Master Degree Project

UNIVERSITY OF SKÖVDE

DETECTION OF PLANT DISEASES IN TOMATO LEAVES

With focus on providing explainability and evaluating user trust

Nishanth Vijay

Supervisor: Joe Steinhauer
Examiner: Tove Helldin

Master Degree Project in Informatics
with a specialisation in Data Science
Spring term 2021

ABSTRACT

The transmission of diseases from unhealthy to healthy plants is one of the most disastrous threats to the agriculture industry. Diseases transferred spread like wild fire and have the potential to infest the whole farm if not detected early. Plant disease detection methods aid in identifying infected plants in their very early stages and also help the user in scaling the identification of plant diseases to a variety of plants in a cost-effective manner. The aim of this thesis is to implement two different machine learning models, namely, Convolution Neural Networks (CNN) and K-nearest Neighbors (KNN) for the application of plant disease detection in tomato leaves.

The two machine learning models were evaluated on four different metrics in order to find the best performing model among the two. The four different metrics were, Accuracy, Precision, Recall and F1-Score. Other than identifying the diseases using the aforementioned machine learning models, this study also focused on providing explainability to the predictions made by the respective models using the Explainable Artificial Intelligence technique, Local Interpretable Model-agnostic Explanations (LIME). In vein of collecting domain specific expertise, a user study was implemented in which the user trust of the AI and XAI models were evaluated and feedback from farmers were collected in order to provide recommendations for future research.

The results on implementing the machine learning models showed that the CNN model performed better than the KNN model in all of the four evaluation metrics and the results from the user study signify that the farmers do not trust the AI and XAI models, however, the user study through the feedback collected from the farmers helps identify areas in which the trust of the farmers can be grown and strengthened.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

For a long time, the agriculture industry has used modern science to meet the food demands of 7 billion people. However, there are numerous threats that people working in the agriculture industry face that threaten the food security of the human society. Some of the threats as we know include, climate change, livestock grazing, plant diseases, etc, (Food and Agriculture Organization of the United Nations, 2017). Among the many threats, the effect of plant disease is truly momentous as it not only causes huge wastage of plants for human consumption but it also immensely affects the health of the human society and the lives of the farmers whose main source of income is from their production of healthy crops (Al-Sadi, 2017; Somowiyarjo, 2011).

During the process of plant harvesting, human experts go through a tedious process of checking and removing mature plants, making sure they aren't affected by any disease and are suitable for human consumption. However, this traditional visual process of identifying the name of the disease a particular plant is suffering from consumes a lot of time and is expensive especially if the farmhouse is big and there are a lot of plants (Gavhale and Gawande, 2014). Furthermore, with the apparent increase of population in the world day by day it is only practical that this process is automated so that the growing demands of the people can be met.

With the dawn of machine learning models, the early identification of plant diseases has been made much easier, less time consuming and cheaper in comparison to the traditional visual identification of plant diseases. A lot of research has been carried out in this domain in recent years, because of which the industry is slowly moving towards replacing the traditional identification of plant diseases with machine learning models (Ngugi, Abelwahab and Abo-Zahhad, 2021).

The aim of this thesis is to implement two different machine learning models, namely, Convolutional neural network (CNN) and K-nearest Neighbor (KNN) on the plant village dataset and also evaluate the aforementioned models based on the following evaluation metrics: Accuracy, Precision, Recall and F1-Score. The study focuses on the disease identification of tomato leaves from the plant village dataset in specific (J and Gopal, 2019). The novelty of this study lies in the fact that this study also aims at providing transparency and explainability for the decisions made by the aforementioned models using the Explainable Artificial Intelligence (XAI) technique, Local interpretable model-agnostic explanations (LIME). The use of XAI in order to explain the predictions made by the machine learning models is very rare and not to be found in many of the research papers in this particular domain, therefore in that regard the speciality of this paper lies in the fact that this paper aims at not only implementing but also explaining and providing transparency to the users on the predictions made by aforementioned machine learning models.

The reason for choosing these models, in particular, lies in the application and model complexity that exists between these models. KNN models are simple models that are known for their short computational time and easy resultant output interpretation and CNN models, on the other hand, are complex models that have the ability to compress images making them into a form that is easier to process making sure that the features essential for getting a good prediction are not compromised. Additionally, other than the difference in application and model complexity between the models, ongoing through various works of literature in the domain of plant disease detection three reasons can be found as to why the comparison between CNN and KNN in this study is important and interesting,

1. There are exactly two studies in the domain of plant disease detection that provide a comparison between the methods KNN and CNN. The first paper is by, Sharma, Verma and Goel (2020) where a survey is conducted in which different Machine learning algorithms and their application in different types of plants are put forth and in the second paper by, Hatuwal, Shakya and Joshi (2020) an implementation between KNN, CNN, Random Forest and Support Vector Machines is presented. In both of the aforementioned studies, the detection of tomato leaves is not included when comparing CNN and KNN, which this study explicitly does. Not only are there very few papers present in this domain comparing CNN and KNN, but there are none comparing the models on tomato leaves. It is also important to mention that the authors of Hatuwal, Shakya and Joshi (2020) also mention the need for implementing their methods on different plant species in their future works.
2. The target user for this study is small-hold farmers and when considering the use case of small-hold farmers, to them both CNN and KNN models are seen as black-box models as their background in Machine learning is little to none, even though in the data science community KNN's are inherently regarded as white-box models. Providing explainability to both the models and not only the CNN model provides great value to the user by showing which model is better than the other and how the explanations differ between the models. It is also important to mention that none of the papers in this domain provide explanation for the decision made by their models which leaves the small-hold farmers to blindly trust the models. This study by providing explainability to both KNN and CNN helps alleviate that.
3. Often in the data science community the CNN models are said to always perform better than the KNN models, which from various literature can be found to be not true. A paper presented by Toghi and Grover (2018) shows the KNN model scoring accuracy of 97% on the MNIST dataset (MNIST is a dataset containing images of handwritten digits) and a paper presented by Wu (2018) shows the CNN model scoring 94% on the MNIST dataset. The dataset used in both papers are identical and contain the same number of training and testing images. From the abovementioned studies, it can be seen that the KNN model performs better than the CNN model hence stating that we cannot certainly say that a CNN model will always perform better than a KNN model, making the comparison in this specific domain important and interesting.

Another important aspect of this study is the execution of a user study. The user study is used as a means to evaluate if whether the farmers find the implemented AI and XAI models trust worthy, comprehensible and as a tool that would improve their plant disease diagnosis process. The user study is also used to receive feedback from the farmers on areas that can be improved and mention potential shortcomings. The feedback from the farmers is used to find areas in which the trust of the farmers can be grown and strengthened. Since one of the future objectives of this study is to implement the plant disease detection system as a mobile application that farmers or any other enthusiast interested in botany can use, feedback from domain specific professionals help the study greatly in identifying the flaws and improving them. Due to the newness of a user study in this domain and in order to not overwhelm participants participating in the user study, for time being this study focuses predominantly on the user trust of the farmers.

CHAPTER 2
BACKGROUND

In this chapter we go over few important concepts and techniques required to understand the problem area.

Before explaining the methods used in the study in detail it is important to provide some context to the chosen models. Convolution Neural Networks and K-nearest Neighbors belong to the same class of machine learning models, namely, supervised machine learning models where the models are trained on labelled datasets and the performance of the models is tested on unseen datasets. CNN and KNN models can be used for both classification and regression tasks. The study of plant disease detection is a Multi-class image classification problem wherein given an image of a diseased plant the classifier has to predict the disease that the plant carries from more than two different class labels. Some of the most commonly used methods in a supervised classification task are Decision Trees, Naive Bayes Classifiers, Support Vector Machines, K-nearest Neighbors and Deep Neural Networks. The selection of a machine learning model solely depends on the nature of the task, however, in recent years, the use of deep learning is becoming widespread across different domains such as facial recognition, object detection, speech recognition and natural language processing (Iqbal and Yan, 2015).

2.1 Artificial Neural Networks

Artificial Neural Networks, abbreviated as ANNs can be described as computational systems that are designed to replicate the human brain's analysis and information processing skills, and just like the human brain an ANN consists of a directed graph with interconnected processing elements known as neurons (Jain, Mao and Mohiuddin, 1996).

There are different types of neural networks such as the Recurrent Neural Network (RNN), Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), etc. Among the many Neural Networks present the most common one is the Multilayer Perceptron network (MLP). A typical MLP network comprises of different node layers, namely, an input layer, one or more hidden layers and finally an output layer. All of the nodes are connected to each other and each node is associated with a weight and threshold. The weight refers to the importance that two nodes connecting have in a network and data is transferred from one layer to another only if the output of an individual node is above the threshold specified (Jain, Mao and Mohiuddin, 1996).

An ANN's training process involves recognizing patterns in the data that is fed into the network, during this supervised learning phase the actual and the desired output of the ANN are compared (Cost function) and the training process is presented a number of times (epochs) until there is little or no difference between the actual and desired output. The Cost function is minimized using a process known as backpropagation, wherein the network works backwards from the output layer to the input layer adjusting the weights of its nodes until the error produced between the actual and desired output is the lowest possible (Jain, Mao and Mohiuddin, 1996).

2.2 Convolutional Neural Networks

Regular Neural Networks such as the Multi-layer Perceptron (MLP), in the past were used for image classification purposes however as the resolution of the images being used to classify became higher and higher the networks became computationally hard to deal with and the number of total parameters used for classification would be far too many.

Convolutional Neural Networks are very similar in working to regular neural networks such as the MLP, however, what changes in Convolutional Neural Networks is that the layers of a CNN have three-dimensional arrangement (width, height and depth) of neurons instead of the standard two-dimensional array and for this simple reason CNNs are widely used on image data for the purpose of classification as the architecture of a CNN is designed to take advantage of the 3d form of an image.

A simple Convolutional Neural Network's architecture consists of three main layers, namely, Convolutional layer, Pooling layer and the Fully connected layer.

The Convolutional layer is regarded as the main building block of a CNN, it consists of learnable parameters known as filters/kernels. The filter is responsible for finding patterns (textures, edges, shapes, objects, etc) in the input image. Each filter slides/convolves over the height and width of the input image, computing the dot product between the filter and the pixels present in the input image. The resultant of a Convolutional layer is a feature map that summarizes all the features found in the input image (Yamashita *et al.*, 2018).

