Detection and Classification of Leaf Diseases in Maize Plant using Machine Learning

MSc Research Project
Data Analytics

Adarsh Jayakumar
Student ID: x18131379

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Adarsh Jayakumar
Student ID:	x18131379
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Cristina Muntean
Submission Due Date:	12/12/2019
Project Title:	Detection and Classification of Leaf Diseases in Maize Plant using Machine Learning
Word Count:	7500
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.

Signature:	
Date:	28th January 2020

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):	

Detection and Classification of Leaf Diseases in Maize Plant using Machine Learning

Adarsh Jayakumar
x18131379

Abstract

One of the major challenges faced by agricultural industry is the need for accurate and early detection of diseases that affect crops. Diseases affect the quality of crops and are capable of wiping out hectares of crop yield resulting in major loss to farmers. Current diagnostic techniques are time consuming and require the presence of highly skilled professionals to analyze the affected plants, understand the symptoms, identify the disease and thereby suggest suitable remedies. The limitations of such techniques have enforced the need to look for alternative techniques which can detect and classify diseases at an early stage. Smart farming using suitable infrastructure can help in tackling and providing solutions to such problems. Data mining techniques, in the recent years have shown great promise in identifying and classifying patterns in similar areas of research. Current research aims to evaluate the performance of algorithms like XGBoost, Gradient Boost, Convolutional Neural Network(CNN) and its architectures like VGG16 and VGG19 coupled with data augmentation and transfer learning against traditional machine learning algorithms like Support Vector Machine(SVM), random forest and measure their effectiveness in identifying and classifying maize plant diseases in terms of accuracy, precision, recall and training time. Training of the models were performed on an open source database containing close to 4000 images, encompassing four distinct classes, including healthy plant images. Out of the models developed, VGG19 architecture of CNN with transfer learning performed the best by achieving an overall accuracy of 95 percent, thereby satisfying the need of building an effective and robust classification model. Also, the performance of the models developed were found to improve with increase in amount of training data. The results obtained using transfer learning techniques on CNN architectures are highly promising and can be extended further to form a comprehensive plant disease identification system that is capable of operating in real world environment. It can thus empower the agricultural community to diagnose diseases and initiate timely treatment without the intervention of trained experts.

1 Introduction

Agriculture is a lifeline for a vast majority of the population in the world with close to 70 percent[1] of people directly dependent on it as means for living. The crops grown by farmers in different regions across the world are mainly based on weather, yield potential, type of soil, etc. Of the crops grown, maize , alternatively known as corn is largely

[1] `https://agriculturegoods.com/why-is-agriculture-important/`

cultivated worldwide and is the third leading crop of the world after rice and wheat, mainly because of its nutritional value and yield potential. Close to 1100 million metric tons were produced last year with countries like India, China and USA being the top cultivators. Although, the numbers look quite impressive, it doesn't paint the full picture as high volumes of the produce are also lost due to various factors.

A good yield not only helps the cultivators, but also significantly boosts the economic growth of the country. However, achieving desired levels of yield is challenging as it is influenced by various aspects like climate, pests, diseases amongst many others. Maize, like majority of the crops grown is very sensitive to such factors with close to 35 percent of its production lost to diseases and pests on average every year(Oerke;2006). Northern leaf blight, Common rust and gray leaf spots are some of the diseases that can create major havoc to these plants. The afore mentioned statistics strongly indicate the need for an early and effective detection of diseases in these plants as negligence and delay can lead to significant losses. Many researches are being done off late in order to tackle these issues and this is currently a research hotspot in the agricultural community.

Traditionally, visual observation by experts were done to diagnose the plant diseases. However, they carried a high risk of subjective perception and were time consuming in nature. In due course of time, spectroscopic and imaging techniques were used with researches like Lu et al. (2017) making use of hyperspectral data of strawberry leaves for identification of diseases. They obtained accuracies close to 70 percent. However, these methods required precision instruments and bulky sensors for analysis which in turn placed the need for expert intervention. Digitization and evolution of machine learning techniques that can detect underlying patterns have been become popular alternatives to diagnose plant diseases in recent years. Conventional classification algorithms like Support Vector machine(SVM), K means clustering , etc. coupled with complex preprocessing and feature extraction techniques have been used and have produced satisfactory results. The rapid advancements and research in this domain have led to development of new brand of models and techniques called deep learning. The introduction of these deep learning techniques into agriculture, and in particular into the field of disease diagnosis, has only started a couple of years back and to a rather limited extent.

The main aim of this research is to build deep learning models based on CNN which shall make use of transfer learning with trainable layers coupled with image augmentation in order to build an efficient model that is capable of generalizing well and producing high levels of accuracy , thus helping us find answers to question like "To what extent are algorithms like Gradient Boost and CNN architectures based on transfer learning effective in identifying and classifying diseases in maize plants?". To the best of my knowledge, these techniques are yet to be fully explored in maize disease classification and can prove to be of great value to farmers to diagnose plants, if desirable results are obtained. Some of the main objectives of this research are:

- To effectively identify and classify maize plant diseases using machine learning algorithms

- To build various models like SVM, Random Forest(RF), Gradient/XGBoost and CNN architectures based on transfer learning which are coupled with data augmentation and identify the best model that can be used to provide solutions to the agricultural community.

- To analyze various train/test split ratios, hyper tune models and analyse their impact on results obtained.

The document is structured in a systematic way where in section 2 is based on understanding and critiquing the related works in the field that support the need for the methodology chosen for current research.Section 3 briefs about the various stages of research development. General design of various models built is discussed in section 4 followed by their implementation in section 5. Evaluation of the results and conclusion is summarized in section 6 and 7 respectively with acknowledgement following them.

2 Related Work

Plant disease detection is an age-old problem that is haunting the agricultural community. Different research approaches have been carried out to accurately detect and classify plant disease of various plants. These include the conventional on-field inspection by experts to the use of spectroscopy and image processing techniques and the more recent use of machine and deep learning algorithms. However, the advantages of these methodologies are not fully utilized due to high cost in case of some methods and lack of robustness in others. Transfer learning and deep learning techniques stack up well against these odds and the current project is implemented using these methods to accurately detect and classify maize disease. Before diving in detail into the specifics of the project, let us understand some of the methodologies used previously and understand their benefits and shortcomings.

2.1 Spectroscopy and its evolution in agriculture domain

Initial researches in the field made use of spectroscopy to identify and classify diseases. Spectral frequencies of plants were analyzed as infected plants exhibited different characteristics in comparison to healthy plants due to difference in amount of light absorbed in the near-infrared(NIR) range which helped in detecting anomalies.

This method was used by Zhang et al. (2012) in his research to detect powdery mildew in wheat plants. Correlation, regression and independent sample t tests were used to analyze the 32 spectral features extracted. As the whole analysis was based on spectral data, it was not effective in early detection of diseases that affected visual characteristics of plant before altering its internal structure.

