Deep Learning-Based Automated Detection of Tomato Plant Diseases: A Comparative Study of Convolutional Neural Network and k-Nearest Neighbor Approaches.

Aarin Badola[1], Arun Singh Bisht[2], Yudhveer Singh Panwar[3], Kunal Gulati[4], Dr. Sandhya Tarwani[5*]

[1,2,3,4,5] Artificial Intelligence-Data Science
Vivekananda Institute of Professional Studies-Technical Campus
aarinbadola633@gmail.com, bishtarun279@gmail.com, yudhveerp10@gmail.com, kunalgulati211@gmail.com, sandhya.tarwani@gmail.com

Abstract. The agriculture sector plays a pivotal role in the global economy, with its performance significantly impacting worldwide markets. Among the myriad challenges faced by the farming industry, plant diseases stand out as a major concern, leading to diminished crop and fruit yields and subsequent monetary losses. These diseases, caused by a variety of factors including viruses, bacteria, and fungi, underscore the critical need for early detection and intervention. In response, agriculture scientists have increasingly turned to machine learning (ML) models for disease detection and classification, with a recent emphasis on deep learning (DL) due to its superior accuracy. The ubiquitous tomato plant, with its global consumption surpassing 210 metric tonnes annually, faces a particularly high risk of disease. In this research paper, we present a deep learning approach for the automated detection of tomato plant diseases using high-resolution images of both diseased and healthy tomato plant leaves. Central to our methodology is image processing, encompassing feature extraction and classification techniques such as convolutional neural network (CNN) and k-nearest-neighbor (kNN) based classification. Leveraging the OpenCV library within the Python programming language, we facilitate seamless image and video processing, object detection, and feature extraction. Our experimentation utilizes several datasets comprising images of tomato plant leaves sourced from authenticated online repositories. Our findings demonstrate the efficacy of the CNN technique, achieving a maximum accuracy of 97.22% and underscoring its potential as a powerful tool in agricultural disease management.

Keyword: Plant Disease Detection, Machine Learning, Deep Learning, CNN, KNN

1 Introduction:

The Indian market is strongly dependent on agriculture, having a notable section of its population indulged in agriculture. The occurrence of plant disease has an adverse effect on agriculture production and standard. If plant disease is not discovered at time, food security will increase. Tomato is one the most important vegetable in daily meals used widely all across the world. Tomato plants are one of the most vulnerable species to diseases that can greatly affect the crop yield and production rate [6]. Losses can be minimized by making informed decisions. In earlier days, agricultural experts used their knowledge and visual inspection to detect diseases, but this process was time-consuming and often produced faulty results. To tackle these challenges, developers and researchers are turning towards advanced technologies, including image recognition and deep learning. Plants which are defected or diseased show some obvious appearances / marks on leaves, steam and fruit. Different diseases show different patterns through which we can identify

the diseases which will help to diagnose the abnormalities. The major reason to work on this paper is to create high quality yielding tomatoes, this will create a healthy chain. By integrating KNN and SVM clustering alongside deep learning techniques, researchers can enhance the accuracy and durability of disease detection models. Moreover, these algorithms play a crucial role in our project by providing insights into factors affecting plant health and productivity. Through a combination of deep learning, CNN, KNN, and SVM clustering, agricultural stakeholders can make informed decisions to ensure sustainable crop production and food security. Moreover, we are training deep learning models to detect multiple diseases at once, which provides an all-inclusive assessment of plant health. This ability is specifically important in agricultural settings where plants are prone to multiple diseases at once [10]. Advanced deep learning is utilized to optimize crop yield along with disease detection. Factors such as plant density, fruit size and ripeness help the model analysis, producing the results showing how the yield can be affected. This information can help farmers make informed decisions about crop management practices such as pruning and fertilization to maximize yield. In the end it can be summarized that deep learning and image recognition can be integrated because of its potential to transform tomato plant disease detection and crop management. By providing farmers with accurate information about plant health, these technologies can help reduce the use of pesticides, minimize crop losses and contribute to sustainable agriculture.