The Pooling layer is another building block of a CNN that is used to perform downsampling in order to reduce the spatial size of the feature map. This is done to reduce the number of parameters, computations in the network and also to make sure overfitting is controlled. There are two different types of pooling in CNN namely, Max Pooling and Average Pooling. Max Pooling returns the maximum value that is present in the portion of the image convolved by the kernel and Average Pooling returns the average of all the values present in the portion of the image convolved by the kernel (Yamashita *et al.*, 2018).

This layer is often followed only after non-linearity is introduced into the network using a Rectified Linear Unit (ReLU) activation function, this layer is used to turn all the negative values present in the feature map into zeros. The addition of ReLU is very important in a CNN as it helps increase the non-linearity in the images. Images as such are known to contain non-linear features such as different objects or the boundaries between those objects, however, when imposing an image through a Convolutional layer to create a feature map, there might be some linearity introduced in the image and in order to bring back the non-linearity in the image the ReLU activation function is used (Zhao *et al.*, 2018).

After adding multiple Convolutional layers and Pooling layers, the fully connected layers are introduced in the architecture. The input coming to the fully connected layer is the flattened output from the final convolutional and pooling layer. The flattening is done to convert the 3d matrix data from the last pooling layer into a 1d array of vectors so that it can be used by the fully connected layers which perform the same operations to an ANN to carry out the final classification and compute the class scores (Yamashita *et al.*, 2018). And in this way, a simple Convolutional Neural network is structured to read and classify images.

2.3 K-nearest Neighbors

The K-nearest Neighbor is a machine learning algorithm that estimates how likely a new sample belongs to a label, based on the voting of the majority labels that the data points nearer to the new sample are found in. The algorithm works by calculating the distance between the new sample and all the other samples after which it sorts the values calculated in ascending order. Based on the 'k' value the majority label is chosen as the prediction. Here 'k' signifies the number of nearest neighbors to be included in voting of the majority label (Guo *et al.*, 2003).

The choice of 'k' in KNN makes or breaks the predictions in the algorithm. If the value of 'k' is too small then the classification is very vulnerable to noise and the model might be overfitted. On the other hand, if the 'k' value is too big then the model most likely will classify any new sample as the majority label all the time (Paryudi, 2019).

KNN is regarded as one of the most simple and intuitive algorithms for classification and it belongs to the class of lazy learner algorithms that do not perform any generalization on the training data until a query on the dataset is made to the system (Guo *et al.*, 2003).

2.4 Local Interpretable Model-agnostic Explanations (LIME)

Most of the machine learning models being used to solve real-world problems nowadays are referred to as black box models. The term black-box models refer to a system that can be viewed only in terms of its input and output, the internal workings of a black box system is unknown and uncontrollable by the user or even the designer. With the growing advancements in machine learning, it is becoming increasingly hard to understand why certain machine learning models are doing what they are doing (Loyola-González, 2019).

The concept of Explainable Artificial Intelligence (XAI) aims at solving the issues of a black-box model by bringing in transparency and explainability into the decisions made by a machine learning model. This helps in the user's experience of using a machine learning model greatly as the users now know and can explain an AI's decision which in turn helps in understanding their model better and provides the user with the freedom to improve the performance of their model or to remove any bias. The use of XAI in machine learning helps in garnering the user's trust and confidence which in most use cases is of top priority (Tjoa and Guan, 2020).

Local Interpretable Model-agnostic Explanations (LIME), is an XAI technique that is used to explain black-box classifiers. LIME is suitable to use on various types of data such as tabular, text and images. The full form of LIME itself paints a good picture as to what LIME means and that is that LIME is (Molnar, 2020),

1. Local, meaning the model is able to provide explanations that really reflect the behaviour of the classifier for a given observation's prediction
2. Interpretable, meaning the model is able to clearly convey its explanations in a manner that makes sense to the user
3. Model-agnostic, meaning the model does make any assumptions about the classifier that it is interpreting and that it can be applied to any black-box model. LIME is model-independent

The working of LIME is explained in the following four steps (Molnar, 2020),

1. Data Permutation – The first step that LIME carries out is that it creates several data points that are closer to the data point that we want to return explanations for. So, for example, in the case of an image, LIME will create several copies of the original image with slight alterations by turning the superpixels of the original image on and off, these types of images are also known as perturbated images.
2. Prediction of perturbated images – In this next step LIME will use our trained black-box model to return predictions for each of the artificially created images/perturbated images.
3. Assigning weights to perturbated images – In this step the LIME model assigns weight to each artificially created image by measuring the distance between the artificially created image and the original image. Closer the distance larger the weight that is assigned which in turn tells the model how important that particular perturbated image is.
4. Finding the important features – The last step in LIME is to find the important features for which a linear regression model is fitted using the weighted perturbated images and just like in traditional linear regression models the coefficients of each feature is returned. The coefficients are then sorted and based on the sorted coefficients, the features in which the largest coefficients are present are deemed as the features that play the most important role in the prediction of our trained black-box model.

There are various types of XAI techniques available in Machine learning such as Anchors, LIME, SHAP (Shapley Additive explanations) and CEM (Contrastive explanations method). From the many XAI techniques that are available the most popular XAI technique used in explaining machine learning models is, LIME.

To provide a short overview of the different XAI techniques (Linardatos, Papastefanopoulos and Kotsiantis, 2021),

1) SHAP is a method inspired from game theory that provides interpretability by calculating the importance of each feature for every individual prediction, in short SHAP when providing an explanation for an instance it considers all possible predictions using various possible combination of inputs (also known as Shapley values). It is used to provide global as well as local explanations, that is, it provides explanations that summarize the entire dataset as well as individual observations.

2) CEM is another XAI technique that is used to generate contrastive explanations for any machine learning model. The speciality of CEM lies in the fact that for a specific prediction to be produced CEM has the ability to not only identify features that must be present but also identify the features that must be absent for that prediction to be produced.

3) Anchor is another model-agnostic approach and it works under a set of if-then rules known as anchors. Essentially, in this method for a given instance's prediction an anchor is defined that presents the local explanation and any changes made to the other features present in the instance does not disturb the prediction value.

In comparison to the many XAI techniques mentioned above, some of the benefits of LIME is that,

1. LIME is faster as it only has to look in its local vicinity to provide explanations
2. It is known to increase the overall interpretability (Dieber and Kirrane, 2020)
3. It is also easier to implement and it is popular (Linardatos, Papastefanopoulos and Kotsiantis, 2021)
4. And it can be applied to any black-box model (Kinkead *et al.*, 2021)

other than the above stated reasons, a research conducted by Malhi *et al.* (2019) on the classification of In-Vivo Gastral images used LIME to provide explanations in order to gain the trust of the health professionals and used CNN for classification of the images. The study states that on providing the explanations returned by LIME to a partner hospital, LIME received positive feedbacks.

To the best of our knowledge the use of LIME in the domain of plant disease detection has never been done before. Due to its infancy in this domain and positive feedbacks received in the paper proposed by Malhi *et al.* (2019), the study would like to execute LIME in the domain of plant disease detection and also collect feedback from farmers on if they trust LIME and detect possible shortcomings in order to improve future iterations of the study.

2.5 Related Works

A research on the recognition of diseases in Paddy leaves using the KNN classifier conducted by (Suresha, Shreekanth and Thirumalesh, 2017) on a database containing 330 images of paddy leaves showed that the proposed work produced a test accuracy of 76.59%. The study used accuracy as its only metric for evaluating the KNN classifier, and not metrics such as precision, recall or f1-score which in this paper is going to be a focus point.

Madhulatha and Ramadevi (2020), conducted a study on the recognition of plant diseases in which a deep Convolution Neural Network model was used and the proposed work was shown to produce an accuracy of 96.50%. The study makes use of the famous AlexNet architecture in order to classify the different plant diseases. The AlexNet architecture is a Neural Network that consists of eight layers of learnable features that is famously used in most image classification use case scenarios. The dataset used in this study makes use of all the images from the plant village dataset which consists of 54,323 images of plant diseases and 38 different disease categories.

Hatuwal, Shakya and Joshi (2020), performed a study on different machine learning models such as, Support Vector Machine (SVM), K-nearest Neighbor (KNN), Random Forest Classifier (RFC) and Convolution Neural Network (CNN) on the detection of plant diseases. Among all the machine learning models the CNN model scored the highest accuracy of 97.89% followed by the RFC which scored 87.436%, SVM which scored 78.61% and finally KNN which scored 76.969%. Unlike the previous studies this paper did make use of precision, recall and f1-score to evaluate its models, however, in the final comparison of the models, the accuracy of all the models were only taken into consideration in choosing the best performing model.

A study conducted by Agarwal *et al.* (2020) on the recognition of diseases in tomato leaves using Convolutional Neural Network showed that the proposed CNN model scored an accuracy of 91.2% in comparison to pre-trained CNN models such as, VGG16 which scored 77.2%, Mobilenet which scored 63.75% and finally the Inception model which scored 63.4%. The proposed CNN model in this study consists of three Convolution layers and three Max pooling layers. This study by (Agarwal *et al.*, 2020) also sheds light on the benefits of not using a pre-trained model in which it finds that the proposed model needed very less storage space of 1.5 MB in comparison to the pre-trained models which needed 100 MB.

CHAPTER 3
PROBLEM SPECIFICATION

3.1 Aim

The aim of this thesis is to,

1. Perform image classification using two different machine learning models
2. Evaluate the performance of the respective machine learning models
3. Provide explainability to the predictions made by the respective machine learning models
4. Evaluate user trust of both the implemented AI and XAI models using a user study and collect feedback from farmers in order to provide recommendations for future research

3.2 Motivation

Plant diseases possess a very devastating threat to the agriculture industry and have the potential of pushing the whole of human society into starvation if not detected early. With the implementation of machine learning models in the domain of plant pathology, the detection of plant diseases will become easier and cheaper helping many farmers in the timely detection of plant diseases, preventing wastage of plants and protecting the transmission of diseases from diseased to healthy plants.

A lot of the research that has been carried out on plant disease detection present a comparative study using different machine learning models but fail to explain the predictions made by their models. In this research we not only provide a comparative study between the workings of a simple and complex model, but we also aim at providing explainability for the predictions made by the models.

