and so SVM will be evaluated to classify and identify the pest on banana leaf.

3.4.4 EfficientNet-B1

EfficientNet-B1 is a type of convolution neural network architecture and scaling approach that uses a compound parameter to scale all the width/depth/resolution dimensions evenly. Unlike traditional methodology, while arbitrarily scales these elements, the EfficientNet scaling approach evenly adjusts network depth and resolution using a predetermined scale coefficient. EfficientNet are basically a group a CNN models which is efficient computationally than other models. The EfficientNet is a family of 8 groups which is B0 to B7, as we move from B0 to B7, the number of parameters remains unchanged, but the accuracy of the models increases (Tan and Le 2019). As efficientnet-b1 is computationally efficient, it will be compared with other deep learning models to classify the pests. The architecture of EfficientNet-B1 is shown below in Figure 6.

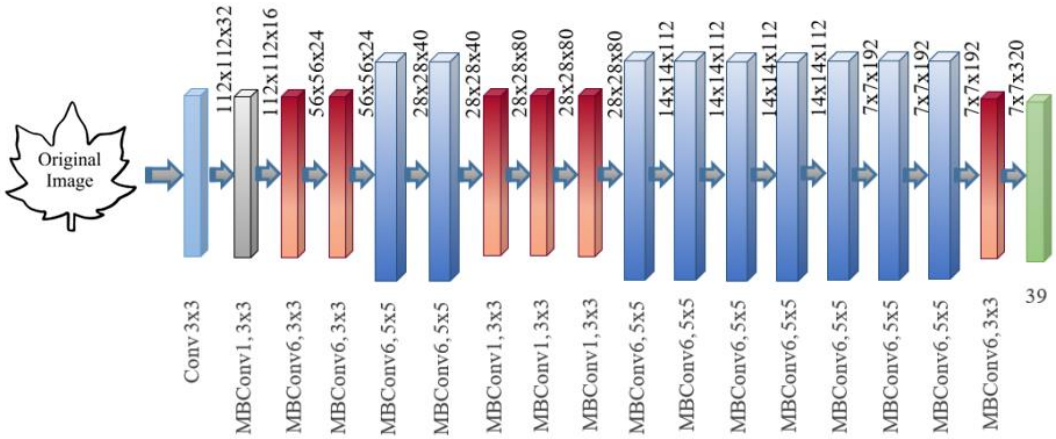

Figure 6: EfficientNet-B1 Architecture

3.4.5 DenseNet201

Recent research has shown that convolutional layers with shorter connections between layers adjacent to the source input image and layers close to output image can be significantly deeper, highly accurate and are more efficient to train the input images. Apart from the convolution and pooling layers, the DenseNet contains two basic layers: the Dense layer and the transition layer. In this research, we adopt this insight and make use of the Dense Convolutional Network (DenseNet), a feed-forward network that joins a layer to every other layer in the model. Unlike standard convolution network, which has a single connection between other layer, DenseNet will have K(K+1)/2 connections. The output of previous features is served as input to the current layer and the feature output is served as input to the next layer in the network. DenseNet addresses the vanishing gradient problem and significantly reduces the overall number of inputs. The DenseNet shows high accuracy when compared with other model, however it takes more computational time than other deep learning models (Huang et al., 2017). Too et al., (2019) compared the efficacy of several models in diagnosing plant disease using transfer learning models. The author made use of Batch Normalization as input to the next layer which solved the problem of overfitting and achieves greater accuracy and lowering train time.

3.4.6 VGG19

VGG19 is a type of transfer learning model of VGG family which has 19 layers. Among the 19 layers, 16 are convolution layers, 3 are fully connected, 5 are maxpool and a softmax layer. The VGG19 makes use of densely connected convolution layers to get higher accuracy. The input to the VGG19 is fixed to 224x224 pixels and it is trained on the ImageNet dataset. VGG19 has widely used for various classification problems and had provided accurate results than other models. As stated in the literature review, (Islam Hoq and Rahman 2019), potato dataset with limited number of input images were used to assess the performance of VGG with limited number of data images. For plant disease classification, VGG19 outperforms other models with higher accuracy and lower execution time per epoch. As a result, VGG19 essentially helps in the reduction of computing time and resources by making use of the pre-trained models for training the new data. Also, as the number of layers rises, the input becomes distorted as it reached the last layer of CNN, hence transfer learning model (VGG19) are used to overcome this limitation.

4 Design Specification

The research flow is shown below in Figure 7:

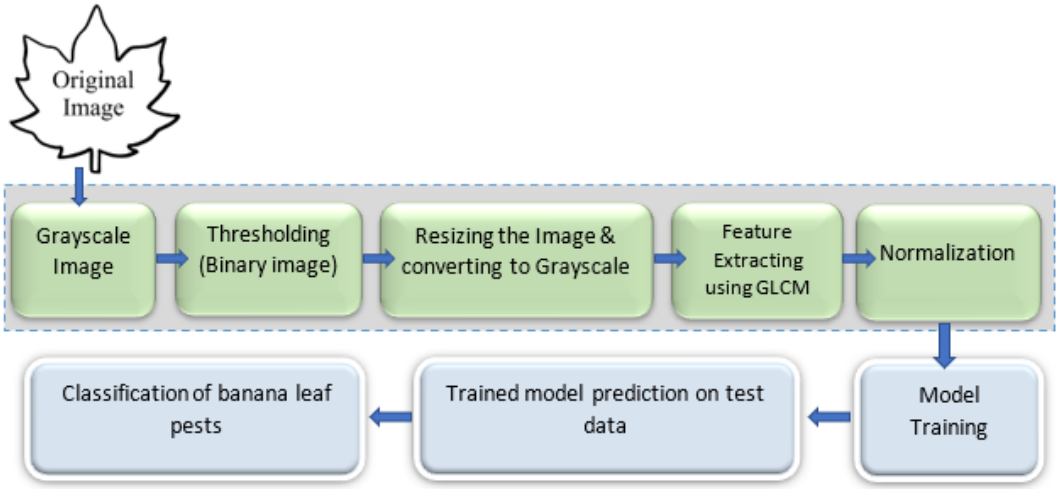

Figure 7: Project Design Flow

- The images of the banana leaf pests are initially acquired from open-source data source (Mendeley repository and Kaggle dataset) that contains images of various pests on banana leaves along with healthy images of banana leaf. All these images are collected in local system and are processed for further evaluation. As the images are acquired from two different data source, renaming of the images is done so that it becomes easier to read the images. The images which are blurred are removed from the dataset as by the model will run unwanted noise by training such images. The images are both in jpeg and jpg format.
- The images are first converted to grey scale and then to binary image to remove the unwanted background from the image. As all the images are of different dimension, the images are resized to a set dimension so that all the images will be in same dimension. After resizing, the images are subjected to grey scale to run the GLCM.
- Grey-level cooccurrence matrix is used as the feature extraction matrix to get all the textural features of the image.
- Feature scaling is done to normalize the data with the help of minmax scaler to get all the features in the same range.
- After normalization of data, simple machine learning models along with deep learning models are evaluated based on precision, recall. F1 score and accuracy.