New technologies named multispectral and hyper spectral imaging aimed at reducing the shortcomings of the previous method by combining spectroscopy and imaging evolved in due course. In research on strawberry leaves by Lu et al. (2017), hyperspectral data was used and algorithms like k-Nearest neighbor(kNN) and Fisher discriminant analysis(FDA) were utilized to detect if the plants were infected by anthracnose crown. Even though they were able to achieve an accuracy of 70 percent, the research had several drawbacks which included the tedious process of data collection using spectroradiometer that necessitated the need for experts that operate it. Further, the data collected can also be flawed due to shadows cast by sunlight leading to reduced performance as stated by the authors.

In another research on wheat by Zheng et al. (2019), hyperspectral data was used for analysis as it was found that spectral data and indices like PRI (Photochemical Reflectance Index) and ARI (Anthocyanin Reflectance index) changed with plant development. The results obtained from their study were better in comparison to previous researches in the field.

Various parameters were tuned and tested in the quest for improving accuracy by other researchers but the inherent problem of data collection and expert knowledge to understand the results remained unaddressed , stressing the need for alternative technologies.

2.2 Use of machine learning in disease identification and classification

Machine learning is a technical advancement that can be used to tackle various problems related to predictions or regressions. This has paved way for its use in different domains and has produced good results. This section shall discuss about its use in disease classification and identify its benefits and limitations.

Some of the algorithms in machine learning that were popular for classification were kNN and SVM and have been used widely for disease classification in plants. As previous researches in the field had placed too much stress on data collection, researchers Dhaware and Wanjale (2017) have tried to mitigate this by making use of images obtained from handheld devices for analysis and incorporating image processing in their methodology which involved segmentation. SVM was used for classification. However, results obtained indicated scope for improvement with usage of techniques like feature extraction.

An accuracy of 90 percent was obtained by Francis et al. (2016) when composed segmentation was used to identify diseases like quick wilt and berry spot in pepper plants. It involved green pixel masking and similarity-based segmentation which was followed by feature extraction. Propagation neural network was used for classification. As the analysis was carried out on small dataset, the results are not totally reliable.

Researchers, Padol and Yadav (2016) made use of k- means algorithm for segmentation. Noises were removed from images using gaussian filter and around 54 features were extracted from it for analysis. This was however time consuming and tedious, and yielded results close to 90 percent making us wonder if such complex procedures had any significant impact.

Global-singular value decomposition(SVD) was used by researchers Zhang and Wang (2016) in their quest to identify cucumber diseases. It proved to be effective in feature extraction. SVM, coupled with an improved recognition method based on watershed algorithm helped them achieve good and accurate results. The whole analysis was conducted on a small dataset due to computational complexity of the techniques involved.

Feature extraction was done using SURF by Aravind et al. (2018) and then k-means was used for clustering them. Gray scale image and its occurrence matrix with histogram served as classifier's input in the study. A variation of this study by Maniyath et al. (2018) involved the use of Histogram of an Oriented Gradient (HOG) as feature extractor. This made use of Hu moments, texture and color histogram for analysis. Both studies produced good results and supported the use of random forest when the amount of data available for analysis is small.

In general, one can comprehend that all the above researches made use of relatively small datasets for analysis and are heavily dependent on pre-processing for accurate classification. This is computationally expensive and requires high domain knowledge. These seem to be the main reasons why we are observing a paradigm shift in recent years as more and more work in this field are making use of deep learning techniques instead of traditional machine learning techniques. These are capable of addressing the bottlenecks and provide better results. Related works on deep learning is scrutinized in the next section.

2.3 Importance of deep learning and transfer learning techniques

Deep learning is a new breed of machine learning algorithm that do not need much preprocessing for analysis. They are capable of extracting features on their own making the process of disease identification simpler.

Studies using transfer learning include the research conducted by Shrivastava et al. (2019) on rice plant. Here CNN architectures were used as feature extractors and output of these were used by SVM for classification. Models produced accuracies of 92 which is commendable. However, the dataset used was of size 619 which is too small for analysis. The limitations of these were addressed making use of data augmentation by authors Coulibaly et al. (2019) which improved the results to a certain extent. Using CNN as feature extractors can however be computationally intensive.Further, the training time of models are not discussed by authors which could have helped us understand their utilities better.

In research on rice plants by Lu et al. (2017), CNN with different number of layers were utilised to identify and classify rice diseases. Although, the techniques achieved good accuracy, the tests were performed on a small dataset of 500 images.Hence, the results obtained must taken with a pinch of salt and further analysis must be done on the same.

An extension of the above research was done by Ferentinos (2018) by making use of dataset of different species of plants. Architectures like Alexnet,VGG,Googlenet which are based on CNN were trained from scratch and resultant models obtained accuracies of about 99 which is mind boggling. In similar lines of analysis by Mohanty et al. (2016), transfer learning versions of Alexnet and Googlenet were used for classification of various plant diseases resulting in accuracies similar to previous study. However, both these models failed miserably when tested on datasets other than the ones under study. The authors suggest that these can be mitigated using data augmentation and training some layers of the architectures with problem specific data which forms the core principle of our research.

2.4 Boosted Trees for image classification

A relatively new breed of algorithms that makes use of boosted trees, their potential is yet to be fully explored. In intial research conducted using these techniques by Gao et al. (2017) for object classification, these techniques performed good and produced accuracies close to 92 percent. Xu and Wang (2019) in their risk analysis of diabetes 2 made use of random forest and xgboost for analysis. The results obtained were highly promising and were close to those obtained by Gao et al. (2017) supporting their use in current research for classifying and identifying plant disease.

The current research aims to address the shortcomings of previous works in this area by making use of data augmentation and transfer learning using trainable layers. The performance of new algorithms like Gradient and XGboost shall be compared against these to find an optimal model that is useful in identifying and classifying maize disease. The methodology and implementation of these techniques is elaborated in further sections.

3 Methodology

At the beginning of a project, its essential to choose a right methodology in order to execute the tasks of a project in an organized manner. The methodology chosen goes a long way in determining the success of the project and therefore, one has to carefully evaluate the various available options before selecting one among them. Data mining projects have an array of methodology choices like Knowledge Discovery in Databases (KDD), CRISP-DM (Cross-industry standard process for data mining), Sample Explore Modify Model and Assess (SEMMA) to name a few. The current research shall make use of KDD as it focusses more on discovering patterns and finding information within the given data rather than describing and understanding the business importance of data. This process seems tailored for academic projects in contrast to commercial projects undertaken by industries where business aspects of data takes priority resulting in CRISP-DM being one of the overwhelming favorites. They also do not stress much on coding and one can proceed to coding phase soon as the business needs of the project are grasped in advance leading to faster implementations.