The motivation for including explainability lies in the fact that most of the machine learning models widely used in this domain are black-box models, which leads to the users using these models for prediction in not trusting and understanding how their models make their predictions. The application of Explainable Artificial Intelligence techniques demystifies these black-box models and allows the users to understand their model predictions better and make decisions on their own as to whether, trust their model or not. The use of XAI helps a great deal in the application of plant disease detection as the transparency and explainability of the models being used is vital in gaining the trust of the workers working in the agriculture industry as their livelihood is dependent on their production of healthy plants.

3.3 Research Question

The research questions that this paper aim to answer are,

1. Which among the two models namely, CNN and KNN, will perform the best on evaluating the models on metrics, Accuracy, Precision, Recall and F1-Score on the chosen dataset?
2. Do the farmers find the explanations provided by LIME for the two AI model's predictions easy to trust?

3.4 Objectives

1. Research the related problem area
2. Implement the pre-processing on the dataset
3. Implement CNN and KNN
4. Train and test CNN
5. Train and test KNN
6. Evaluate the performance of both the models
7. Apply XAI technique on both the models in order to find the features responsible for the prediction
8. Design and execute a user study

3.5 Hypothesis

The study hypothesises that the CNN model may perform better than the KNN model as CNNs are known to perform well with a large amount of data and on image classification use cases. With regards to the user study, the study hypothesises that the farmers may trust the implemented AI and XAI model's predictions and explanations due to the positive feedbacks it received in Malhi *et al.* (2019) and in comparison to the two models, it is also hypothesised that the farmers may trust KNN with LIME more than CNN with LIME due to its simplicity and the distrust the CNN with LIME due to its complexity.

3.6 Use Case

Information technology systems such as mobile phones can be used to take the pictures of the leaves and used to detect and display the disease of the plant through the seamless user interface provided by a mobile application. Applications such as these help the target users save time and also enable users in testing more of their plants at a fast phase. The implementation of such an application is not included in this study and is reserved for future works.

3.7 Target User

In general, this application can be used by anyone who is involved in the agriculture industry growing any type of plants. However, this study considers its primary target users as small-hold farmers who have no education on how to identify plant diseases. This application will be utmost useful to the small-scale farmers who have no knowledge regarding the differences between different plant diseases as this study not only predicts the plant diseases but also explains the features that were responsible for the prediction made, which puts users with zero knowledge regarding plant diseases in great comfort in knowing and trusting the predictions. The reason for choosing LIME is because it helps increase model interpretability which is vital as this study is directed towards users who have absolutely no knowledge on how to detect early stages of plant diseases and no background in machine learning. An XAI technique that is easy to interpret helps the users understand the explanations with no help. A study conducted by Dieber and Kirrane (2020) where the understandability of the output produced by LIME for tabular data was tested on people who had no prior knowledge in LIME and LIME showed to increase the overall interpretability of the model.

CHAPTER 4
METHODS AND IMPLEMENTATION

4.1 Selection of Methods and Metrics

Plant disease detection can be performed using different classifiers and a multitude of techniques have been used in the past for this purpose. In this thesis, the classifiers that were used for performing the detection were the Convolution Neural Network (CNN) and K-nearest Neighbor (KNN). The XAI technique used for providing explainability for the predictions made by the classifiers was Local Interpretable Model-agnostic Explanations (LIME). The evaluation of the aforementioned models was done using the following metrics: accuracy, precision, recall and f1-score. Each classifier was evaluated using the same four evaluation metrics and the results of both the classifiers were used to find the best performing model on the disease detection of tomato leaves from the plant village dataset.

4.2 Dataset

The dataset chosen for this paper contains 10,000 images of tomato leaves with 10 different categories. Each category contains the same number of images. The images of plant diseases are of the size 256x256 and the dataset does not contain any missing images. The dataset is commonly referred to as the plant village dataset and can be found on the public website, Mendeley data [1].

Baterial Spot	Early Blight	Healthy	Late Blight	Leaf Mold
				
				
Septorial Leaf Spot	Spider Mites Two Spotted Spider Mite	Target Spot	Mosaic Virus	Yellow Leaf Curl Virus

Fig 1 - 10 different types of tomato leaf diseases

[1] Link for the dataset - https://data.mendeley.com/datasets/tywbtsjrjv/1

4.3 Libraries Used

Table 1 - Libraries used in the implementation of CNN and KNN

NumPy	Is a library in python programming language that helps in managing large, multidimensional arrays and matrices. This library also enables the operation of mathematical functions on the created arrays or matrices
Cv2	Is an open source computer vision and machine learning library that is used to solve computer vision problems such as reading or resizing an image
Sklearn	Is a open source machine learning library that includes various classification and regression algorithms such as, K-nearest Neighbors, Support vector machines, random forests, etc. This machine learning library was introduced by
Keras	Is an open source neural network library that is designed to enable the fast implementation of deep neural networks with in the python interface itself
Matplotlib	Is a plotting library used in python for creating static and animated visualizaitions
Lime	Is an open source library in python programming language that is used to implement the LIME model that explains black box classifiers

4.4 Pre-processing of the Dataset

The pre-processing step for any machine learning model is of great importance and ideally shapes the performance and results of the models chosen. In this report, the following were the steps that were carried out in order to make sure that the models produced optimal results.

1. Read images – cv2.imread() was used to read the images from the root directory that was specified
2. Resize images – cv2.resize() was used in order the change the dimensions of the images. The images were 256x256 in size initially and were resized to 150x150 due to the inability of the system's (Random Access

Memory) RAM to carry out operations on all 10,000 images of size 256x256 using both CNN and KNN model.

3. After reading and resizing the images, we then convert the images into an array form using np.array()
4. The labels of each plant images are then also mapped to a unique value using LabelBinarizer()
5. Finally, the plant village dataset is split into two different sets, namely, train and test set with a 75:25 ratio respectively

4.5 Configuration of the Classification Models

4.5.1 Convolution Neural Networks

The architecture of CNN used for plant disease detection in this paper was as follows, the first block contains a Convolutional layer with 32 filters of size 3 x 3 and the activation function used was the ReLU activation function. We then follow the operation by performing batch normalization, and choosing the Max Pooling layer with a pool size of (3,3) and adding a dropout layer with 25% dropout.

Batch normalization was performed in order to speed up the convergence of the neural network, it is generally applied after each individual layer so that the output of the previous layer can be normalized allowing for each individual layer present in the network to perform learning independently (Garbin, Zhu and Marques, 2020). Dropout layer is a technique used to prevent the model from overfitting by randomly switching off some sections of the neurons. When some sections of the neurons are switched off the incoming as well as the outgoing connections from the neurons are also switched off and this results in the betterment of the model in learning and allows for the model to not generalize to the test dataset (Garbin, Zhu and Marques, 2020).

The second block in the network contains two convolutional layers with 64 filters of size 3 x 3 with ReLU activation function and batch normalization. After which, Max Pooling layer with pool size (2,2) and a dropout layer with 25% dropout was added.

The third block in the network contains two convolutional layers with 128 filters of size 3 x 3 with ReLU activation function and batch normalization. After which similar to the second block, a Max Pooling layer with pool size (2,2) and dropout layer with 25% dropout was added.

To complete the model we now move on to constructing our Fully connected layer but before that, a flattening operation was performed to convert the 3D output of the last convolution layer into 1D form before feeding it into the Fully connected layers for classification.

For the classification of the features extracted from the last convolution and pooling layer, a dense layer was added with 1024 neurons with ReLU activation function, batch normalization and dropout layer of 5% dropout to improve the results of the model. The Final layer in the neural network was the logits layer which will return the values of the predictions made by the model and for this a dense layer with 10 neurons (as there are 10 classes) with Softmax activation function was used. The Softmax activation was used in this layer as we are using a multi-label classification model (Yamashita *et al.*, 2018).

Table 2 - Parameters used in the CNN model

Parameter	**Setting**	**Description**
Epochs	25	Defines the number of iterations that the whole training dataset goes through the network
Batch size	256	Refers to the number of training data that is used during each epoch before weights in the network are updated
Learning rate	1e-3 (0.001)	This parameter determines how much the weights have to change based on the error observed
Optimizer	Adam	Optimization functions are used in machine learning to reduce the cost/loss function and by reducing the cost/loss function we can in turn achieve minimal differences between the predicted and actual output. Adam is one such optimization technique used instead of the traditional stochastic gradient descent

The architecture and the parameters of the CNN discussed above in Table 2 were chosen after performing many experimentations on the chosen dataset and the settings that produced the best results were chosen at last. The reason the learning rate was set to 0.001 is that for complex problems such as image classification smaller learning rates are often good at producing higher generalization accuracies even though they produce larger training times it is worth the wait as they are known to improve the overall accuracy of the model (Wilson and Martinez, 2001).

The batch size of the model was found by experimenting different batch sizes on the model. In the first experiment, the batch size was set to 32 and model produced an accuracy of 94.9%. In the second experiment, the batch size was set to 128 and the model produced an accuracy of 93.1% and in the third time of the experiment the batch size was set to 256 and the model produced an accuracy of 98.5%. A study by Rolnick *et al.* (2018) investigates the effect of batch size on the robustness of noise in Neural Networks during training. The study runs on a 2-layer ConNet on the MNIST dataset and the experiment is implemented on varying batch sizes from 32 to 256. The study found that increasing the batch size provided great robustness to the noise in the dataset.

The Adam optimizers are one of the most popular optimizers used in Neural Networks and they are widely known for producing great results when used for image classification purposes (Ault, 2020).

4.5.2 K-nearest Neighbors

The value of 'k' in KNN determines the performance of the model and as mentioned earlier if a small value of 'k' is chosen then we increase the chances of the model overfitting and if a large value of 'k' is chosen then we increase the chances of the model classifying the majority class all the time. All numbers between 1 to 25 were tested to find the best value of 'k'. Two graphs were plotted using the help of Matplotlib Library. The first graph shows the relation between the 'k' value and the accuracy produced by the model. The second graph shows the relation between the 'k' value and the error produced by the model.