2 Literature Review:

Sladojevic et al. utilized Deep Learning and leveraged convolution networks to detect plant disease in various plants for which they put forward an effective solution. The model they built using CNN recognized 13 different types of plant diseases. The experimental outcomes from this re-search reveal that the model utilized in this case can prove satisfactory results and obtain an average accuracy of 96.3% [1]. K.P. Ferentinos employed Deep Learning and developed a convolution neural network to perform plant disease identification and evaluation using multiple images of healthy and diseased plant leaves. Several algorithms were opted and different model architectures were trained using a publicly available database comprising of 87,848 images and containing 25 different plants in a set of 58 distinct classes and the best accuracy came out to be at 99.53% [2].

S. Mohanty et al. detected 14 species and 26 diseases by collecting 54,306 images from a public dataset which included both diseased and healthy plants. These images were employed into training and testing datasets to model a convolution neural network with an accuracy of 99.35% [3]. F. Qin et al. used algorithms for pattern recognition which were based on image processing for the determination and analysis of the four specific types of diseases in alfalfa leaf, to improve the disease management in the alfalfa industry. 12 legion image segmentation methods were applied to obtain a total of 129 features and an SVM model was built using 45 most important features of the 129 obtained. The results indicated that the ac-curacy of this SVM model came at 94.74% [4]. Chai et al. combined computer vision and digital image processing and various approaches were analysed for the purpose of detection of four specific tomato plant leaf diseases that are late blight, early blight, Mold and leaf spot. Multiple leaf images were secured by setting up an image acquisition system. Various features were extracted by applying intensive image pre-processing to evolve a discriminant system and selecting 12 of the 18 variables were selected for the purpose of developing the Bayesian discriminant system. The outcome extracted reveals that an accuracy of 94.71% was achieved for the testing sets [5].

Vishnu. S et al. discuss how plant disease classification without the help of any dedicated technique can be very time consuming as well as costly specifically in remote areas and developing countries. These types of problems require an image processing-based solution that is divided into four phases which are; colour space transformation, image segmentation, feature extraction and finally feeding the pre-defined model along with its extracted features [6]. Murk Chouhan et al. propose

a deep learning-driven model which they call plant disease diagnosis tool The model takes images of leaves of plants and detects various diseases. They have created a convolution neural network and used its multi-layered functionalities to train the model by extracting features from the images. A maximum of 98.3% testing accuracy was achieved by the model [7]. Anshul Bhatia et al. utilized Extreme Learning Algorithm which is a subset of Machine Learning for detection of plant disease. They have used a practical situation-based dataset called Tomato Powdery Mildew Disease (TMPD) dataset. Multiple techniques for the purpose of resampling like Importance Sampling, Synthetic Minority Over-sampling Technique (SMOTE), Random Under Sampling (RUS), and Random Over Sampling (ROS) were applied to balance the dataset. Best results were obtained using the Importance Sampling technique which gave the peak values for Area Under Curve and Classification Accuracy that came out at 88.57% and 89.19% respectively [8].

3 METHODOLOGIES:

Deep Learning is a part of Machine Learning and it is utilized because of its high efficiency and capabilities for training Convolutional Neural Networks [CNN] while learning from large datasets. Deep Learning has made the task of feature extraction and engineering easier and faster. Deep Learning utilizes multiple layers through which data is passed within the neural network. Deep Learning in particular utilizes Artificial Neural Networks [ANN] which tries to replicate the human brain where information is passed through neurons and synapses.

3.1 SYSTEM OVERVIEW:

The block diagram gives the general overview of the system and the step-by-step procedure followed in this paper as shown in Figure 1.