Current project shall make use of image data of maize leaf which is organized into various folders based on diseases for classification. This data shall be loaded and preprocessed by resizing to a standard size which is later split into test and train data. Models are developed on training data that are also augmented in order to prevent overfit and further generalize the data, thus enabling them to predict accurately on unseen data. The performance of these models is then evaluated on test data. The detailed explanation of each of the steps followed under KDD is provided in following subsection. The process flow followed is depicted as shown in Figure 1.

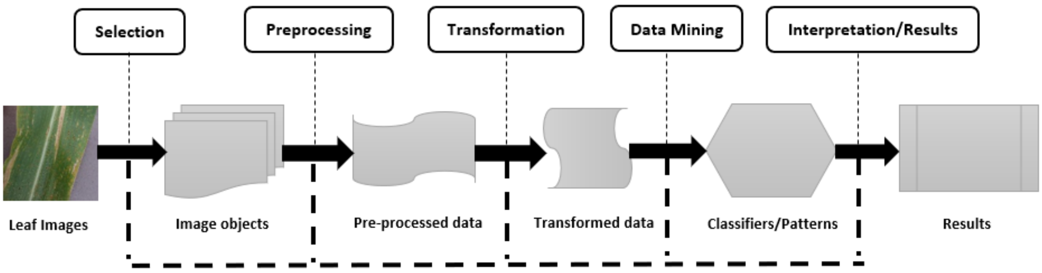


Figure 1: Process flow of KDD

3.1 Data Selection

One of the major challenges of research projects in agricultural domain is the lack of data that are publicly usable and is discussed by authors Arsenovic et al. (2019) in their study. Most of the researches done are on proprietary data that are collected by authors. This is a time consuming and expensive proposition in itself. It is only recently that a public project initiative called Plant Village[2] made such data available for public. It contains leaf images of various plants that are healthy or are affected by diseases, categorized and

[2] https://github.com/spMohanty/PlantVillage-Dataset

labelled into various folders. This data is available on GitHub and has been used for analysis by various authors. Current research only uses data related to maize plant from the above dataset. Although the data is available in various formats like colored, grey scale and segmented; only colored data has been used for the study as they seem to have produced better results in many previous researches Mohanty et al. (2016). The data used has images that belongs to four different categories namely gray leaf spot, common rust, northern leaf blight and healthy maize and their sample images are as shown in Figure 2

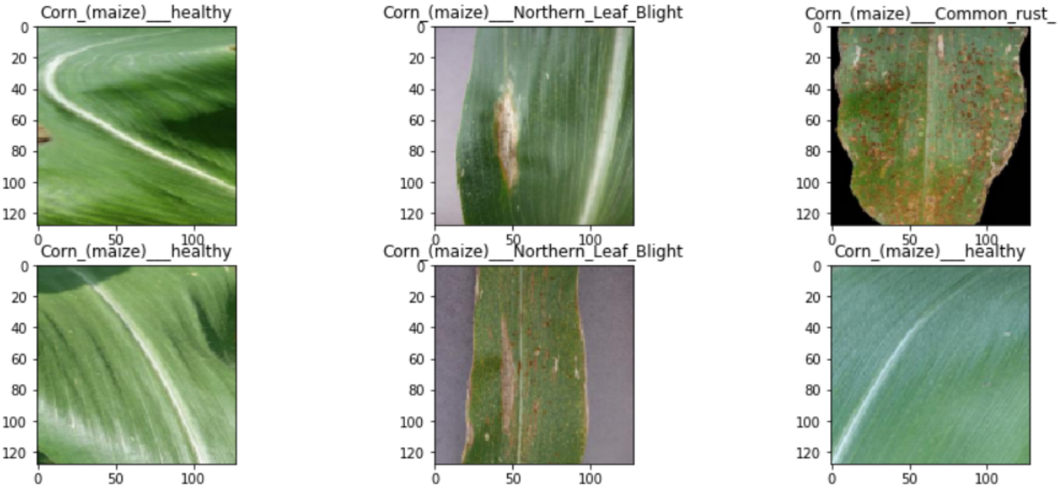

Corn_(maize)___healthy
Corn_(maize)___Northern_Leaf_Blight
Corn_(maize)___Common_rust_
Corn_(maize)___healthy
Corn_(maize)___Northern_Leaf_Blight
Corn_(maize)___healthy
0
20
40
60
80
100
120
0
50
100

Figure 2: Maize Disease Images

3.2 Data Pre-Processing

The dataset on the whole had data related to 14 different plant species. As the current research focusses only on maize plant, other folders were removed from the data repository resulting in a sample size of 4000 images spread across four folders as shown in Figure 3. These images were then loaded by recursively parsing through these folders and reading image files located in them. This was followed by resizing them to satisfy the dimensionality constraints of various algorithms and to provide uniformity so that the classification algorithms perform well. They were also labelled based on the class they belonged using python libraries like Label Encoder. Further, a bar chart was plotted to check for distribution of images in various folders. It showed that data was almost uniformly spread across all the classes which would enable us to build efficient models as the data is not biased. Models were then built on this data and validated for test train splits of 70:30 and 80:20.

3.3 Data Transformation

A variety of data transformations have been used as for the models to avoid overfitting on the train dataset provided and to be able to provide high validation accuracies on the test dataset. The loaded images are initially converted to NumPy arrays containing their RGB values and normalized so that classification models can process them. Additionally, image augmentation technique has been utilised in the current research and is implemented

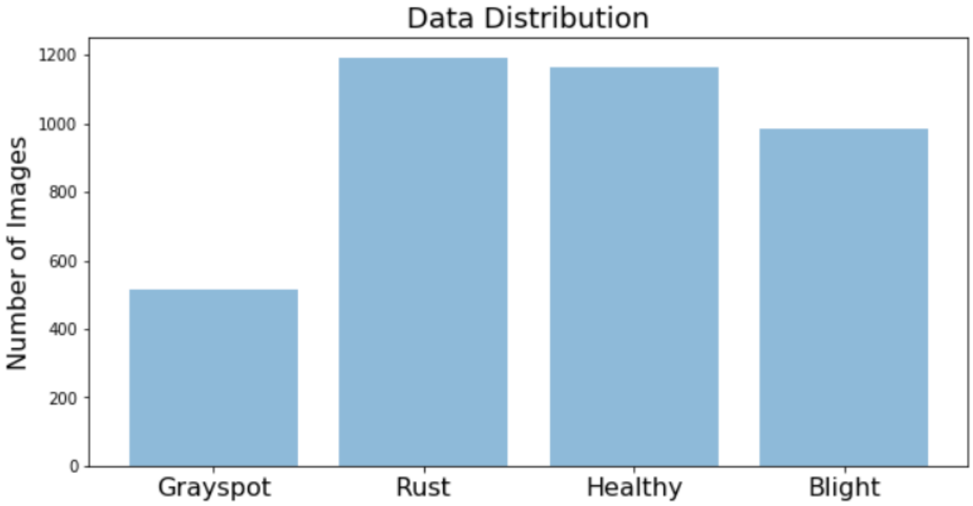


Figure 3: Data distribution across different classes

making use of keras image preprocessing module called ImageDataGenerator. Some of the features used from this module include:

- rotationrange: Used to rotate the loaded image by the amount of degrees specified
- widthshiftrange: Move the image along the horizontal axis with typical values lying between 0 and 1
- heightshiftrange: Move the image along the vertical axis with typical values lying between 0 and 1
- shearrange: Used to stretch the image
- zoomrange: Zooms randomly on certain areas of image, enabling algorithms to train better on those highlighted characteristics
- horizontalflip: Used to flip image horizontally. Very useful in generalization of data
- fillmode: Indicates how the newly created pixels has to be filled with 'nearest' being the most common technique

Augmentation has only been applied on training data so that the deep learning algorithms developed do not overfit the training data and hence can generalize well on unseen test data which was a major issue for researches as mentioned in section 2.2.

3.4 Data Mining

Different models were developed and evaluated to find an optimum model that performs better in comparison to others, both in terms of accuracy and also performance. The models that were built were based on two approaches. First approach involved developing models like SVM, RF , Gradient Boost and XGBoost from scratch by training them purely on data used in the project. The second approach involved development of deep learning models derived from CNN named VGG with different number of layers. These utilize transfer learning techniques and have pretrained weights from image net dataset. Only the final few layers were retrained with the data used in the project and used for analysis. Further, the parameters of all the implemented models were tuned using grid search technique or by running several tests to find the optimum model that can identify and classify maize diseases. Literature reviews indicate that Gradient Boost and transfer learning with retrained layers are yet to be explored in this domain of study. These have been implemented to check if they can provide better performance than the currently existing models.

3.5 Evaluation

As the research is based on classification, metrics that are used to evaluate model performance include accuracy. precision recall and f-1 score. All the developed models, i.e. those built from scratch and those built using transfer learning are compared based on these metrics to determine the best model. A brief explanation of what each of these metrics mean in the context of current project is described below.

- Accuracy[3]: It is a de-facto standard in evaluating the efficiency of models that are based on classification and indicates the number of predictions that are accurately made by the model over the total number of samples. It is an indication of the percentage of test samples that were accurately classified by the model . It is an important metric and enables us to find answers to some of the research objectives that were stated above.

- Precision[4]: It is ratio of number of cases in which the model could accurate predict the class of test data over the sum of predictions that were in fact correct and those that the model assumed to be correct but was wrong and is mathematically calculated as shown in Equation (1)

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{1}$$

- Recall[5]: Another evaluation metric that shows the percentage of cases that model was correct in predicting over the sum of correct predictions and incorrectly classified cases as shown in Equation(2)

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

- F1-Score[6]: Both precision and recall are calculated for every class available in the data as this project is a multiclass classification problem. The harmonic mean of both these values form the F1 score of the model and is a robust way to compare the models that are developed. Mathematically, it is calculated as shown in Equation (3)

$$F1 - Score = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \tag{3}$$

[3] https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

[4] https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

[5] https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

[6] https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

4 Design Specification

The project makes use of a two-tier design as shown in Figure 4 as the dataset used is publicly available and has not been created specifically for the project using any api/cloud-based approach, thus eliminating the need for a three-tier design. Python was chosen as the main tool for development, mainly because of its ease of use and also the availability of a large number of libraries like keras that can be easily utilized to achieve the functionalities and build models indicated in the figure. Jupyter notebook served as the Integrated Development Environment (IDE) in business logic where pre-processing and transformation of data was done.This was followed by implementation of models like SVM, RF, Gradient Boost, XGBoost and CNN architectures like VGG16 and VGG19 that are based on transfer learning. The results obtained are then visualized in the presentation layer using matplotlib libraries of python enabling us to easily compare and evaluate the models that are developed. The general architecture of different models developed in business logic tier are briefly described below.

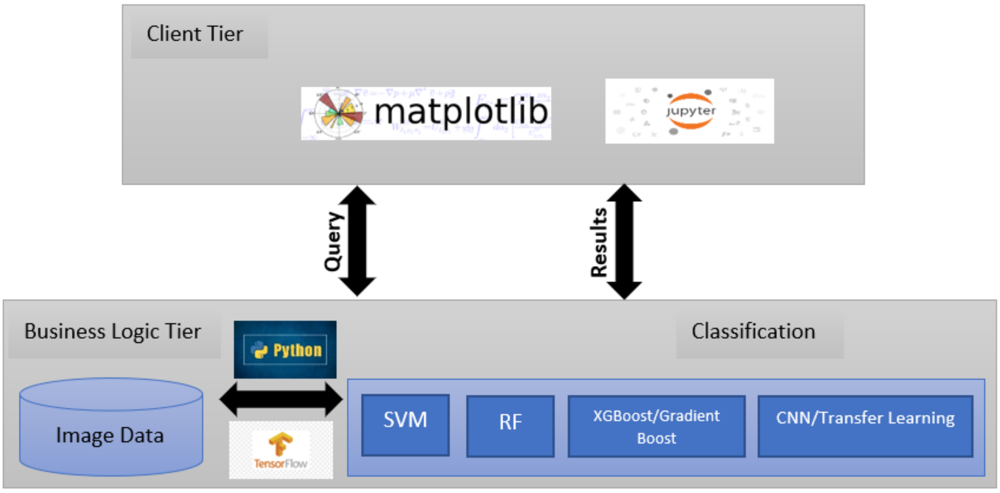


Figure 4: Design Strategy

4.1 Support Vector Machine

It is part of supervised learning technique which is used for handling both regression and classification problems. The main aim of this algorithm is to find a hyperplane that best separates the features of different classes under consideration. The data points close to the hyper plane are called support vectors and the algorithm tries to maximize this distance by using the best fit line. If the classes are not linearly separable, as is the case with maize disease classification, it tries to achieve better results by using soft margins and kernel tricks with polynomial and rbf kernels being the most popular.

4.2 Random Forest

It is an ensemble model that contains several decision trees as building blocks. The whole concept of this algorithm is based on random sampling of training data and the reason they are quite popular is because they are robust and do not overfit training data. Decisions are based on large number of uncorrelated trees and this would outperform individual constituent models.

4.3 Gradient Boost and Extreme Gradient Boost (XGBoost)

Both of these models are ensemble models based on decision trees that are used for performing supervised learning tasks like classification and regression. They are based on gradient boosting architecture and use weighted average technique to make a strong learner from a collection of weak learners. They work iteratively and the new models developed are trained to minimize the errors obtained from previous models. This is continued until the errors cannot be further reduced. One of the optimized versions of gradient boost that was developed recently is the XGBoost. As they are boosted and do not overfit, they seem tailored for classification and is used for identifying and classifying maize disease.