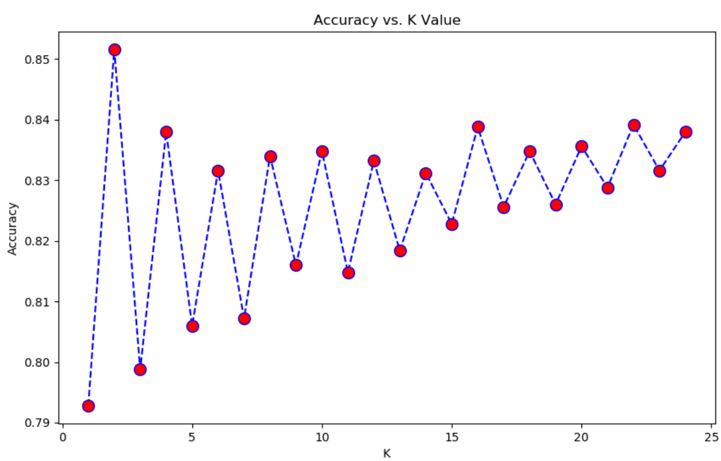


Fig 2 - Accuracy plotted against different K values

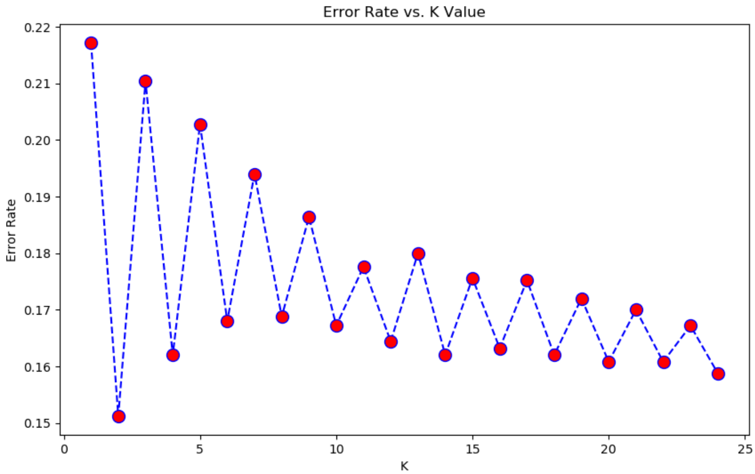


Fig 3 - Error rate plotted against different K values

The best value of 'k' is chosen from observing both the above graphs and a 'k' value that has high accuracy and low error rate is used for the experiment. From Fig 2, we can see that at k = 21 the accuracy of the model reaches 83.6% and the error produced from Fig 3 can also be found to be the least with a value of 0.163. Hence, for this study the value of 'k' was chosen as 21 and the distance was calculated using Euclidean distance.

Table 3 – Parameters used in the KNN model

Parameter	**Setting**	**Description**
K	21	Number of nearest neighbors to be included in the voting of the majority class

4.6 Configuration of the XAI Model

4.6.1 Local Interpretable Model-agnostic Explanations

We make use of the lime image package to implement our LIME model and at first, we create an object 'explainer'. This object makes use of the method explain_instance() that takes in 3D image data and the model's predictor function (model.predict). Based on the prediction from the trained model the explanations are returned.

As mentioned earlier, once the images are perturbated, the trained black-box model is used by LIME to predict each one of the artificially created images and weights are assigned based on the proximity of perturbated images to the image of interest. After assigning weights to each one of the perturbated images a linear classifier is used to find the most important features in the image.

In order to visualize the explanations. The method 'get_image_and_mask()' was used. This method takes in the 3D image data passed earlier and returns (image, mask) tuple. Where 'image' refers to the 3D numpy array representing the image data and 'mask' refers to the 2D numpy array representing the features in the image responsible for the prediction made by the model.

The first 5 parameters that LIME require can be seen in Table 4,

Table 4 – Parameters used for providing explainability in LIME

Parameter	**Setting**	**Description**
Image data	images	This parameter holds the image that you want to return explanations for. The name of the setting depends on the directory in which the image is present in
Predictor function	model.predict	This refers to the trained model's predictor function that was implemented when the model was created
Top labels	top_labels = 3	The trained model predicts the top 'n' labels that the image might belong to. The labels are arranged from the highest probability to the lowest probability of belonging to a label.
Number of samples	num_samples = 1000	Refers to the number of artificial data points to be created to fit the linear model

Here we take a look at the parameters needed for visualizing the explanations in Table 5,

Table 5 – Parameters used for visualizing the explanations in LIME

Parameter	**Setting**	**Description**
Label	explanation.top_labels[0]	The top most label (which contains the highest probability) is returned
Positive only	positive_only = True (mask 1) positive_only = False (mask 2)	If true, positive only returns the superpixels that contribute towards the prediction of the label If false, returns both the positive and negative superpixels that contribute towards the prediction of the label
Number of features	num_features = 5 (mask 1) num features = 10 (mask 2)	Number of superpixels to be used in the explanations
Hide rest	hide_rest = True (mask 1) hide_rest = False (mask 2)	If true, Turns the non-explanation sections of the image into black If false, doesn't do anything to the image

The output from LIME contains two images 'mask 1' and 'mask 2' as seen above. Mask 1 returns the features that were responsible for the prediction made by the model and Mask 2 returns the features that increase and decrease the probability of that image belonging to the predicted label.

4.7 User Study

A survey style user study was incorporated in order to enquire and collect feedback from professionals in the domain (farmers) on the trustworthiness of the implemented AI and XAI model's predictions and explanations. The survey contains questionnaires and short text fields. The questionnaire was used to verify if whether the users find the predictions of each of the AI models and explanations provided by the XAI model to be trustworthy, consistent and comprehensible and the short text fields were used in order to collect additional feedback from the users. In order to measure the agreement of each question in the questionnaire with the user the Likert scale approach was adopted and this approach was adopted due to its popularity and value that it brings to a user study in quantifying a particular user's opinion (Bishop and Herron, 2015). The Likert scale used in this study had the following pointers to collect responses from the users,

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

The primary objective of this user study is to evaluate user's trust for the implemented AI and XAI models and since there is no well-established methodology to evaluate XAI systems the metrics used in this study were adopted from Khodabandehloo *et al.* (2021). The adopted evaluation metrics were,

- Human-machine task performance (HMTP), which measures if whether the use of the tool makes the end-users more successful in their task
- Explanation satisfaction (ES), which measures the end-user's satisfaction and understandability of the machine's explanations
- User trust and reliance (UTR), which measures the end-user's trust and reliance in the machine's explanations

Using the above-mentioned metrics and Likert scale approach the user study was designed. The user study was split into three main sections, the first section contained the AI model's predictions without any explanations from the XAI model and the second section contained the AI model's predictions with explanations from the XAI model and the third section contained the results, that is the actual disease labelled for each prediction.

The participants were shown a small sample size of predictions from the AI and XAI models and were asked whether or not if they trust and understand the predictions and explanations. Questions from the survey are available down below.

4.7.1 Survey Questions

For AI model's prediction without any explanations the questions presented were as follows,

1. I trust that the AI model is offering me the correct predictions (UTR)
2. The predictions presented by the AI model are comprehensible (ES)
3. The identification of diseases in tomato leaves would be easy with this AI model (HMTP)

The users are then shown the AI model's prediction with explanations from the XAI model (LIME) and asked the following questions,

1. I trust that the AI model is offering me the correct explanations (UTR)
2. The explanations presented by the AI model are comprehensible (ES)
3. The identification of diseases in tomato leaves would be easy with this AI model (HMTP)

The users are then shown the results (actual disease labelled) and asked the following question,

1. On viewing the results my trust in the AI models has increased (UTR)
2. If you have any additional feedback, please provide them below

Since farmers are the domain experts it was ideal for farmers to be chosen as participants for the user study as they have domain specific experience in diagnosing plant diseases and would be able to provide in depth feedback on the right and wrongs in plant disease detection. Additionally, the farmers participating in the survey were also asked to mention their years of experience in a small text field. This data would help the research in distinguishing what a farmer who has been working for less than 2 years in the domain thinks about the detection tool and what a farmer who has been in the domain for more than 2 years thinks of the detection tool. The participants were shown five predictions from each AI model and asked the above stated questions and their responses were registered.

The user study was presented through a google form and published on the Facebook group, ‘World Agriculture Group for farmers’ for the farmers in the group to participate. The user study was chosen as a survey for this very reason as surveys can be easily distributed through online forums without the need for the participants to be present at one place at one time. Surveys encourage mass participation and do not demand a lot of time from the participants unlike interviews. An introductory google document was also given to the participants and it covered the following topics, participant’s objective for the user study, purpose of the research, purpose of the user study and the user study design, participation criteria, ethical aspects, contact information and links to the survey.

The above mentioned google document also contained a brief introduction to the AI models, XAI model and the evaluation metrics used in the user study for the participants to read if they were interested. The submitted google document can be found under Appendix A.

CHAPTER 5
RESULTS

5.1 Results from the Classification models

Convolution Neural Network

Table 6 - CNN evaluation metrics

Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
98.5	93	93	93

K-nearest Neighbor

Table 7 - KNN evaluation metrics

Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
83.6	90	84	86

The following results present the Accuracy, Precision, Recall and F1-Score of both the CNN and KNN model. In order to understand the concept of the above-mentioned evaluation metrics, it is important to understand what True positives (TP), True negatives (TN), False positives (FP) and False negatives (FN) are.

For example for an image of a tomato leaf that contains the disease 'Early blight', the confusion matrix looks like as shown in Fig 4. The confusion matrix shown below is for one of the 10 categories/diseases.

		Actual disease	
		Early blight	Not Early blight
Predicted Disease	Early blight	TP	FP
	Not Early blight	FN	TN

Fig 4 - Confusion matrix for 'Early blight'

If the model correctly predicts the image of the plant as containing the disease then the outcome is known as a True positive (TP) outcome. If the model correctly predicts the image of the plant as not containing the disease then the outcome is known as True negative (TN) outcome.

If the model incorrectly predicts the image of the plant as containing the disease then the outcome is known as a False positive (FP) outcome. If the model incorrectly predicts the image of the plant as not containing the disease then the outcome is known as a False negative (FN) outcome.

5.1.1 Accuracy

This metric is calculated by dividing the total number of correct predictions which is the True positive and True negative outcomes, with the overall total number of samples (Liu *et al.*, 2014). As we can observe from Table 6 and Table 7, the CNN model scored a high accuracy of 98.5% in comparison to the KNN model which scored 83.6%.