5 Implementation

For the classification and identification of banana leaf disease, many machines learning models have been developed on different datasets. The main limitation observed in most of the research was the datasets used, the images were restricted and had only few diseases to classify. Furthermore, from the literature review it was found that classification using SVM, KNN, Random Forest, VGG19, DenseNet201 and EfficientNet-b1 showed good results for other classification problems. So, these models are tested on classification of disease on

banana leaf with five different disease type. The dataset used for classification is procured from two open data source which contains images of banana leaf disease.
This section explains the end-to end steps involved for image processing, image segmentation and feature extraction.

5.1 Implementation of Support Vector Machine, Random Forest and KNN

- To start with, the images of the Banana leaves were procured from Kaggle data repository and Mendeley repository which is open data source. The images were merged into a folder and noise, or unwanted images were removed from the dataset. Some images from the Kaggle dataset had camera watermark on them they were cropped manually to avoid unwanted training to machine learning models. As the images were merged into a folder, renaming of the images were done to make it easier to read and further analysis can be undertaken.
- Initially, to read the images from the dataset, OpenCV method is utilized. OpenCV method is open-source library for machine learning to process images. When an image is loaded using OpenCV method, the image by default is loaded in blue, green, red format but for the image to further evaluate we require the image in Red, Green, Blue format, for this we use the cvtColor() method.
- The images obtained for this study varied in size, which was uneven and so as a result resizing of the images is performed. Image resizing is performed using resize() method to make all the images in same dimension.
- Following colour conversion and resizing, the images is prepared for feature extraction. For obtaining the textural features, we require all our images in grey scale as it has only single dimension. The grey scale has range in between (0 to 255).
- Next to obtain the desired region of interest, which is the infected region of banana leaf, image segmentation is done. Image segmentation is done using Adaptive Threshold (Thresholding) which converts the image in the range 0 and 1. If a pixel in an image has a value lesser than threshold i.e 0 is considered as background image otherwise the pixel is considered as the foreground pixel i.e 1 which is the infected part of the banana leaf. The image obtained after greyscale is also called Binary image.
- After thresholding, feature extraction of the images is done using GLCM and four textural statistical features were extracted which were further used for classification. The features obtained were subjected to scaling as to scale all the images equally, thus no feature will have higher effect over other in the dataset. Feature scaling was performed using the MinMax scaler and thus we get normalized images.
- The normalized data from minmax scaler is then separated into training and testing sets and then it is ready for classification using machine learning models. The Figure 8 shows the overall image processing process.

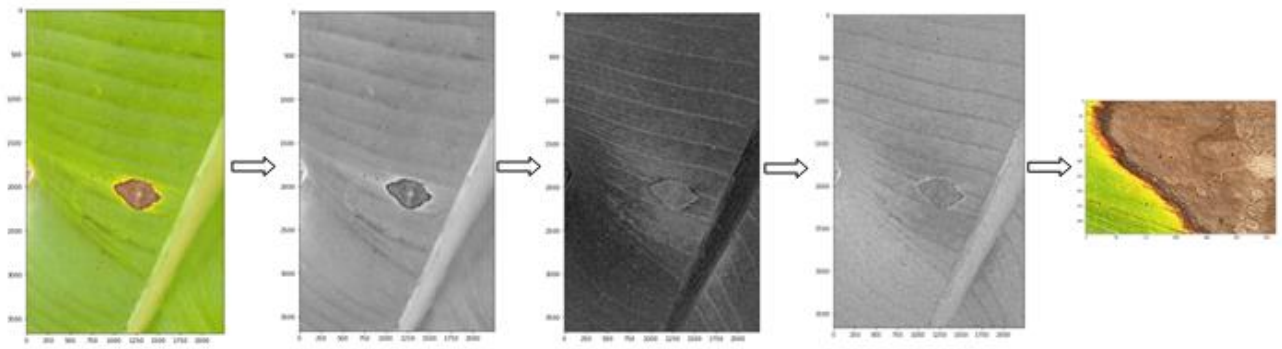

Figure 8: Banana Leaf Disease Image Processing

5.1.1 Random Forest Classifier

After image segmentation and feature extraction techniques is completed, the dataset is ready to perform classification.
Hyperparameter tuning is done before implementing the model to find the best values for each model. This is done using the GridSearchCV method. The parameters are passed as list to the model and the best performed parameter is given by the method. Figure 9 shows the GridSearchCV for random forest model. Thus, the parameter which gives highest performance for random forest is selected as best parameter.

```
from sklearn.model_selection import GridSearchCV

param = {'n_estimators': [1,5,10,20,50,100]}
random = RandomForestClassifier()
clf = GridSearchCV(random, param)
clf.fit(x, y)
sorted(clf.cv_results_.keys())
print(f"Random Forest Classifier Best paramaters {clf.best_params_}")
print(f"Random Forest Classifier Model score {clf.best_score_}")
```

```
Random Forest Classifier Best paramaters {'n_estimators': 100}
```

Figure 9: Random Forest Parameter Evaluation

After completing the hyperparameter tuning, the model is evaluated and the best score for the model is found. The Random Forest model is then tested on the testing dataset and confusion matrix along with classification report is displayed.

5.1.2 K-Nearest Neighbor

KNN is in the sklearn package which is used to load various machine learning models. Similar to Random Forest, Hyperparameter tuning is done for KNN to find the best possible parameter. Again, GridSearchCV is in which user specified input are given and among them best possible parameters are shown. Figure 10 shows the GridSearchCV method for KNN. The best possible value for k is decided and used to find the best score for the model. The k value is the segregation medium which decides the class of the input image.

```
from sklearn.model_selection import GridSearchCV

param = {'n_neighbors': [7,9,13,17,19]}
KNN = KNeighborsClassifier()
clf = GridSearchCV(KNN, param)
clf.fit(x, y)
sorted(clf.cv_results_.keys())
print(f"KNN Best paramaters {clf.best_params_}")
print(f"K-Nearest Neighbors Classifier Model score {clf.best_score_}")
```

```
KNN Best paramaters {'n_neighbors': 19}
```

Figure 10: Best Parameter Search

Model Evaluation is done after running the model using confusion matrix and classification report.