4.4 Deep Learning based on CNN

Convolutional Neural Network, popularly known as CNN is a part of many deep neural models that are utilized for image classification and is inspired by biological neurons that empower the brain. While normal machine learning models discussed above can be improved in terms of execution time by using many feature extraction techniques like Histogram of Oriented Gradients(HOG) , Scale Invariant Feature Transform(SIFT); they are complex, computationally intensive and need expert knowledge to be implemented. They tend to place a lot of importance for image pre- processing which may not be feasible when looking at real world deployment of machines that can identify and classify disease. Further, these techniques may fail to separate the leaf from its natural background leading to results that are unreliable. CNN helps bypass the process of feature extraction and segmentation as they are capable of extracting features from the provided image using various layers and are proven to be more effective than other algorithms in similar areas of study. The classification step here makes use of fully connected layers where each neuron provides connection to all learned feature maps. In order to accurately compute the class scores, the connected layers are based on SoftMax activation function. This function takes a vector of features which have been learnt from the learning process of the training data set and outputs the probability of the test image belonging to one of the classes. Functionalities of some of the layers used in the design are briefly described below:

- Convolution layer: A layer used to extract features from an image by making use of filters that learn from small squares of input data. It receives the input maize leaf images in the form of pixel values that is convolved with the filter to extract low level characteristics of image like curves and edges and generates a feature map.

- Pooling Layer: It is used to reduce the dimensionality of image which reduces the computational power needed for successive layers. Max pooling is used in the project and it selects the maximum value in a region based on filter size.

- Fully Connected Layer (Dense) : These are part of the final layers of CNN are capable of recognizing features that are highly correlated with output class. The output is a one-dimensional vector obtained by flattening the results of previous pooling layers.

- Dropout layer : Used to reduce overfitting of the model by randomly discarding certain set of neurons in that layer.

- SoftMax layer: final layer in the network that helps in classifying the input images of maize into various classes based on the characteristics learnt by the network

4.5 Transfer learning

It is a niche in the deep learning domain that is gaining prominence off late. It basically works by transferring knowledge acquired from data in one domain into other domains. The main advantages include the lack of need for huge datasets to train the model and less computational power as model weights are already pretrained . They also generalize well on new data as they are built to prevent overfitting .The models used in the current research, namely VGG16 And VGG19 are pre trained on ImageNet database containing 1.6 million images belonging to over 1000 classes. These models are known to provide best results and excellent performance when used for image classification as they have learnt several features from the huge database it had been trained on. Only the final few layers of the model have been retrained using maize dataset to make the model predict better. A brief description of the architectures chosen for transfer learning is given below:

- VGG16 and VGG19 These deep learning architectures make use of only 3x3 convolutional layers that are piled on top of one another. VGG16 entails 13 convolutional layers and makes use of 5 max pooling layers in its architecture. The number 16 and 19 indicate the number of weighted layers in the network. The final layer is a dense layer containing 4096 neurons. A SoftMax layer is used for final classification. A pictorail representation of VGG16 architecture is as shown in Figure 5

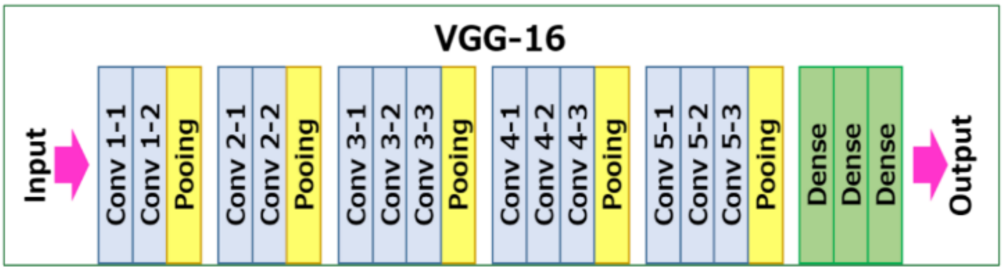


Figure 5: VGG16 architecture

5 Implementation

This section discusses in detail the steps that were performed to implement the various image classifier models for classification of diseases of maize plant using images of its leaves.

5.1 Development Environment

R and python are the two popular languages used in the field of machine learning. The current research project was entirely implemented using python (3.7.4) with Jupyter notebook serving as the Integrated Development Environment (IDE). This was mainly because of the availability of large number of deep learning libraries like keras and tensor flow on python that make implementation simpler and easy. Other necessary packages like numpy, scikit, etc were also installed in due course of implementation using pip command. Matplotlib and other utilities were finally used for analysis of the results.

5.2 Data Handling

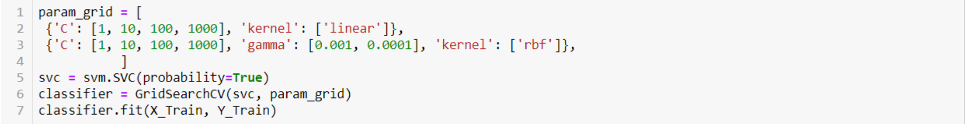
As discussed in section 3, the dataset necessary for developing the project was downloaded from GitHub and preprocessed by removing unwanted folders and retaining information pertaining only to maize plant. They were then transformed into numpy arrays and split into test and train tests before building various models. Even though, several models have been developed in similar areas of research , it must be noted that they are not robust and perform rather poorly on unseen data. This is being addressed by making use of image augmentation functionality of Image generator class of keras library and transfer learning. This would make the models more robust and generalize better on unseen data.

5.3 Model Implementations

All models discussed in section 4 are implemented and the parameters and data split used for each of them is discussed below.

5.3.1 SVM

It is one of the basic machine learning models implemented for classifying maize diseases and makes use of svm module from the library, sklearn. It has several parameters like kernel, gamma, etc. that can be hyper tuned to achieve better results. GridsearchCV from the same sklearn module has been used to find the best values for some of the parameters in order to make the model perform best. These tunings are computationally expensive and take a lot of time and indicated that rbf kernel performed best for the classification with a cost of misclassification (C) value of 1. A snapshot of the gridsearch used is as shown in Figure 6.



```
param_grid = [
 {'C': [1, 10, 100, 1000], 'kernel': ['linear']},
 {'C': [1, 10, 100, 1000], 'gamma': [0.001, 0.0001], 'kernel': ['rbf']},
         ]
svc = svm.SVC(probability=True)
classifier = GridSearchCV(svc, param_grid)
classifier.fit(X_Train, Y_Train)
```

Figure 6: Grid Search parameters for SVM

5.3.2 Random Forest

It is an ensemble model which is implemented using the RandomForestClassifier from the sklearn.ensemble library. Like every other model, it too has several parameters that can be hyper tuned. Although the algorithm runs in a blink when run on default parameters, it was taking a lot of time to run the gridsearchcv task to find the optimum parameters. Hence the test was discarded, and default values were used. The default values included 10 for n_estimators which is the number of trees in the forest , ‘gini’ as criterion for loss calculation and bootstrap value of true indicating that samples were drawn with replacement.