In this report since the dataset chosen at hand contains labels that contain the same number of samples, accuracy proves as a useful measure in comparing both models. However, in scenarios where the dataset is imbalanced accuracy cannot be proven useful due to the paradoxical finding known as the Accuracy paradox.

5.1.2 Precision

This metric is calculated by dividing all the correctly predicted predictions (True positives) with the total positive predictions (True positives and False positives) (M and M.N, 2015). This metric highlights the exactness of a classifier and answers the question, out of all the samples that the classifier predicted as being 'Early blight', how many of them were actually 'Early blight'. As we can observe from Table 6 and Table 7, the CNN model scores a precision score of 93% and the KNN model scores 90%. This proves that among the two classifiers, the CNN model will significantly return more relevant results than irrelevant ones in comparison with the KNN model.

5.1.3 Recall

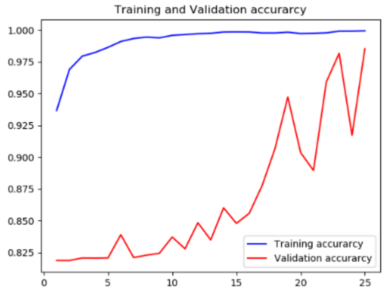
This metric is calculated by dividing all the correct predictions (True positives) with the total number of True positives plus the False negatives (M and M.N, 2015). This metric is used to highlight the sensitivity of the model, meaning among the samples labelled as 'Early blight' in the dataset, how many did the classifier recognise as 'Early blight' in its predictions. As we can observe from Table 6 and Table 7, the CNN model scores a recall of 93% and the KNN model scores 84%, proving that the CNN model is able to return most of the relevant results in the plant village dataset in comparison to the KNN model.

5.1.4 F1-Score

Since in both precision and recall the CNN model has produced better scores than the KNN model, comparing the F1-Score of the models is hardly necessary as F1-Score is a harmonic mean of both precision and recall (Liu *et al.*, 2014) and is used to serve as alternative metric that helps in deciding if a model that produces high precision and low recall is more suitable than a model that produces low precision and high recall. However, from Table 6 and Table 7 we can see that the CNN model even in the F1-score metric performs better than the KNN model, producing an F1-Score of 93% and the KNN model producing a score of 86%.

The training vs validation accuracy and the training vs validation loss were also plotted for the CNN model in which we can observe from Fig 5 and Fig 6 that as the training accuracy increases the validation accuracy increases and as the training loss decreases the validation loss also decreases signifying that the model is neither underfitting nor overfitting and that the model is constantly learning.

Underfitting in machine learning models occurs when a model performs poorly on the training dataset and performs poorly on the validation/testing dataset as well. Overfitting in machine learning models occurs when the model performs really good on the training dataset but performs poorly on the validation/testing dataset. A good fit model is one that learns the training dataset and is able to suitably generalize the validation dataset well (Zhang, Zhang and Jiang, 2019).



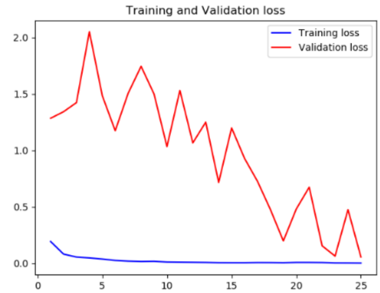


Fig 5 – Training Vs Validation Accuracy **Fig 6** – Training Vs Validation Loss

In conclusion, to answer our first research question, from the above evaluation metrics we can observe that the results are as hypothesized and CNN does in fact produce better results than the KNN model, which makes the CNN model a better suit for application in plant disease detection.

The explanations as to why CNN performed better than the KNN model can be found from the core architecture that the CNN model is equipped with, namely, the convolutional layers, pooling layers and fully connected layers. In a CNN the image goes through many learnable layers in order to understand the different aspects of the image such as edges, shapes, objects, etc. Meanwhile, the KNN model only looks at the neighbors present near the image that is to be classified in order to predict its label. The model complexity that lies behind both these models is one of the reasons why the CNN model outperformed the KNN model.

However, when comparing the training time of both the models, the KNN model easily wins over the CNN model. The KNN model took approximately an hour to finish its training and the CNN model took approximately 4 hrs to finish its training. If the KNN model had produced similar or closer results to the CNN model, in that case, considering the training time the KNN model would have been the best pick.

5.2 Results from the XAI model

Fig 7 - Original image of a tomato leaf with 'Late blight' disease

The above image Fig 7 is an example of what is fed into LIME for it to provide explanations on the prediction made by the CNN and KNN model.

Implementation of LIME on CNN

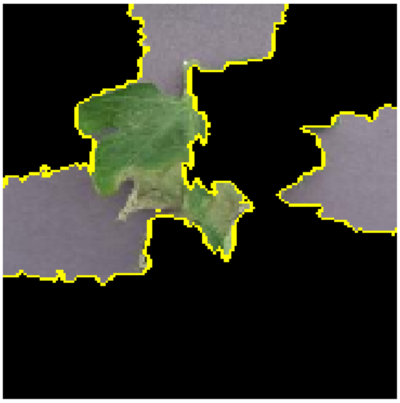


Fig 8 – Features produced by LIME **Fig 9** – Probabilities produced by LIME

The output produced by LIME in order to explain the CNN model's predictions can be seen above. The image to the left (Fig 8) showcases the features that are responsible for the CNN model predicting Fig 7 as 'Late blight' and the image to the right (Fig 9) showcases the parts that increase the probability (green) of the image belonging to 'Late blight' and parts that decrease the probability (red) of the image belonging to 'Late blight'.

Implementation of LIME of KNN

Fig 10 - Features produced by LIME **Fig 11** - Probabilities produced by LIME

The same image Fig 7 is fed into LIME in order to explain the KNN model's predictions and just as seen for in the CNN model, the left image (Fig 10) showcases the features responsible for the KNN model's prediction and the image to the right (Fig 11) shows the parts that increase the probability (green) of the image belonging to 'Late blight' and parts that decrease the probability (red) of the image belonging to 'Late blight'.

Now in vein of answering our second research question, the results from the user study are presented in the next section.

5.3 User Study

The user study received a total of 7 responses from farmers present in the "World Agriculture group for farmers" Facebook group and the responses for each question are visualised using the plot_likert package and shown below. The work experience of all the 7 farmers were between 5-10 years, which makes their feedback valuable and important.

For CNN's and KNN's predictions without explanation the user study results were as follows,

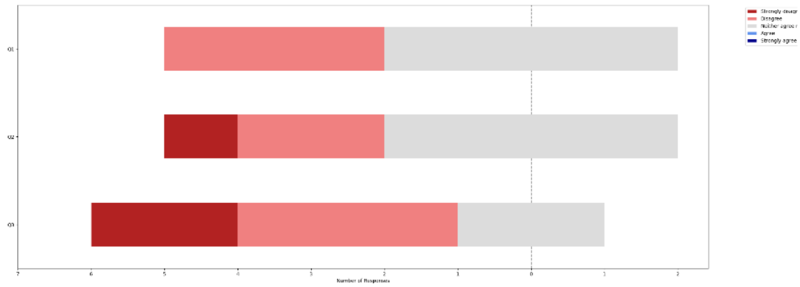


Fig 12 - Convolutional Neural Networks (without explanations)

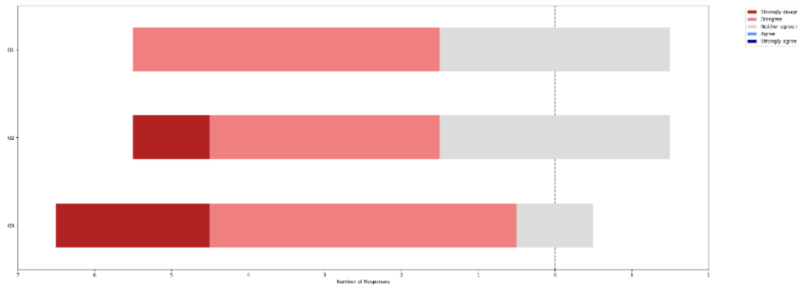


Fig 13 - K-nearest Neighbors (without explanations)

For CNN's and KNN's predictions with explanations from LIME the user study results were as follows,

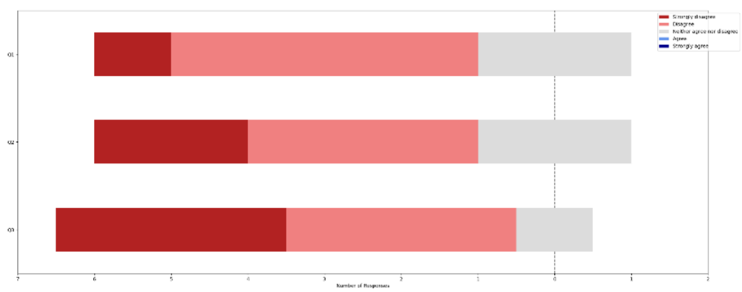


Fig 14 - Convolutional Neural Networks (with explanations from LIME)

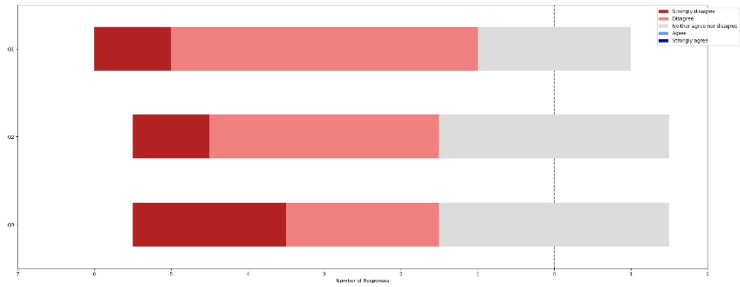


Fig 15 - K-nearest Neighbors (with explanations from LIME)

When the users where shown the results (actual disease labelled) for each prediction and asked if their trust had increased in the AI models their responses were as follows,