5.1.3 SVM

SVM has the kernel function which are the mathematical functions to perform operation on the data. The various types of kernels are Gaussian, Sigmoid, Radial Basis Function, linear, non-linear kernels. GridSearchCV method is used to find the optimal kernel for the SVM classifier. It is also used to find the best parameter which is required for carrying out the classification. Thus, the kernel takes the data and converts the data into the required format. Figure 11 shows the GridSearchCV method for Support vector machine classifier.

```
from sklearn.model_selection import GridSearchCV

param = {'kernel':['rbf','linear','poly','sigmoid'], 'C':[5,100,500,1000]}
svc = svm.SVC()
clf = GridSearchCV(svc, param)
clf.fit(x, y)
sorted(clf.cv_results_.keys())
print(f"SVM Best paramaters {clf.best_params_}")
print(f"SVM Model score {clf.best_score_}")
```

```
SVM Best paramaters {'C': 1000, 'kernel': 'rbf'}
```

Figure 11: Kernel and Parameter selection

5.2 Implementation of EfficientNet-B1, VGG19, DenseNet201

5.2.1 EfficientNet-B1

Initially a function is created to resize the images and to convert the images into array. The dataset is then split into Train and Test dataset. Data Augmentation is performed to increase the dataset size by Flipping the image Horizontally, Rotating the image. There are various other ways like Zooming, Adjusting Distortion, Vertical shift etc. which can be performed within Data Augmentation. The EfficientNet-B1 model is trained on Imagenet with pooling

as ‘max’. The Dropout layer is added to avoid overfitting by the model. After Dropout layer, two Dense Layers are added which is used to for classification of the images. Fine-tuning is done by setting the model.trainable to TRUE. The activation function used are ‘Relu’ and ‘softmax’ and learning rate is set to 0.0005. Adam optimizer is used in this model to lower the loss function. The ‘relu’ or Rectified Linear Unit is used to increase the computational speed. Figure 12 shows the summary of EfficientNet-B1 model.

```
model.summary()
```

```
Model: "model_26"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_85 (InputLayer)       [(None, 360, 360, 3)]     0

 sequential_64 (Sequential)  (None, 360, 360, 3)       0

 efficientnetb1 (Functional)  (None, 1280)             6575239

 dropout_39 (Dropout)        (None, 1280)              0

 dense_112 (Dense)           (None, 16)                20496

 dense_113 (Dense)           (None, 5)                 85

=================================================================
Total params: 6,595,820
Trainable params: 6,533,765
Non-trainable params: 62,055
```

Figure 12: EfficientNet-B1 model summary

5.2.2 VGG19

Similarly, for implementing VGG19, the images are appended into a single list and the labels are appended to another list. The images are reshaped into 3-D space as required for input to convolution neural network. The image size is set to 224x224 which is required for VGG19 model. The images and the label are converted into array. Label Binarizer is used which is included in the sklearn package to accept the categorical variables and encode them into the dummy variables. After encoding the labels, the data is split into training and testing set. Strategy.scope() method is used to distribute the training load on different GPUs. The data is trained on ‘imagenet’. MaxPool2D layer is used with strides set to 2. Flatten layer is used to convert the 2-dimensional result from MaxPool2D layer into a single dimension. Dense layer is used with activation function ‘softmax’ and the model is compiled using ‘Adam’ optimizer. The batch size is set to 64 and 12 epochs are used to train the model. The learning rate is set to 0.000001. The Figure 11 gives the summary of VGG19 model

```
Model: "sequential_63"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 vgg19 (Functional)          (None, 7, 7, 512)         20024384

 max_pooling2d_73 (MaxPoolin  (None, 3, 3, 512)        0
 g2D)

 flatten_48 (Flatten)        (None, 4608)              0

 dense_103 (Dense)           (None, 5)                 23045

=================================================================
Total params: 20,047,429
Trainable params: 4,742,661
Non-trainable params: 15,304,768
```

Figure 13: VGG19 model summary

5.2.3 DenseNet201

For implementation of DenseNet201 model, the input shape of the image is set to (224,224,3). The model is trained using the ImageNet database. For DenseNet201, 2 dense layers are used. First Dense layer is set to 128 with activation function as 'relu' while the other Dense layer activation function is set to 'softmax'. The model is compiled using the 'adam' optimizer with metrics to evaluate as accuracy. The learning rate is set to 0.000001 like VGG19 model. For training the data, 12 epochs are used, and batch size is set to 64.

6 Evaluation Results

The evaluation of the machine learning models is done using f1-score, precision, recall and accuracy. The testing set will be used to compare the prediction of the model with the trained dataset. As the dataset is imbalanced, so f1-score is the best performance evaluation metrics as it takes into account both precision and recall. As precision considers True positive and recall considers False positive. Both are equally important when evaluating a imbalanced dataset. The evaluation report of all the model is as follows:

6.1 K-NN, Support Vector Machine and Random Forest classifier

6.1.1 K-NN

This section evaluates the machine learning models which are implemented and analysis of the models are done to classify and identify the banana leaf disease. In this study, for evaluating the performance of the model's various metrics like the f-1 score, precision, recall, confusion matrix, Validation Accuracy, Validation Loss are used. The computational time of the deep learning models are also discussed in this section to evaluate the performance. The confusion matrix and classification report of K-NN is shown in Figure 14.

Performance Evaluation

	precision	recall	f1-score	support
cordana	0.59	0.50	0.54	34
healthy	0.33	0.07	0.12	42
pestalotiopsis	0.32	0.22	0.26	32
sigatok	0.71	0.83	0.76	116
xamthomonas	0.65	0.91	0.76	69
accuracy			0.63	293
macro avg	0.52	0.51	0.49	293
weighted avg	0.58	0.63	0.59	293

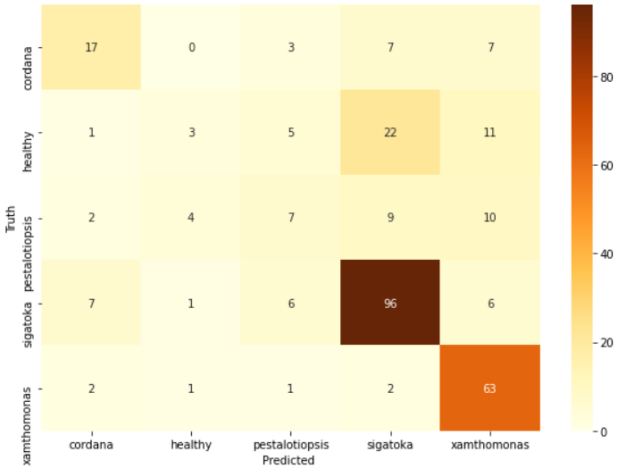

Figure 14: K-NN Classification report and Confusion matrix

From the classification report we can see that the model can classify well for sigatoka and xamthomonas. The precision and recall for healthy leaves are lower than the rest of the diseases. The precision for sigatoka is 71% which means it 71% positively predicted and recall is 83% which means 83% the class is precited positive. The model shows an overall accuracy of 63%.