5.3.3 Gradient Boost and Extreme Gradient Boost

These are ensemble models based on decision trees that follow boosting architecture as discussed in section 4. They have been implemented using the GradientBoostingClassifier from sklearn.ensemble library and XGBClassifier from xgboost library respectively. These have been hyper tuned using gridsearch to improve efficiency. Some of the parameters that have been tuned included learning rate and n_estimators which is the number of trees used for ensemble. The tuning proved to be computationally expensive and took a lot of time to complete. It indicated that model performed well with learning rate of 0.1 and n_estimators at 100 for both Gradient and XGBoost.

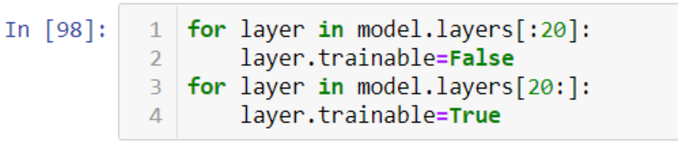
5.3.4 CNN and transfer learning using VGG16 and VGG19

At first, a basic CNN was developed by making using deep learning library called keras which comes prebuilt with all the layers like convolution, max pooling, etc. CNN with multiple layers were developed and checked for accuracy. Having tried various combinations of layers ranging from 3 to 6, five-layered CNN produced the best result when it came to classifying the maize diseases. Dropout layers were also used to avoid overfit and softmax was used as activation function on dense layer at the end. Relu activation was used for hidden layers. Other parameters like epoch, batch size, learning rate and optimizers were also varied in order to test the model. Epochs of 15,32, 64; learning rate of 0.01 and 0.1 and optimizers like adam, nadam and SGD were some of the parameters that were varied during development of CNN.

For developing models based on transfer learning, keras offers several architectures that are pretrained on Image net data and can be downloaded using their respective libraries. The pretrained weights of these architectures are then downloaded onto the machine running the algorithm. VGG16 and VGG19 architectures were used in the current project for analysis . VGG is known to produce accurate results based on research conducted in section 2 and hence has been used for analysis. However, they are huge and occupy larger space in comparison to other architectures (300 Mb).

The models downloaded were then modified by making the top 4 layers trainable and freezing the rest of the layers on image net weights as shown in Figure 7. In place of model’s dense layers, a global average pooling layer with couple of dense layers were used and ‘softmax’ classifier with four classes was used for final classification. Parameters like batch size and epoch along with learning rates were varied for analysis. They were tested for batch size of 32 and 50 with epochs of 15 and 30. Learning rate of 0.001 with higher batch size performed better in the tests.

The models were run using an image size of 224*224 pixels which is the default size for these architectures. Every model implemented was also evaluated for different train test split ratios i.e. 70/30 and 80/20 Validation for deep learning based models were done for every epoch to verify if they were overfitting on the training data.



```
In [98]:  for layer in model.layers[:20]:
              layer.trainable=False
          for layer in model.layers[20:]:
              layer.trainable=True
```

Figure 7: Training layers for transfer learning

6 Evaluation

This section provides a detailed analysis of all the results that are obtained with regards to achieving objectives of the project. The transfer learning models that have been developed were built to demonstrate the feasibility and robustness of such techniques in the field of disease classification . Boosting algorithms like Gradient Boost and XGBoost are relatively new and are yet to be fully explored in this field. They are improved versions of random forest and each tree built tries to reduce the error of previously built trees. The accuracy, recall, precision and f1-score obtained for developed models based on data split ratio of 70:30 is as listed in the Table 1.

Table 1: Results for 70:30 data split

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
VGG19	95%	93%	92%	93%
VGG16	94%	93%	90%	91%
Random Forest	85%	78%	77%	77%
XGBoost	89%	83%	81%	82%
SVM	86%	66%	73%	69%
GradientBoost	87%	81%	78%	80%
5-Layer CNN	79%	84%	80%	72%

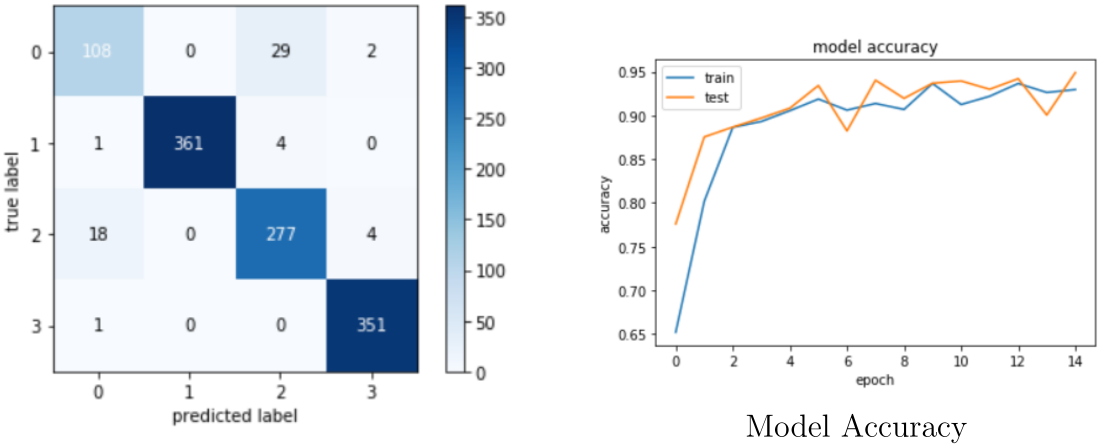
The value of these metrics when analyzed on a train test split of 80:20 is as shown in Table 2.