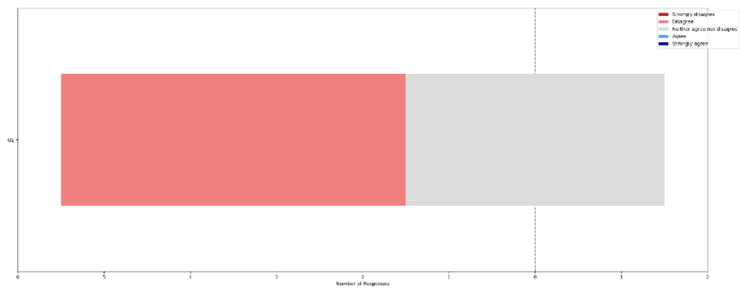


Fig 16 - CNN and LIME (results)

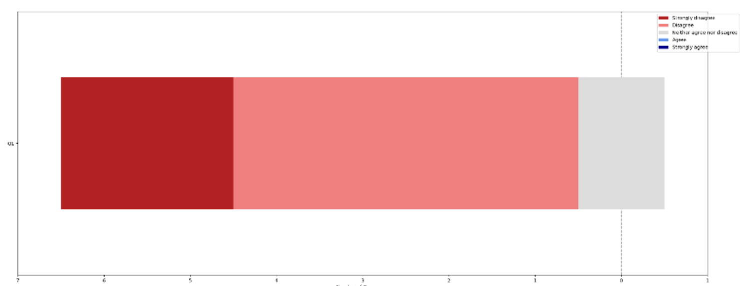


Fig 17 - KNN and LIME (results)

Based on the above responses each metric (UTR, ES and HMTP) are calculated by assigning individual numbers ranging from 1 to 5 to each Likert scale pointer (Chakrabartty, 2014). For example, 1 is assigned as strongly disagree, 2 is assigned as disagree and so on until 5 for strongly agree. The scores once computed are divided by the number of responses in order to make the comparison and evaluation easier. The below table shows the scores computed for each metric.

Table 8 – Score computed for each evaluation metric

Metric	CNN (without explanation)	KNN (without explanation)	CNN (with explanation)	KNN (with explanation)	CNN and LIME (results)	KNN and LIME (results)
UTR	2.57	2.42	2.14	2.14	**2.42**	1.85
ES	2.42	2.28	2	**2.28**	-	-
HMTP	2	1.85	1.71	**2.14**	-	-

Additional feedback received from farmers are listed below,

<u>Convolutional Neural Network</u>

1. *"The use of VOC in this study is absent making it hard for me to trust the legitimacy of the detection"*
2. *"Detection of plant disease depends on time of the month, soil type and environmental conditions on which data is missing"*
3. *"Explanations by LIME are a bit random and unclear"*
4. *"Correctly predicts 3 out of 5 diseases but is not good at explaining how"*

<u>K-nearest Neighbors</u>

1. *"Visual explanations from LIME cover almost the whole image of the leaf when the fungal infections can be found only in certain areas"*
2. *"Tough to rely when it gets only 2 images right"*
3. *"Produces convincing symmetrical explanations that are unclear why"*

Comparing CNN with LIME and KNN with LIME, on observing the user trust reliance score from Table 8, we can see that upon showing the results to the farmers that they trust the CNN with LIME model more than the KNN with LIME model as the CNN with LIME model's score is higher. With regards to explanation satisfaction and human-machine task performance the KNN with LIME model is what the farmers seem to prefer from looking at Table 8.

However, based on the feedback received and looking at the Likert scale visualisations, it can be observed that there aren't any farmers who've agreed or strongly agreed with the any of the UTR questions asked. Especially based on the feedback received it is clear that the LIME's explanations for both the models is not adequate enough to gain the trust of the farmers completely. Therefore, to answer our second research question, No, the farmers do not find the explanations provided by LIME for the respective AI model's predictions easy to trust. The reasons for the above results from the user study are discussed in the next section in three parts as user study limitations, XAI model and analysis of results from the user study.

CHAPTER 6
DISCUSSION

In this chapter we discuss the limitations of the user study, discuss problems found in the XAI model, analyse results from the user study, discuss ethical aspects of the research and finally, include a summary of the entire thesis.

6.1 User Study Limitations

A major limitation of the user study was inadequate amount of participants. Only 7 farmers were able to participate in the survey and with such a low number of participants it becomes hard to tell if the results from the user study are because of less explainability and interpretability from the XAI models or rather due to the participant's collective fear that their jobs might someday be taken by Artificially intelligent robots.

Another limitation of the user study was the lack of data to support and make the farmers trust in the AI and XAI models. As mentioned by one of the farmers during the process of conducting the user study and also in the user study feedback, the current generation of farmers use something known as Volatile organic compounds (VOC) in order to detect the disease of plants. Plants as they 'breathe' naturally release VOCs but when they are diseased the concentration and type of VOC changes which is what is used to determine the disease of a plant (Jansen *et al.*, 2011). Hence, data on the VOC concentration and type of VOC would have helped the farmers in trusting the models more as it is data that they use and are familiar with. Data on soil type, environmental conditions and time of the month are also important in such classifications in providing confidence to the farmers as certain plant diseases can be found only at certain time of the month and only on certain soil types, it is widely known in the plant pathology community that environmental factors such as increase or decrease in temperature, humidity or moisture contents are known to cause diseases in plants (Agrios, 2005).

6.2 XAI Model (LIME)

As discussed earlier in the study, we known that LIME creates simulated data around the original prediction through the process known as perturbation in its process to generate explanations for a single prediction. The simulated data created through the process of perturbation is done randomly, which means that every time LIME is executed the explanations returned for a single prediction will be different. This instability in LIME was found while testing the XAI model numerous times trying to return explanations for a single prediction and can also be seen in LIME's explanations for CNN's predictions in the survey (attached in Appendix B). In user trust where stability is of utmost importance, LIME's instability is a critical issue that makes its explanations not trust worth to the users.

A research conducted by Zafar and N. Khan (2019) also mentions this instability in LIME and proposes an alternative method known as Deterministic Local Interpretable Model-Agnostic Explanations (DLIME) that uses hierarchical clusteing to group the training data and makes use of KNN instead of a linear classifier to return explanations. The study makes use of the health care dataset from the UCI repository. The results from the study signify that even though DLIME is successful in producing stable explanations, it's stability depends on the number of samples present in the dataset. The study further adds that in the future it looks forward to solving this issue while also maintaining the stability of model explanations and experiment with other data types such as image data and text data.

Example of LIME's instability in generated explanations for CNN's prediction,

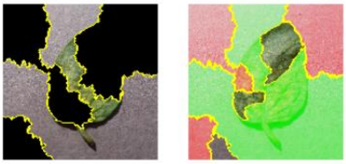


Fig 18 - First explanation by LIME

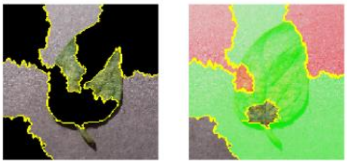


Fig 19 - Second explanation by LIME

On executing LIME twice the above two images are the explanations returned by LIME for a single prediction made by CNN. The plant in the image contains the, Two-spotted spider mite disease.

Through the execution of LIME in this study, we've found that the LIME's instability in generated explanations not only occurs in numeric data as mentioned by Zafar and N. Khan (2019) but also in image data. In future iterations of this study it would be interesting to solve LIME's instability and conduct the user study with the farmers once again.

6.3 Analysis of results from the user study

Since the results from the survey are a reflection of farmer's opinion there is no one single explanation as to why the farmers trusted or understood one model more than the other. However, few plausible explanations are,

1. Model complexity – the farmers might have preferred the KNN model in terms of ES and HMTP just because it was easier to understand while the CNN model was a little complex to understand
2. General aversion towards AI - as mentioned earlier, there is a possibility that the farmers who participated carried a severe dislike towards AI and hence trusted the AI and XAI models less
3. LIME's instability

4. Incomplete data to support – missing data on VOC, soil type, environmental conditions and time of the month as mentioned by the farmers in the user study feedback

In future iterations of the study it is believed that, improvements in the above following areas would certainly improve the user trust of the farmers.

6.4 AI Model (KNN and CNN)

On viewing the results from the user study one of the questions that rises is, how would have the results from the user study changed if the hyperparameter settings of the implemented models were differnet?

In this study due to it's time constraint and the ongoing pandemic the machine learning models implemented were optimized to the best of our abilities, For example in the case of the KNN model which scored 83.6%, the value of 'K' was chosen by testing the model on all numbers between 1 and 25, and the 'K' that was finally selected produced the highest accuracy and lowest error rate. Similarly, in the case of the CNN model which scored 98.5%, the hyperparameters of the model were experimented multiple times until a stable model that was neither underfitting nor overfitting was produced.

There is a chance that the results from the user study would have produced alternate results given the machine learning models were implemented differently or different machine learning models were chosen. However, it cannot be said with certainty that tuning the hyperparameters of the model or use of a different classification model, would produce same or alternate results in the user study without executing and reconducting the user study, as the results are a reflection of the participant's opinions and depend on multiple factors as observed before.

6.5 Ethical Aspects

According to (Wohlin *et al.*, 2012) we address the four main ethical principles in line with the work done in this thesis,

1. Informed consent – this principle states that subjects involved in the study must be given access to all the information regarding the study before they make their decision to participate or not participate in the experimentation. The participants involved with the user study were briefed on the security of their identity and given all information regarding the purpose of the study, participants objective and how their feedback from the user study will be used and how long will it be used for. Only the participants who consented to the briefing attended the user study and provided feedback.

2. Scientific value – this principle states that the study must contain enough scientific value for it to motivate participants to expose themselves to the risks of the study. A lot of research has been conducted in the domain of plant disease detection where multiple machine learning models are implemented however none of them provide explainability and transparency to these otherwise known black-box models. This research provides scientific value by introducing the XAI technique, LIME which allows for the users to understand their model's predictions. Also, through the help of the participant's feedbacks in the study future machine learning enthusiasts when working on plant disease detection systems now know what to look for in order to earn the trust of farmers and improve their detection systems.
3. Confidentiality – this principle states that researchers must take all the precautions needed in order to protect the data and make sure sensitive information are not shared. The dataset used in this paper is open-source, which means that it is open to use and experiment by anyone interested in the data science community and the participants data used for the user study are stored in a secure server and erased upon submitting the final version of the thesis.
4. Beneficence – this principle states that the benefits of the study must overweigh the risks associated with the study. The risks of not detecting a plant disease sooner are that it carries the possibility to cause global starvation, cause a dent in the agriculture industry and affect the health of the whole human society. By making use of machine learning models in this study we are able to detect plant diseases sooner than they appear at a large scale and furthermore. With the help of LIME, the study adds explainability to the machine learning models and by conducting a user study on the implemented AI and XAI models the study was able to collect domain specific feedback from the farmers. In this study, the benefits certainly overweigh the risks associated with the study.