6.1.2 Support Vector Machine

	precision	recall	f1-score	support
cordana	0.59	0.29	0.39	34
healthy	0.32	0.14	0.20	42
pestalotiopsis	0.33	0.06	0.11	32
sigatok	0.64	0.88	0.74	116
xamthomonas	0.68	0.90	0.78	69
accuracy			0.62	293
macro avg	0.51	0.46	0.44	293
weighted avg	0.56	0.62	0.56	293

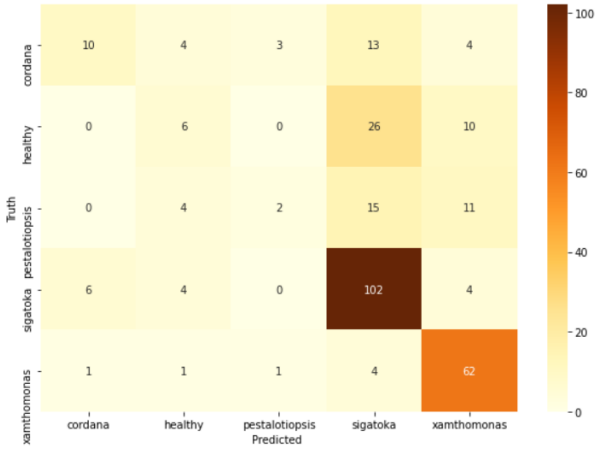

Figure 15: SVM Classification report and Confusion matrix

Similarly, from the classification report we can see that the model shows higher precision, recall and f-1 score for Sigatoka and Xamthomonas and it shows an overall accuracy of 62%. The precision is around similar to K-NN while the recall value is increased for all the class.

6.1.3 Random Forest Classifier

	precision	recall	f1-score	support
cordana	0.69	0.65	0.67	34
healthy	0.54	0.45	0.49	42
pestalotiopsis	0.41	0.28	0.33	32
sigatok	0.80	0.79	0.80	116
xamthomonas	0.69	0.88	0.77	69
accuracy			0.69	293
macro avg	0.62	0.61	0.61	293
weighted avg	0.68	0.69	0.68	293

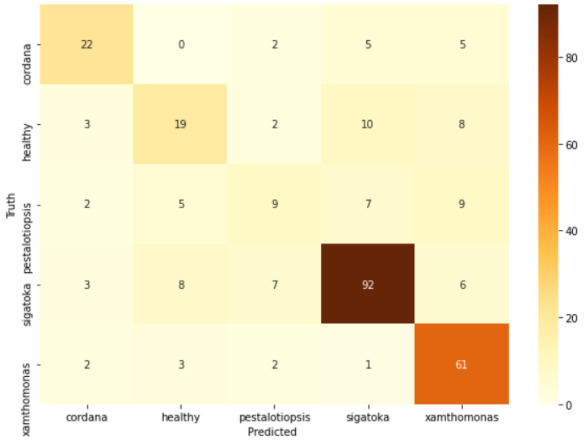

Figure 16: KNN Confusion Matrix and Classification Report

The Random Forest shows higher performance than K-NN and SVM. From the classification report we can see improved values for precision when compared to the other two models. The f1-score for Random Forest is around 63% which is improved. The Confusion matrix shows diagonal like shape which means the images are well classified.

6.2 EfficientNet-B1, DenseNet201, VGG19

6.2.1 EfficientNet-B1

The Validation sparse categorical accuracy obtained while running 1 Epoch of the EfficientNet-B1 model is 91.12%. The loss of the efficientnetb1 model is 0.319. As the accuracy was lower when compared to previous literature review, DenseNet201 model is applied. The validation accuracy and validation loss is shown in Figure 17.

```
50/50 [==============================] - ETA: 0s - loss: 0.6626 - sparse_catego
rical_accuracy: 0.8006  WARNING:tensorflow:Can save best model only with val_sp
arse_categorical_crossentropy available, skipping.
50/50 [==============================] - 5694s 115s/step - loss: 0.6626 - spars
e_categorical_accuracy: 0.8006 - val_loss: 0.3198 - val_sparse_categorical_accu
racy: 0.9112
```

Figure 17: EfficientNet-B1 Accuracy and Loss

The sample testing images of EfficientNet-B1 is shown below in Figure 18. We can see the model is correctly identifying the correct disease of banana plant.

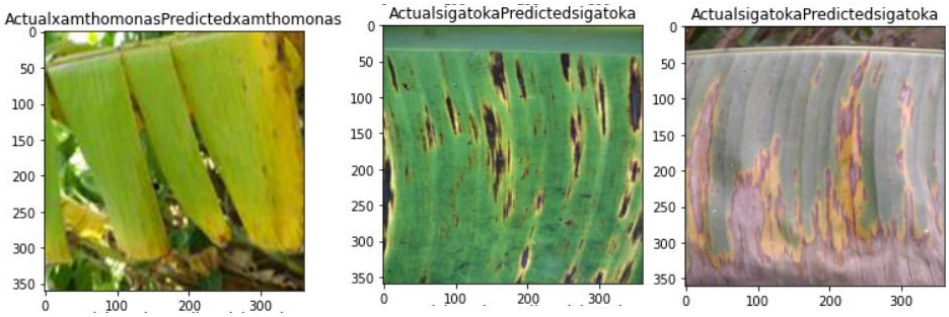

Figure 18: EfficientNet-B1 Testing images

6.2.2 DenseNet201

The DenseNet201 showed an f1-score of 95% which is higher than EfficientNet-B1 model. The validation loss is 0.20, which is again lower than EfficientNet-B1 model. The classification report and confusion matrix are shown in Figure 19. The DenseNet201 validation accuracy and loss is shown in Figure 20.

	precision	recall	f1-score	support
cordana	0.95	1.00	0.98	21
healthy	0.94	0.86	0.90	36
pestalotiopsis	0.91	0.94	0.93	33
sigatoka	0.99	0.95	0.97	83
xamthomonas	0.92	0.98	0.95	62
accuracy			0.95	235
macro avg	0.94	0.95	0.94	235
weighted avg	0.95	0.95	0.95	235

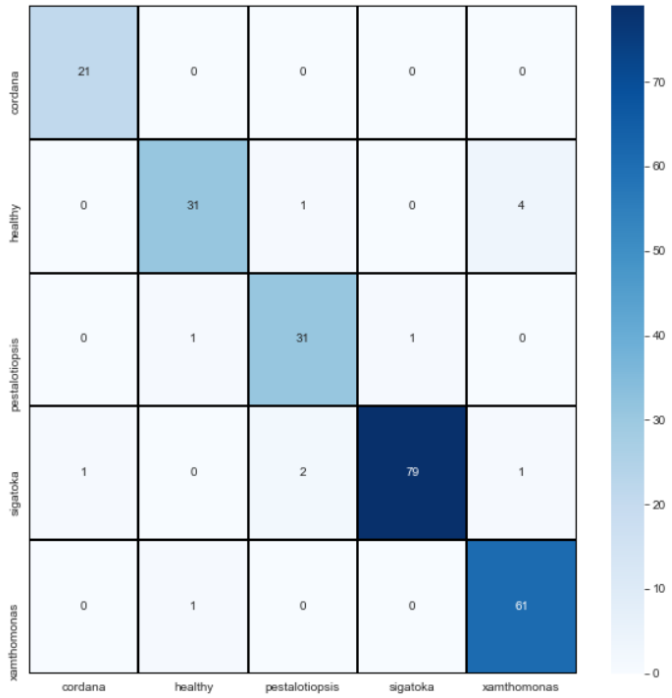

Figure 19: Classification report and Confusion Matrix

From the Classification Report, we see higher f1-score, precision and recall than other machine learning models. From the Training v Validation accuracy, we see the accuracy is increasing gradually while Training v Validation loss is decreasing gradually so the epoch is set to 12 to avoid overfitting of model. The total computational time for DenseNet201 model is 35 minutes which is lower than rest of the models.