Table 2: Results for 80:20 data split

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
VGG19	95%	94%	93%	93%
VGG16	94%	93%	90%	91%
Random Forest	83%	74%	74%	74%
XGBoost	90%	85%	82%	83%
SVM	91%	86%	83%	84%
GradientBoost	88%	83%	80%	80%
5-Layer CNN	89%	90%	79%	78%

6.1 Model Comparison

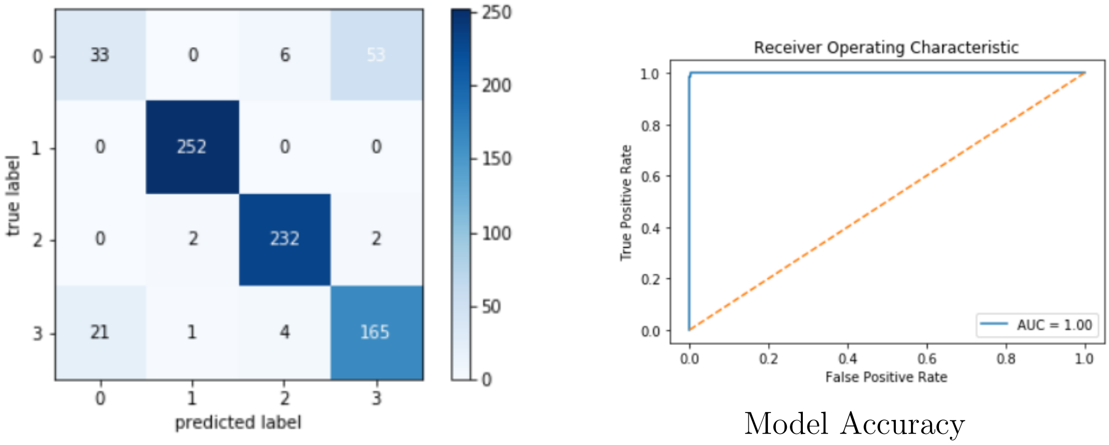
From the values shown in Table 1 and 2, we can conclude that VGG19 transfer learning architecture works best for classifying the maize disease as it has better accuracy, precision



Model Accuracy

Confusion matrix

Figure 8: Confusion Matrix and Model Accuracy for VGG19



Model Accuracy

Confusion matrix

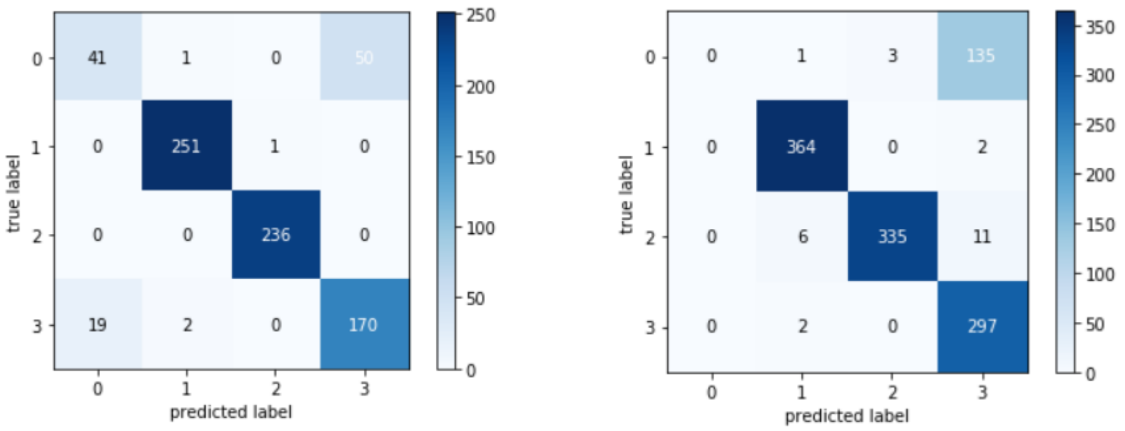
Figure 9: Confusion Matrix and ROC Curve for Gradient Boost

and recall scores in comparison to other models. It scored an accuracy of 95 for both the training splits of 70:30 and 80:20 with VGG16 following closely at second place.This results confirm that transfer learning techniques based on CNN are better suited for identification and classification of maize plant than traditional models thereby answering our research question .With pretrained image net weights and only modifying the final few layers, the models developed can perform better in real world scenario. The training data also contained augmented images which improved the model efficiency and the values of accuracy and other metric values substantiate it. The confusion matrix and model accuracy plot(learning curve) for a 70:30 split for VGG19 which performs best in classifying maize disease is as shown in Figure 8

Similarly, the confusion matrix and receiver operating characteristics for Gradient Boost is as shown in Figure 9 . Their performance was relatively average when compared to transfer learning methods but were on par with other algorithms in the study.

6.2 Analysis of different test train splits

Skimming through values in Table 1 and 2, we can conclude that all models in general perform better with increase in amount of training data. This can further be stressed



Confusion matrix-80:20

Confusion matrix-70:30

Figure 10: Confusion Matrix for 80:20 and 70:30 split for SVM

using classification matrix of models developed. For illustration, lets consider the case of SVM.The confusion matrix of SVM for 80:20 and 70:30 split is shown in Figure 10. One can see that SVM predicts class zero which is gray spot leaf disease with 41 correct guesses on an 80:20 split compared to 0 accurate classifications of this disease on a 70:30 split indicating that more training data improves model performance.

This huge difference is however not observed in case of transfer learning models. VGG19 for example has a precision score of 84 and 92 for detecting gray leaf spot thereby suggesting that transfer learning techniques are not heavily dependent on quantity of training data and generalise well. They were able to learn better from the data provided and are able to differentiate diseases that are remotely similar in terms of appearance and classify them accurately.

6.3 Execution Time

Apart from evaluating the models based on accuracy and other related metrics, training time was also compared to understand how each of these models stack up against each other. As the final layers of the transfer learning based algorithms were retrained, model training time of these were relatively more with VGG19 taking 580 seconds per epoch compared to 780 seconds per epoch of VGG16. This indicates that VGG19 is better, both in terms of training time and accuracy when compared with VGG16. Among other models, random forest executed the fastest with a training time of only 22 seconds with SVM and XGboost taking 18 and 44 minutes respectively. Grid search significantly shot up the training times of these algorithms with SVM running for close to 6 hrs, even for a limited number of parameters chosen. This indicates that GPU enabled cores must be used for such analysis as they can help cut down these training time significantly.

6.4 Discussion

The primary objective of the research was to build models that are efficient in identifying and classifying maize plant diseases. A thorough literature review was done to understand the current limitations and research gaps. These mainly revealed the need for models that could generalise well on unseen data. Transfer learning solutions are tailor made for such issues with models like VGG16 and VGG19 developed to address this problem in the current research. Along with this models like Gradient Boost and XGBoost were developed and compared against conventional models based on SVM,RF using metrics like accuracy,precision, recall and f1-score. Necessary pre-processing which involved resizing of images and converting them to numpy arrays were done before building the models. Augmentation of images and validation for every epoch run was done in the case of deep learning models to ensure that they are robust.Grid research and parametric analysis was also performed to find optimum parameters for developed models.

Results indicated that transfer learning models were better when it comes to predicting almost every class of disease on which it was trained.This was not the case when it came to conventional algorithms like SVM and RF as they performed miserably when it came to diseases that looked similar in terms of symptoms , namely gray leaf spot and northern leaf blight. Models based on Gradient Boost and XGBoost also performed better in this scenario but were not as good as VGG based models.