Societal Impacts

The societal impacts of this research can be split into two separate categories, positive and negative. The positive impacts of this research are that this research allows small-hold farmers to be independent and not be dependent on expert knowledge which in turn allows for cheaper, large-scale identification of plant diseases and also helps the small-scale farmers make profits out of their production of healthy. Making the identification of plant diseases accessible to anyone and everyone helps meet the food demands of the human society and this directly helps society as a whole to escape starvation and health problems.

The negative impacts of this research are the effects it has on different members of the society such as the jobs of the experts who traditionally used to identify plant diseases visually or the organizations that employ people to detect plant diseases. A possible solution to this particular negative impact is incorporating the knowledge of the experts in improving the predictions of the model as newer diseases emerge. In this way both the job security of the experts is restored and at the same time the models are also constantly being improved. It is important to mention that the societal impacts of this research have not been tested and the above stated societal impacts are plausible impacts.

Security

In accordance to GDPR, the study was conducted in a transparent and confidential manner. The anonymity of the participants was protected (confidential) and the participants were briefed on how their information will be used, how long will it be used for and how it will be stored (transparency).

The data from the user study were stored in a secure server and the participants were promised that upon submission of the thesis that their data would be erased. The identity of the farmers was not revealed at any point in the study.

6.6 Summary

The aim of this paper was to implement two different machine learning models for the detection of plant diseases and provide explainability for the predictions made by each machine learning model.

The models that were chosen for this study were the, Convolution Neural Network and K-nearest Neighbor. In order to evaluate the models, metrics such as Accuracy, Precision, Recall and F1-score were used in this study. The explanablity of models was done using the Explainable Artificial Intelligence technique, Local Interpretable Model-agnostic Explanations. Both the models were tested on the same dataset called the plant village dataset.

On comparing both the models using the performance metrics, the results show that the CNN model performs better than the KNN model in the application of plant disease detection. The CNN model scored an accuracy of 98.5%, precision of 93%, recall of 93% and f1-score of 93%, while the KNN model managed to score an accuracy of 83.6%, precision of 90%, recall of 84% and f1-score of 86%.

Explanations were generated for both CNN and KNN model's predictions using LIME and a user study was conducted to evaluate if the farmers trust and understand the implemented model's predictions and explanations. Results from the user study higlight the areas that farmers look for when diagnosing plant diseases and the results also indicate that the predictions and explanations from AI and XAI models are not adequate enough to trust the models for the purpose of detecting plant diseases. The feedback from the farmers help the study in identifying areas that might help boost trust in farmers and in turn help improve future iterations of the study.

In short the key contributions of this thesis were,

1. Implementation of LIME to generate explanations for plant diseases
2. Execution of a user study to evaluate user trust of farmers
3. Identification of areas that could possibly help user trust grow in farmers
4. Identification of LIME's instability for image data

CHAPTER 7
CONCLUSIONS

On implementing two machine learning models, Convolutional Neural Networks (CNN) and K-nearest Neighbors (KNN) on the disease detection of tomato leaves from the plant village dataset and also evaluating the aforementioned model using the following metrics: Accuracy, Precision, Recall and F1-Score, the study shows that CNN model performs better than the KNN model in the plant disease detection of tomato leaves by outperforming the KNN model is all of the four evaluation metrics. The study also makes use of the XAI technique Local Interpretable Model-agnostic Explanations (LIME) in order to provide explainability to the predictions made by the models. With the execution of a user study, this study is able to get feedback from farmers on if they trust the aforementioned AI and XAI models. The results from the user study indicate that the farmers find the predictions and explanations from AI and XAI models inadequate and therefore do not trust the implemented tools for the detection of plant diseases. However, through additional feeback the farmers highlight areas that could possibly help improve and trust the AI and XAI models.

7.1 Future Works

The dataset used in this study makes use of only tomato leaves from the plant village dataset. Due to the lack of enough Random Access Memory (RAM) storage the study had to be limited to only 10,000 images. In the future, it would be great to test the implementation of both the CNN and KNN model and also use LIME on the whole plant village dataset containing multiple different plants in order to bring detection and explainability to a wide variety of plants.

Another work that this study would like to pursue in the future is to provide a comparative study on different XAI techniques and implement a user study in order to find out which XAI technique provides the best explainability, transparency and interpretability.

With the addition of data on Volatile organic compounds, soil types, environmental conditions and time of the month as mentioned by farmers through feedback from the user study, the user trust of the detection tool is expected to grow a little higher. As discussed earlier in the use case of this study, a working application that is capable of taking pictures of plants and detecting plant diseases in real-time is the ideal goal and will prove to be of great use to the farmers and botany enthusiasts.

REFERENCES

Agarwal, M. *et al.* (2020) 'ToLeD: Tomato Leaf Disease Detection using Convolution Neural Network', *Procedia Computer Science*, 167, pp. 293–301. doi: 10.1016/j.procs.2020.03.225.

Agrios, G. N. (2005) 'chapter ten - ENVIRONMENTAL FACTORS THAT CAUSE PLANT DISEASES', in Agrios, G. N. (ed.) *Plant Pathology (Fifth Edition)*. San Diego: Academic Press, pp. 357–384. doi: 10.1016/B978-0-08-047378-9.50016-6.

Al-Sadi, A. (2017) 'Impact of Plant Diseases on Human Health', *International Journal of Nutrition, Pharmacology, Neurological Diseases*, 7, pp. 21–22. doi: 10.4103/ijnpnd.ijnpnd_24_17.

Ault, R. (2020) 'Optimization Study of an Image Classification Deep Neural Network', *Final Report*, p. 10.

Bishop, P. and Herron, R. (2015) 'Use and Misuse of the Likert Item Responses and Other Ordinal Measures', *International Journal of Exercise Science*, 8, p. Article 10.

Chakrabartty, S. N. (2014) 'Scoring and Analysis of Likert Scale: Few Approaches', *Jr. of Knowledge Management & Information Technology,* 1.

Dieber, J. and Kirrane, S. (2020) 'Why model why? Assessing the strengths and limitations of LIME.', *CoRR*.

Food and Agriculture Organization of the United Nations (ed.) (2017) *The future of food and agriculture: trends and challenges*. Rome: Food and Agriculture Organization of the United Nations.

Garbin, C., Zhu, X. and Marques, O. (2020) 'Dropout vs. batch normalization: an empirical study of their impact to deep learning', *Multimedia Tools and Applications*, 79(19), pp. 12777–12815. doi: 10.1007/s11042-019-08453-9.

Gavhale, M. and Gawande, U. (2014) 'An Overview of the Research on Plant Leaves Disease detection using Image Processing Techniques', *IOSR Journal of Computer Engineering*, 16, pp. 10–16. doi: 10.9790/0661-16151016.

Guo, G. *et al.* (2003) 'KNN Model-Based Approach in Classification', in Meersman, R., Tari, Z., and Schmidt, D. C. (eds) *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 986–996. doi: 10.1007/978-3-540-39964-3_62.

Hatuwal, B. K., Shakya, A. and Joshi, B. (2020) 'Plant Leaf Disease Recognition Using Random Forest, KNN, SVM and CNN', *POLIBITS*, 62, p. 7.

Iqbal, M. and Yan, Z. (2015) 'SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY', *International Journal of Soft Computing*, 5, pp. 946–952. doi: 10.21917/ijsc.2015.0133.

J, A. P. and Gopal, G. (2019) 'Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network', 1. doi: 10.17632/tywbtsjrjv.1.

Jain, A. K., Mao, J. and Mohiuddin, K. M. (1996) 'Artificial neural networks: a tutorial', *Computer*, 29(3), pp. 31–44. doi: 10.1109/2.485891.

Jansen, R. M. C. *et al.* (2011) 'Detection of Diseased Plants by Analysis of Volatile Organic Compound Emission', *Annual Review of Phytopathology*, 49(1), pp. 157–174. doi: 10.1146/annurev-phyto-072910-095227.

Khodabandehloo, E., Riboni, D. and Alimohammadi, A. (2021) 'HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline', *Future Generation Computer Systems*, 116, pp. 168–189. doi: 10.1016/j.future.2020.10.030.

Kinkead, M. *et al.* (2021) 'Towards Explainable CNNs for Android Malware Detection', *Procedia Computer Science*, 184, pp. 959–965. doi: 10.1016/j.procs.2021.03.118.

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2021) 'Explainable AI: A Review of Machine Learning Interpretability Methods', *Entropy*. doi: 10.3390/e23010018.

Liu, Y. *et al.* (2014) 'A Strategy on Selecting Performance Metrics for Classifier Evaluation', *International Journal of Mobile Computing and Multimedia Communications*, 6, pp. 20–35. doi: 10.4018/IJMCMC.2014100102.

Loyola-González, O. (2019) 'Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View', *IEEE Access*, 7, pp. 154096–154113. doi: 10.1109/ACCESS.2019.2949286.

M, H. and M.N, S. (2015) 'A Review on Evaluation Metrics for Data Classification Evaluations', *International Journal of Data Mining & Knowledge Management Process*, 5(2), pp. 01–11. doi: 10.5121/ijdkp.2015.5201.

Madhulatha, G. and Ramadevi, O. (2020) 'Recognition of Plant Diseases using Convolutional Neural Network', in *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 738–743. doi: 10.1109/I-SMAC49090.2020.9243422.

Malhi, A. *et al.* (2019) 'Explaining Machine Learning-Based Classifications of In-Vivo Gastral Images', in *2019 Digital Image Computing: Techniques and Applications (DICTA). 2019 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–7. doi: 10.1109/DICTA47822.2019.8945986.

Molnar, C. (2020) *Interpretable Machine Learning*. Lulu.com.

Ngugi, L. C., Abelwahab, M. and Abo-Zahhad, M. (2021) ‘Recent advances in image processing techniques for automated leaf pest and disease recognition – A review’, *Information Processing in Agriculture*, 8(1), pp. 27–51. doi: 10.1016/j.inpa.2020.04.004.