Accuracy and Loss

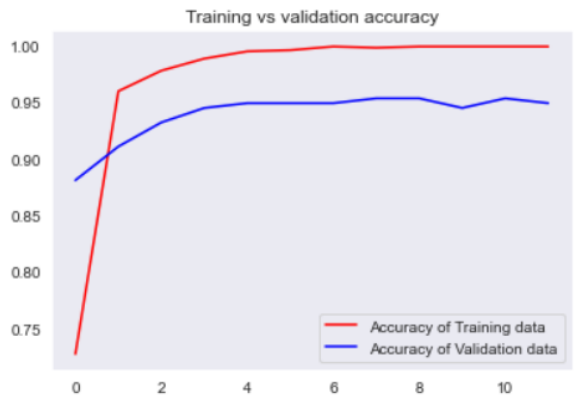

Figure 20(a): Training & Val Accuracy

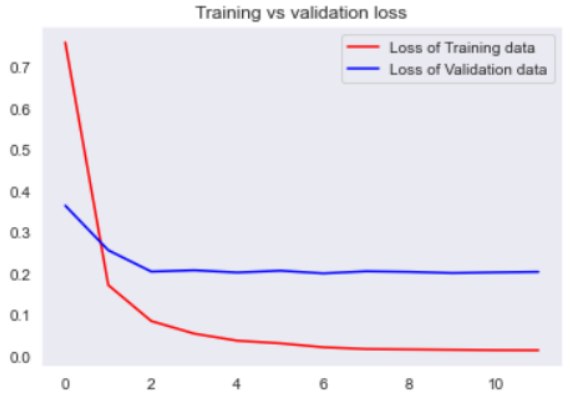

Figure 20(b): Training & Val Loss

The sample testing of the banana leaf disease using DenseNet201 is shown in Figure 21.

Figure 21: Sample test image using DenseNet201

6.2.3 VGG19

VGG19 showed highest accuracy among all the models implemented (EfficientNet-B1 and DenseNet201). It showed an accuracy of 95% with 94% precision and recall percentage. The validation loss is also lower than the other two models which is 0.164. The classification report and Confusion matrix is shown in Figure 22. The training and validation accuracy & loss is shown in Figure 23.

	precision	recall	f1-score	support
cordana	1.00	0.90	0.95	21
healthy	0.87	0.92	0.89	36
pestalotiopsis	0.91	0.91	0.91	33
sigatoka	0.99	0.98	0.98	83
xamthomonas	0.95	0.97	0.96	62
accuracy			0.95	235
macro avg	0.94	0.93	0.94	235
weighted avg	0.95	0.95	0.95	235

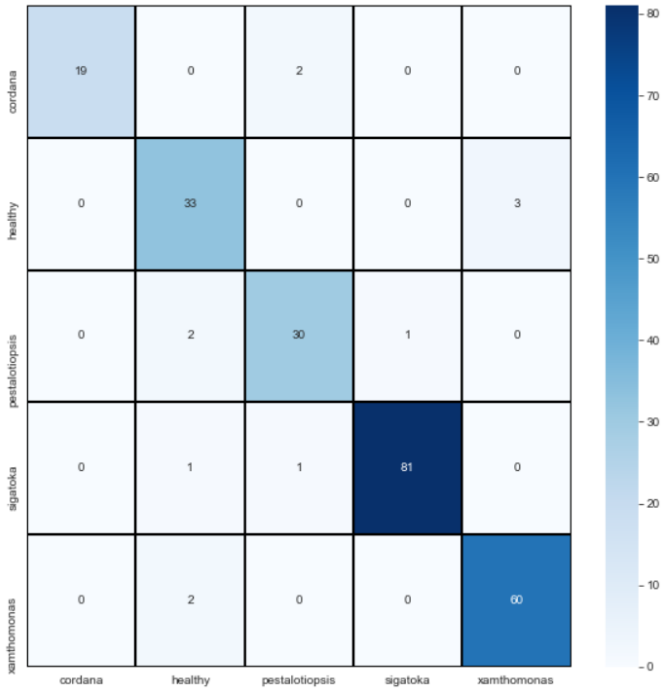

Figure 22: Classification report and Confusion matrix of VGG19

From the classification report, we see that VGG19 has the highest f1-score of around 94% and highest precision when compared with all the models. The computational time required to run the model is 45 minutes which is lower than EfficientNet-B1 model. The Training and validation accuracy and loss is shown in Figure 23. From the graph, there is no overfitting as the validation accuracy is increasing gradually and validation loss is decreasing gradually. The epoch is set to 12.