Although the research was successful in achieving its objectives,it is not short of limitations and faced some roadblocks in due course of its development. One major roadblock was the lack of data captured from different places and backgrounds which could have been used to further test the model and improve on it. Also, the models developed were trained on three types of disease classes. This might result in models under performing when it makes predictions on unseen classes of maize disease. Images with multiple leaves or those affected by multiple diseases can also be a challenging affair and their classifications are also not tested. Apart from this, training the last few layers of transfer learning or running the grid search for developed algorithms were highly time consuming and can be improved by making use of higher configuration machines or by running the models using GPU.The amount of training time taking by these algorithms was one of the reasons for training them on images of reduced dimensions. Better accuracy could probably have been achieved if higher dimension images were used for training the models

7 Conclusion and Future Work

The results obtained, as discussed in previous section highlight and support the use of CNN based transfer learning technique in order to accurately identify and classify maize leaf diseases, thus satisfying the objectives stated for the project. As image augmentation and transfer learning are used, it is a given that these models are capable of generalising well and can predict better on unseen data. VGG 19 performed best among the models developed. However, they took a lot of training time which is concerning. Gradient Boost and XGboost, although not as good as transfer learning based models were also quite effective in detecting and classifying maize plant disease.Its therefore a tradeoff between accuracy and training time and models can be chosen based on priority of requirement.

The research conducted can be extended further by testing other transfer learning based models available. In future, object detection can also be added to make the models more diverse and identify pests or insects that may be affecting the plants. The models developed are however trained on single leaf data and may not perform well on images with multiple leaves which also presents scope for improvement.Finally, the developed models can be deployed on smart phones using smaller architectures of transfer learning like MobileNet or by using cloud based platforms, thus enabling the farmers to detect and identify diseases at the click of a button and monitor them regularly by providing suitable remedies. This would be cost effective and could become a revolutionary change that could go a long way in helping the agricultural community at large.

8 Acknowledgement

I would like to express my heartfelt gratitude to my mentor Dr.Cristina Muntean whose experience, constant guidance and advice has helped me complete the research project successfully. I would also like to take this opportunity and thank the faculty members of NCI who have imparted knowledge and made me more proficient in the field of data analytics.Lastly,i thank my family and friends,for having faith and supporting me throughout my journey, without which none of this was achievable.

References

Aravind, K. R., Raja, P., Mukesh, K. V., Aniirudh, R., Ashiwin, R. and Szczepanski, C. (2018). 'Disease classification in maize crop using bag of features and multi-class support vector machine', *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, pp. 1191–1196, IEEE Xplore Digital Library.doi:10.1109/ICISC.2018.8398993.

Arsenovic, M., Karanovic, M., Sladojevic, S., Anderla, A. and Stefanovic, D. (2019). 'Solving Current Limitations of Deep Learning Based Approaches for Plant Disease Detection', *Symmetry 2019* , 11, pp. 939.

Coulibaly, S., Kamsu-Foguem, B., Kamissoko, D. and Traore, D. (2019). 'Deep neural networks with transfer learning in millet crop images', *Computers in Industry* ,108, pp. 115-120.

Dhaware, C. G. and Wanjale, K. H. (2017). 'A modern approach for plant leaf disease classification which depends on leaf image processing', *2017 International Conference on Computer Communication and Informatics, ICCCI 2017,* 1-4 Jan 2017, pp. 5–8, IEEE Xplore Digital Library.doi:10.1109/ICCCI.2017.8117733.

Ferentinos, K. P. (2018). 'Deep learning models for plant disease detection and diagnosis', *Computers and Electronics in Agriculture* , 145, pp. 311-318.

Francis, J., Anto Sahaya Dhas D and Anoop B K (2016). 'Identification of leaf diseases in pepper plants using soft computing techniques', *2016 Conference on Emerging Devices and Smart Systems (ICEDSS)* . Namakkal, India,4-5 March 2016, pp. 168–173, IEEE. doi: 10.1109/ICEDSS.2016.7587787.

Gao, X., Fan, S., Li, X., Guo, Z., Zhang, H., Peng, Y. and Diao, X. (2017). 'An improved XGBoost based on weighted column subsampling for object classification', *2017 4th International Conference on Systems and Informatics, ICSAI 2017.* Hangzhou, China, 11-13 November 2017, IEEE Xplore Digital Library. doi:10.1109/ICSAI.2017.8248532.

Lu, J., Ehsani, R., Shi, Y., Abdulridha, J., de Castro, A. I. and Xu, Y. (2017). 'Field detection of anthracnose crown rot in strawberry using spectroscopy technology', *Computers and Electronics in Agriculture,* 135, pp. 289-299 .

Maniyath, S. R., Vinod, P. V., Niveditha, M., Pooja, R., Prasad Bhat, N., Shashank, N. and Hebbar, R. (2018). 'Plant disease detection using machine learning', *Proceedings - 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control, ICDI3C 2018.* Bangalore, India, 25-28 April 2018, pp. 41-45, IEEE Xplore Digital Library. doi: 10.1109/ICDI3C.2018.00017.

Mohanty, S. P., Hughes, D. P. and Salathé, M. (2016). 'Using deep learning for image-based plant disease detection', *Frontiers in Plant Science,* 7, pp. 1419 .

Padol, P. B. and Yadav, A. A. (2016). 'SVM classifier based grape leaf disease detection', *Conference on Advances in Signal Processing, CASP 2016.* Pune, India, 9-11 June 2016, pp. 175-179, IEEE Xplore Digital Library. doi: 10.1109/CASP.2016.7746160 .

Shrivastava, V. K., Pradhan, M. K., Minz, S. and Thakur, M. P. (2019). 'Rice plant disease classification using transfer learning of deep convolution neural network', *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives,* XLII-3-W6, pp. 631-635.

Xu, Z. and Wang, Z. (2019). 'A Risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier', *11th International Conference on Advanced Computational Intelligence, ICACI 2019.* Guilin, China, 7-9 June 2019, pp. 278-283, IEEE Xplore Digital Library. doi: 10.1109/ICACI.2019.8778622.

Zhang, J. C., liang Pu, R., hua Wang, J., jiang Huang, W., Yuan, L. and hua Luo, J. (2012). 'Detecting powdery mildew of winter wheat using leaf level hyperspectral measurements', *Computers and Electronics in Agriculture*, 85, pp. 13-23, ScienceDirect. doi: 10.1016/j.compag.2012.03.006 .

Zhang, S. and Wang, Z. (2016). 'Cucumber disease recognition based on Global-Local Singular value decomposition', *Neurocomputing* , 205, pp. 341-348, ScienceDirect. doi: 10.1016/j.neucom.2016.04.034 .

Zheng, Q., Huang, W., Cui, X., Dong, Y., Shi, Y., Ma, H. and Liu, L. (2019). 'Identification of wheat yellow rust using optimal three-band spectral indices in different growth stages', *Sensors (Switzerland)* , 19(1) .