Paryudi, I. (2019) ‘What Affects K Value Selection In K-Nearest Neighbor’, *International Journal of Scientific & Technology Research*. Available at: /paper/What-Affects-K-Value-Selection-In-K-Nearest-Paryudi/34c2817df22843d5dc1e4047086617d2d28dd7b6 (Accessed: 6 June 2021).

Rolnick, D. *et al.* (2018) ‘Deep Learning is Robust to Massive Label Noise’. Available at: https://openreview.net/forum?id=B1p461b0W (Accessed: 6 June 2021).

Sharma, V., Verma, A. and Goel, N. (2020) ‘Classification Techniques for Plant Disease Detection’, *International Journal of Recent Technology and Engineering*, 8(6), pp. 5423–5430. doi: 10.35940/ijrte.F9902.038620.

Somowiyarjo, S. (2011) ‘Plant disease problems on smallholder farms in Asia’, *Australasian Plant Pathology*, 40(4), pp. 318–319. doi: 10.1007/s13313-011-0075-5.

Suresha, M., Shreekanth, K. N. and Thirumalesh, B. V. (2017) ‘Recognition of diseases in paddy leaves using knn classifier’, in *2017 2nd International Conference for Convergence in Technology (I2CT). 2017 2nd International Conference for Convergence in Technology (I2CT)*, pp. 663–666. doi: 10.1109/I2CT.2017.8226213.

Tjoa, E. and Guan, C. (2020) ‘A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI’, *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21. doi: 10.1109/TNNLS.2020.3027314.

Toghi, B. and Grover, D. (2018) *MNIST Dataset Classification Utilizing k-NN Classifier with Modified Sliding Window Metric*.

Wilson, D. R. and Martinez, T. R. (2001) ‘The need for small learning rates on large problems’, in *IJCNN’01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222). International Joint Conference on Neural Networks (IJCNN’01)*, Washington, DC, USA: IEEE, pp. 115–119. doi: 10.1109/IJCNN.2001.939002.

Wohlin, C. *et al.* (2012) *Experimentation in Software Engineering*. Berlin Heidelberg: Springer-Verlag. doi: 10.1007/978-3-642-29044-2.

Wu, H. (2018) ‘CNN-Based Recognition of Handwritten Digits in MNIST Database’, p. 8.

Yamashita, R. *et al*. (2018) ‘Convolutional neural networks: an overview and application in radiology’, *Insights into Imaging*, 9(4), pp. 611–629. doi: 10.1007/s13244-018-0639-9.

Zafar, M. R. and Khan, N. (2019) *DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems*. ResearchGate.

Zhang, H., Zhang, L. and Jiang, Y. (2019) ‘Overfitting and Underfitting Analysis for Deep Learning Based End-to-end Communication Systems’, in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP). 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6. doi: 10.1109/WCSP.2019.8927876.

Zhao, G. *et al*. (2018) ‘Rethinking ReLU to Train Better CNNs’, *2018 24th International Conference on Pattern Recognition (ICPR)*. doi: 10.1109/ICPR.2018.8545612.

Appendix A

The document sent to the participants contained the following information

In recent years AI tools have been of great use in providing assistance in various domains such as, automobile industry, health industry and the finance industry (Suguna, 2021). In this study an AI model is used in order to assist the farmers in identifying the name of plant diseases in tomato leaves. The chosen AI models were trained using the images of tomato leaves from the plant village dataset. The dataset chosen can be viewed using the following link - https://data.mendeley.com/datasets/tywbtsjrjv/1.
The user study provided contains a small sample size from the dataset where the chosen AI model's and XAI model's predictions and explanations for each of the tomato leaves is presented. The user study is split into three separate parts, part one, in which the AI model's predictions without explanations are presented and part two, in which the AI model's predictions with explanations are presented and finally, part three in which the results (actual disease labelled) for each prediction are presented.

Participants objective
As an expert in this domain the overall objective for you is to provide feedback as to whether the predictions made by each of the two AI models (CNN and KNN) and the explanations provided by the XAI model (LIME) for those models are consistent, trustworthy and comprehensible.

Purpose of the research
In short, the purpose of this study is to evaluate two different AI models and find which among the two perform the best in plant disease detection and also explain the predictions of each of the AI models using an XAI model. The results from this study will be used to automate the process of plant disease detection and make the detection tool accessible to everyone by implementing it as a mobile application.

User study purpose and design
The survey is used as a measure to verify if the implemented AI and XAI models are interpretable and explainable. The survey will approximately take no more than 5 mins to fill out and is designed using the Likert scale approach.

Participation criteria
The participants of this survey are strictly expected to be farmers and also carry knowledge regarding plant pathology

Ethical aspects
The user study abides by the GDPR regulations and the participant's identities are not revealed at any point in the study. The responses from the participants are stored securely in a google server and upon the completion of the final submission of the research paper the responses from the participants are erased.

Contact information
If you have any clarifications regarding the survey or would like to be updated on the future developments of this research, please contact me through my facebook ID.

Links to the survey
Model 1 - https://forms.gle/KUEwBVoAePR24zMAA
Model 2 - https://forms.gle/bUHhJHxAzvpBGJXE6

Appendix B

Design of the survey

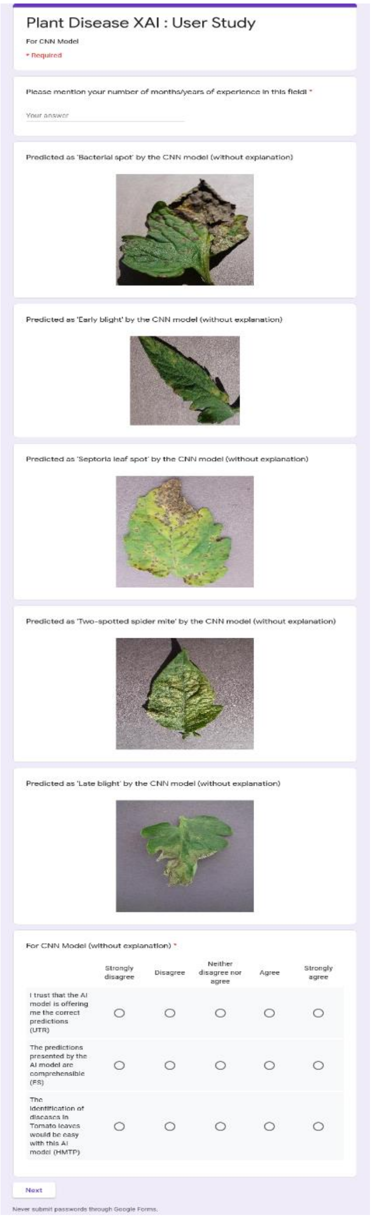


Fig 20 - CNN's predictions (without explanation)

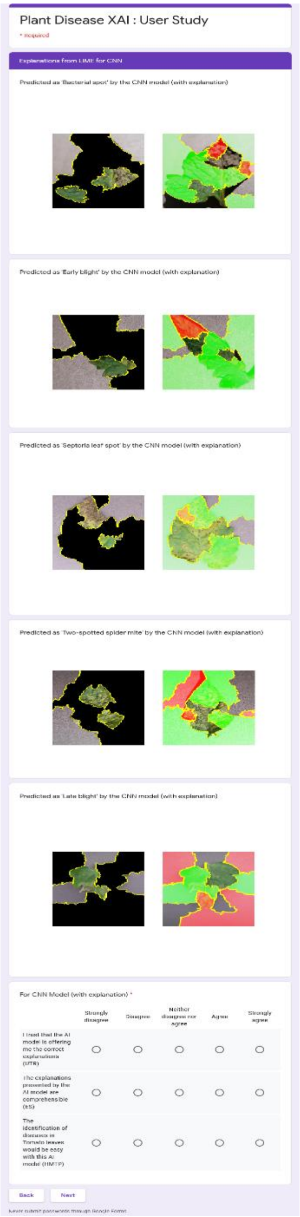


Fig 21 - CNN's predictions (with explanations)

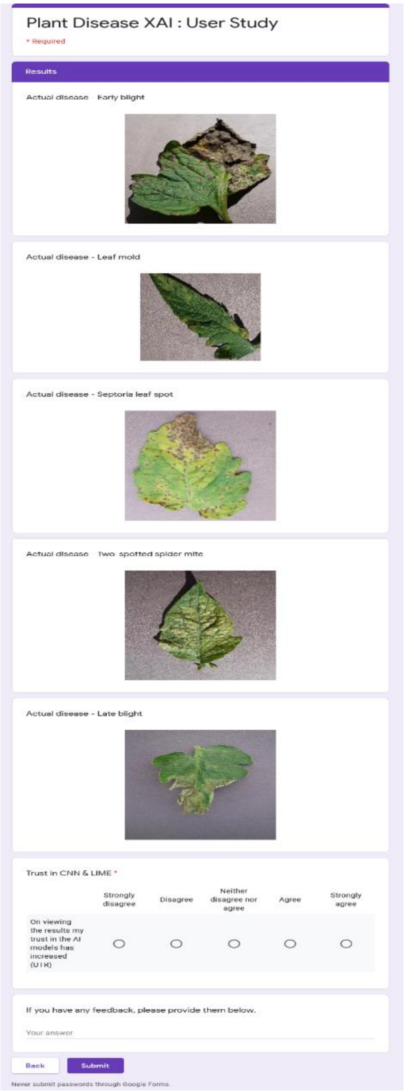


Fig 23 – Results (CNN and LIME)

The above figures are a screenshot of the google form that was sent to the farmers in order to collect their feedback. Similarly, for the KNN model another survey with a similar design was sent to the farmers.

The actual surveys can be accessed right below,

CNN Model [2]

KNN Model [3]

[2] https://forms.gle/KUEwBVoAePR24zMAA

[3] https://forms.gle/bUHhJHxAzvpBGJXE6

Appendix C

The code for the entire project is uploaded to GitHub and can be found in the following link [4].

The GitHub folder uploaded contains the following python files,

1. Implementation of CNN & LIME
2. Implementation of KNN & LIME
3. User study Visualization for CNN & LIME
4. User study Visualization for KNN & LIME

[4] https://github.com/nishvjay/Plant_disease_detection-.git