Accuracy and Loss

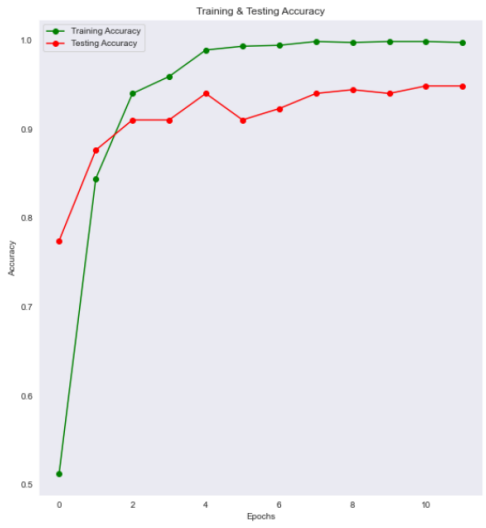

Figure 23(a): Training & Val Accuracy

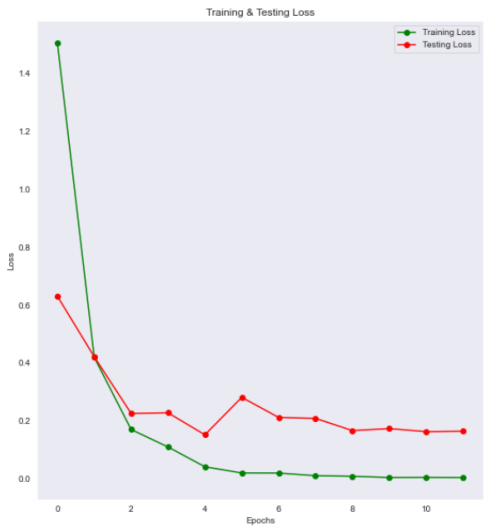

Figure 23(b): Training & Val Loss

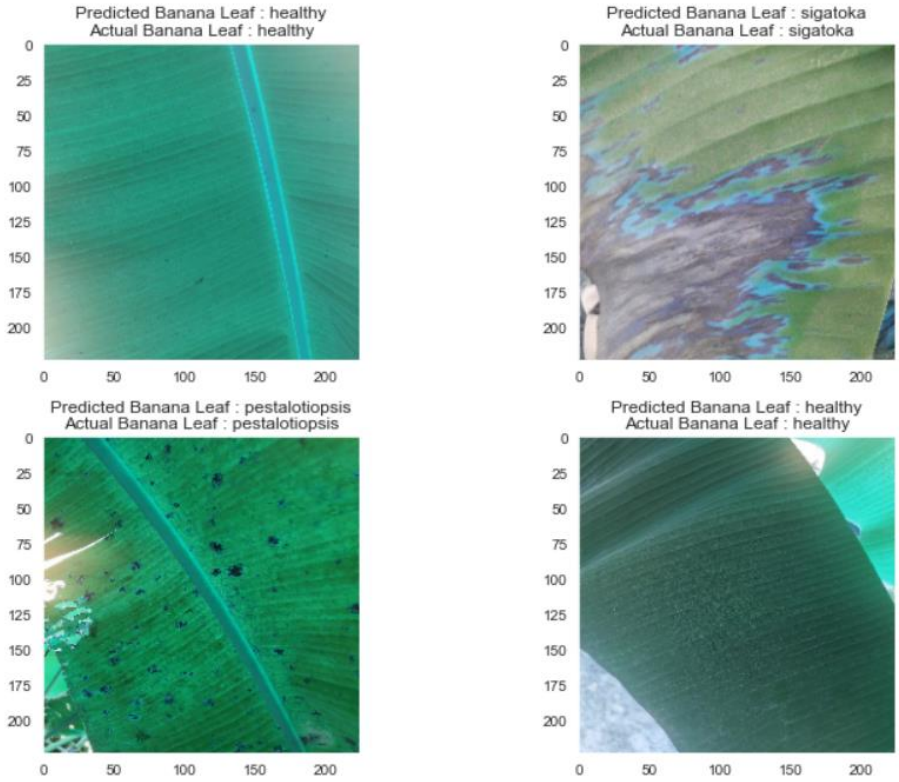

Figure 24: Sample test images using VGG19

The sample testing of banana leaf disease using VGG19 is shown below in Figure 24. Almost all the disease class are very well classified.

7 Discussion

The aim of the research was to identify and classify the banana leaf pest using machine learning models. These traditional statistical implemented models are compared with deep learning models with modern neural networks. Haralick's feature were used to implement the machine learning models. The challenges faced while evaluated these models were:

- Image dataset – As the data was gathered from two different data source, the biggest challenge was to merge the image of different resolution into a folder. The images were of high resolution so high processing power was required to classify the disease image and implement it using complex models.
- The time taken to run complex deep learning models was high. The time taken to run EfficientNet-B1, DenseNet201 and VGG19 was 50 minutes, 35 minutes, and 45 minutes respectively. So high computing power was required to implement the deep learning models.

Different techniques were executed to process the images conversion of image to grey scale, Gaussian and Adaptive thresholding method was used to segment the infected image from the background image, suing GLCM technique to extract four features from the image and implement machine learning models. The Figures 25, 26 ,27, 28 below shows the comparison evaluation of accuracy, precision, recall and f1-score respectively.

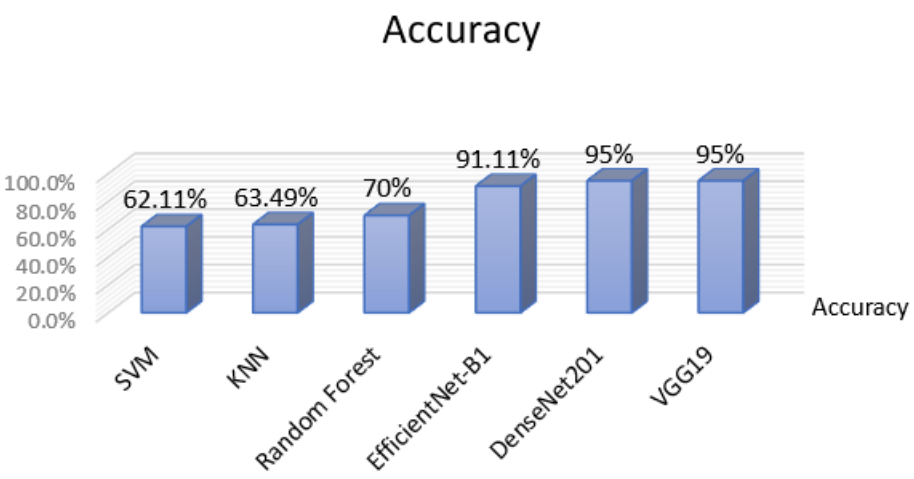

Figure 25: Accuracy Comparison

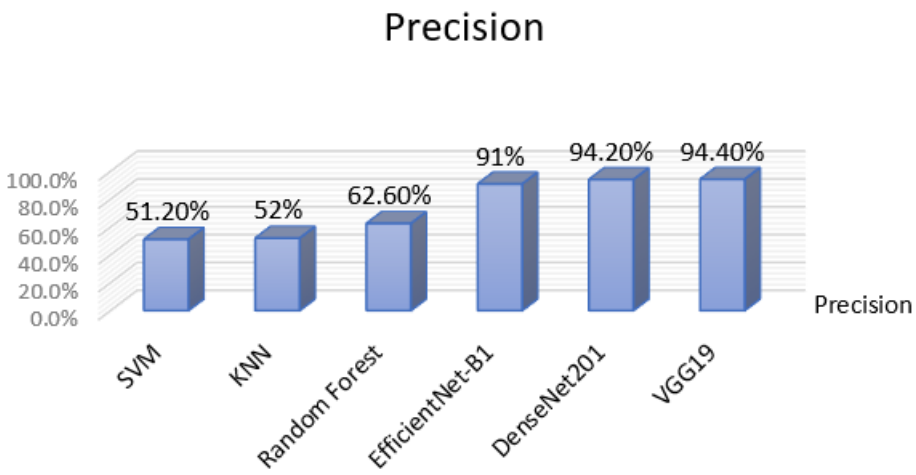

Figure 26: Precision Comparison

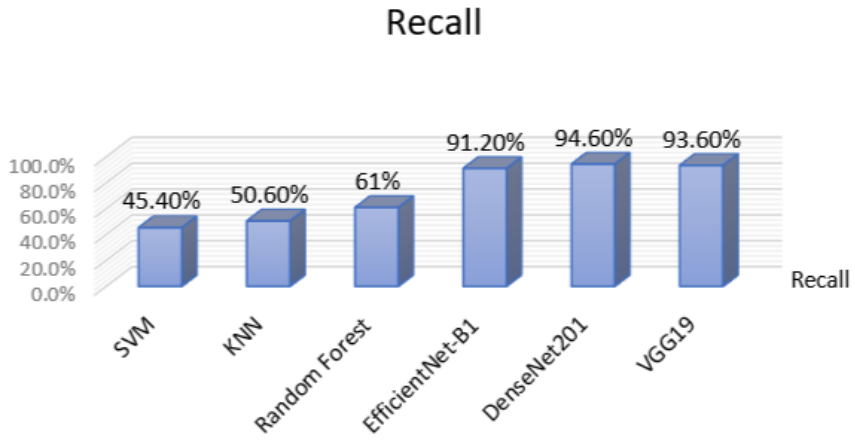

Figure 27: Recall Comparison

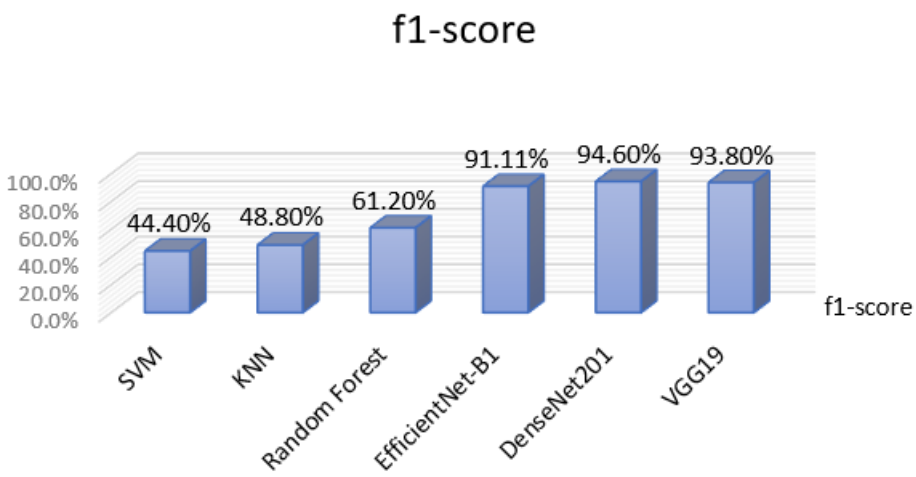

Figure 28: f1-score Comparison

Comparing with the reviewed papers is shown in below Table:

Author	**Plant type**	**Technique Used**	**Accuracy**
Vipinadas & Thamizharasi	Banana	SVM	No accuracy
Ramesh et al. (2018)	Random Plant leaves	HOF Feature Extraction Random Forest Naïve Bayes Support Vector Machine	 70% 57.61% 40.33%
Hu et al. (2016)	Potato	Hyperspectral Image	95%
Islam et al. (2017)	Potato	RGB Image	95%
Amara et al (2017) Poornam and Francis (2021)	Banana All Plant leaves	LeNet CNN	Grey-scale85% Color – 92% 89%
Yogesh Rokade (2022)	**Banana**	**EfficientNet-B1** **DenseNet201** **VGG19**	**92%** **95%** **95%**

Hence, comparing with the review papers, the model performed well considering the dataset was taken from two different data source and addition of new banana leaf diseases.

8 Conclusion and Future Work

The research project was carried out to identify and classify the banana leaf pests so that certain precautionary actions can be carried out to help farmers save from loss and eventually help increase the GDP of countries depending on agricultural sector. Traditional methods were visiting the field and examining with naked eye, the method was time taking and costly for farmers. So, by using machine learning technique farmers can just upload the banana leaf image to know the status of the leave whether it is healthy or diseased and if diseased, what kind of disease the banana leaf is infect with. Thus, saving a lot of time and money from the process and help produce good quality and quantity of food.

The accuracy obtained from traditional machine learning models: SVM, KNN and Random Forest classifier using GLCM as feature extraction method was 62.11%, 63.49% and 70% and the f1-score of all the models was around 44.40%, 48.80% and 61% respectively. Even though KNN and SVM have high accuracy but have low f1-score which means they less efficient than Random Forest Classifier which has higher accuracy and higher f1-score compared to the other two models. The Deep learning models showed an accuracy of 91.11%, 95% and 95% for EfficientNet-B1, VGG19 and DenseNet201 respectively. The f1-score of VGG19 and DenseNet201 was approximately same. So both these models can be used over other models to identify and classify banana leaf pest.

Although the implemented research project functioned effectively and were capable of achieving previous study work by including real environment field images but to make model more efficient more labels of disease can be added by visiting agricultural management sectors to train model with different diseases. Further by adding more diseases, analysis can be done like which is the most common and product damaging pest/disease on banana leaf. Application can be deployed so that it becomes easier for farmers to download the application and run it on their mobile phone. Also, hybrid neural network models can be implemented which require high end system/processor.

Acknowledgement

I would like to express my heartfelt gratitude to Prof. Abubakr Siddig, for all his guidance, support, assistance for the duration of my thesis period. His approach and recommendation in solving complex queries were the key factors that inspired me and kept me motivated throughout the Research Project. I would like to thank my family for their support and blessings all throughout my Masters. Special gratitude to my sister who trusted and believed in me while pursuing my Masters. Finally, I would like to thank all my friends for their encouragement throughout the research project.

References

Agarwal, A., Sarkar, A. and Dubey, A.K., 2019. Computer vision-based fruit disease detection and classification. In *Smart Innovations in Communication and Computational Sciences* (pp. 105-115). Springer, Singapore.

Alazawi, S.A., Shati, N.M. and Abbas, A.H., 2019. Texture features extraction based on GLCM for face retrieval system. *Periodicals of Engineering and Natural Sciences (PEN)*, *7*(3), pp.1459-1467.

Amara, J., Bouaziz, B. and Algergawy, A., 2017. A deep learning-based approach for banana leaf diseases classification. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband*.

Bandi, S.R., Varadharajan, A. and Chinnasamy, A., 2013. Performance evaluation of various statistical classifiers in detecting the diseased citrus leaves. *International Journal of Engineering Science and Technology*, *5*(2), pp.298-307.

Barbedo, J.G.A., 2019. Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering, 180, pp.96-107.

Bhimte, N.R. and Thool, V.R., 2018, June. Diseases detection of cotton leaf spot using image processing and SVM classifier. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 340-344). IEEE.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R., 1999, March. The CRISP-DM user guide. In *4th CRISP-DM SIG Workshop in Brussels in March* (Vol. 1999). sn.

Chaudhari, V. and Patil, M., 2020, August. Banana leaf disease detection using K-means clustering and Feature extraction techniques. In *2020 International Conference on Advances in Computing, Communication & Materials (ICACCM)* (pp. 126-130). IEEE.

Dey, N., Zhang, Y.D., Rajinikanth, V., Pugalenthi, R. and Raja, N.S.M., 2021. Customized VGG19 architecture for pneumonia detection in chest X-rays. *Pattern Recognition Letters*, *143*, pp.67-74.

Dhingra, G., Kumar, V. and Joshi, H.D., 2018. Study of digital image processing techniques for leaf disease detection and classification. *Multimedia Tools and Applications*, *77*(15), pp.19951-20000.

Duong, L.T., Nguyen, P.T., Di Sipio, C. and Di Ruscio, D., 2020. Automated fruit recognition using EfficientNet and MixNet. *Computers and Electronics in Agriculture*, *171*, p.105326.

Haralick, R.M., Shanmugam, K. and Dinstein, I.H., 1973. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), pp.610-621.

Hatuwal, B.K., Shakya, A. and Joshi, B., 2020. Plant Leaf Disease Recognition Using Random Forest, KNN, SVM and CNN. *Polibits*, *62*, pp.13-19.

Hossain, E., Hossain, M.F. and Rahaman, M.A., 2019, February. A color and texture based approach for the detection and classification of plant leaf disease using KNN classifier. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-6). IEEE.

Hu, M.H., Dong, Q.L., Liu, B.L. and Malakar, P.K., 2014. The potential of double K-means clustering for banana image segmentation. *Journal of Food Process Engineering*, *37*(1), pp.10-18.

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

Indriani, O.R., Kusuma, E.J., Sari, C.A. and Rachmawanto, E.H., 2017, November. Tomatoes classification using K-NN based on GLCM and HSV color space. In *2017 international*

conference on innovative and creative information technology (ICITech) (pp. 1-6). IEEE.

Islam, F., Hoq, M.N. and Rahman, C.M., 2019, November. Application of transfer learning to detect potato disease from leaf image. In *2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)* (pp. 127-130). IEEE.

Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V. and Kaur, M., 2021. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, *39*(15), pp.5682-5689.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), pp.255-260.

Joshi, A.A. and Jadhav, B.D., 2016, December. Monitoring and controlling rice diseases using Image processing techniques. In 2016 International Conference on Computing, Analytics and Security Trends (CAST) (pp. 471-476). IEEE.

Kaur, S., Pandey, S. and Goel, S., 2018. Semi-automatic leaf disease detection and classification system for soybean culture. *IET Image Processing*, *12*(6), pp.1038-1048.

Koonce, B., 2021. EfficientNet. In *Convolutional neural networks with swift for tensorflow* (pp. 109-123). Apress, Berkeley, CA.

Kumar, D., 2020. Feature extraction and selection of kidney ultrasound images using GLCM and PCA. *Procedia Computer Science*, *167*, pp.1722-1731.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S. and Bochtis, D., 2018. Machine learning in agriculture: A review. *Sensors*, *18*(8), p.2674.

Mohanty, S.P., Hughes, D.P. and Salathé, M., 2016. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, *7*, p.1419.

Mokhtar, U., Ali, M.A., Hassanien, A.E. and Hefny, H., 2015. Identifying two of tomatoes leaf viruses using support vector machine. In *Information systems design and intelligent applications* (pp. 771-782). Springer, New Delhi.

Nagayets, O., 2005. Small farms: current status and key trends. The future of small farms, 355, pp.26-29.

Panigrahi, K.P., Das, H., Sahoo, A.K. and Moharana, S.C., 2020. Maize leaf disease detection and classification using machine learning algorithms. In *Progress in Computing, Analytics and Networking* (pp. 659-669). Springer, Singapore.

Pantazi, X.E., Moshou, D. and Tamouridou, A.A., 2019. Automated leaf disease detection in different crop species through image features analysis and One Class Classifiers. *Computers and electronics in agriculture*, *156*, pp.96-104.

Poornam, S. and Francis, S.D.A., 2021. Image based Plant leaf disease detection using Deep learning. *Int. J. Comput. Commun. Inf*, *3*, pp.53-65.

Rahamathunnisa, U., Nallakaruppan, M.K., Anith, A. and KS, S.K., 2020, March. Vegetable disease detection using k-means clustering and svm. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1308-1311). IEEE.

Ramesh, S., Hebbar, R., Niveditha, M., Pooja, R., Shashank, N. and Vinod, P.V., 2018, April. Plant disease detection using machine learning. In *2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C)* (pp. 41-45). IEEE.

Raut, S.P. and Ranade, S., 2004. Diseases of banana and their management. In *Diseases of Fruits and Vegetables: Volume II* (pp. 37-52). Springer, Dordrecht.

Shrivastava, S. and Hooda, D.S., 2014. Automatic brown spot and frog eye detection from the image captured in the field. *American Journal of Intelligent Systems*, *4*(4), pp.131-134.

Simon, C.M.L., 2011. The Role of Irrigation in Banana Production in St. Vincent and the Grenadines. In *29th West Indies Agricultural Economics Conference, July 17-21, 2011, Saint Vincent, West Indies* (No. 187832). Caribbean Agro-Economic Society.

Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR.

Tian, L., Xue, B., Wang, Z., Li, D., Yao, X., Cao, Q., Zhu, Y., Cao, W. and Cheng, T., 2021. Spectroscopic detection of rice leaf blast infection from asymptomatic to mild stages with integrated machine learning and feature selection. Remote Sensing of Environment, 257, p.112350.

Too, E.C., Yujian, L., Njuki, S. and Yingchun, L., 2019. A comparative study of fine-tuning deep learning models for plant disease identification. *Computers and Electronics in Agriculture*, *161*, pp.272-279.

Tumang, G.S., 2019, July. Pests and diseases identification in mango using MATLAB. In *2019 5th International Conference on Engineering, Applied Sciences and Technology (ICEAST)* (pp. 1-4). IEEE.

ur Rahman, H., Ch, N.J., Manzoor, S., Najeeb, F., Siddique, M.Y. and Khan, R.A., 2017. A comparative analysis of machine learning approaches for plant disease identification. *advancements in life sciences*, *4*(4), pp.120-126.

Vipinadas, M.J. and Thamizharasi, A., 2016. Banana leaf disease identification technique. *International Journal of Advanced Engineering Research and Science*, *3*(6), p.236756.

Yu, X., Zeng, N., Liu, S. and Zhang, Y.D., 2019. Utilization of DenseNet201 for diagnosis of breast abnormality. *Machine Vision and Applications*, *30*(7), pp.1135-1144.