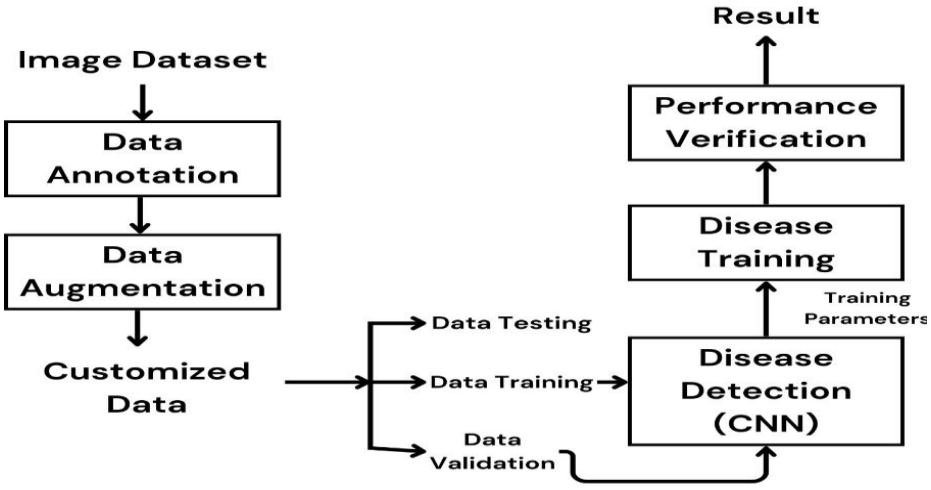

Fig 1: Proposed Methodology for Disease Discussion

3.2 Data Collection and Annotation:

The dataset used for the training and the testing set includes several images of both healthy and diseased tomato plants. This is a public dataset collected from a trusted source and contains pictures from different instances and different surroundings (e.g. lighting, intensity, angles etc.). The images in this particular dataset depict leaf diseases in tomato plants. The dataset classifies images under different names so that it becomes easier to distinguish between various diseases (some diseases may look alike depending on the disease progression). The primary goal of this annotation process is to identify the classes and segregate the images in various classes depending on the particular infection. The description of the dataset is available in Table 1.

Table I Overview of the dataset

Dataset Title	Description	Source	Size	URL
Plant Village	This dataset contains multiple images of diseased as well as healthy tomato plant leaves.	Kaggle	248 MB	https://www.kaggle.com/arjuntejaswi/plant-village

3.3 Image Pre-Processing:

This step is essential for the actual functioning of the model. Image pre-processing means manipulating an image before it is fed into the computer vision and it is required for enhancing the quality of image by suppressing undesired distortions, noise reduction, etc. This process becomes even more essential as it helps in increasing the accuracy of the ML model many folds.

3.4 Dataset Discussion:

The dataset used to perform plant disease detection contains classes and each class has multiple images of various tomato leaf diseases and a total of 16,066 has been used which are further separated into training, testing and validation datasets. Table 2 provides description of these classes and dataset.

Table II Dataset Elucidation

Class	Plant Name	Health Status	Disease Name	Images(numbers)
C_0	Tomato	Healthy	None	1591
C_1	Tomato	Diseased	Late Bright	1920
C_2	Tomato	Diseased	Early Blight	1000
C_3	Tomato	Diseased	Bacterial Spot	2127
C_4	Tomato	Diseased	Yellow leaf curl virus	3232
C_5	Tomato	Diseased	Two Spotted spider Mite	1676
C_6	Tomato	Diseased	Mold	952
C_7	Tomato	Diseased	Target Spot	1414
C_8	Tomato	Diseased	Mosaic Virus	373
C_9	Tomato	Diseased	Septoria Leaf Disease	1781

The Table provides a quick study of the dataset and classes within it. The number of images in these classes lies between 300 and 3000. Multiple diseases of the tomato plant leaf are available in the dataset and corresponding to them is the number of instances they have occurred in the dataset. Figure 2 shows that the most spread disease is the Yellow Leaf Curl Virus and the least spread is the Mosaic Virus. Below are attached some of the images from the dataset that show the diseased tomato plants.

Figure 2 Sample of Various Tomato Plant Leaves from Dataset

4 Algorithms Used:

4.1 Convolution Neural Network

We have employed convolution neural networks for the task of disease detection and classification in tomato plant leaves. Convolution neural networks area class of deep neural networks that give particularly efficient results on images or visual imagery. Image classification tasks heavily utilize CNN to produce results with highest accuracy as it captures the repeating patterns in the images and works on them to do the classification. CNN majorly utilizes two layers; Convolution Layer and Pooling Layer as shown in Figure 3. The convolution layer applies convolution operations on the images to detect patterns in an image using filters, each filter captures some feature from the images. After the convolution layer, the pooling layer is deployed to reduce the spatial dimension of the features and retain the most important information.

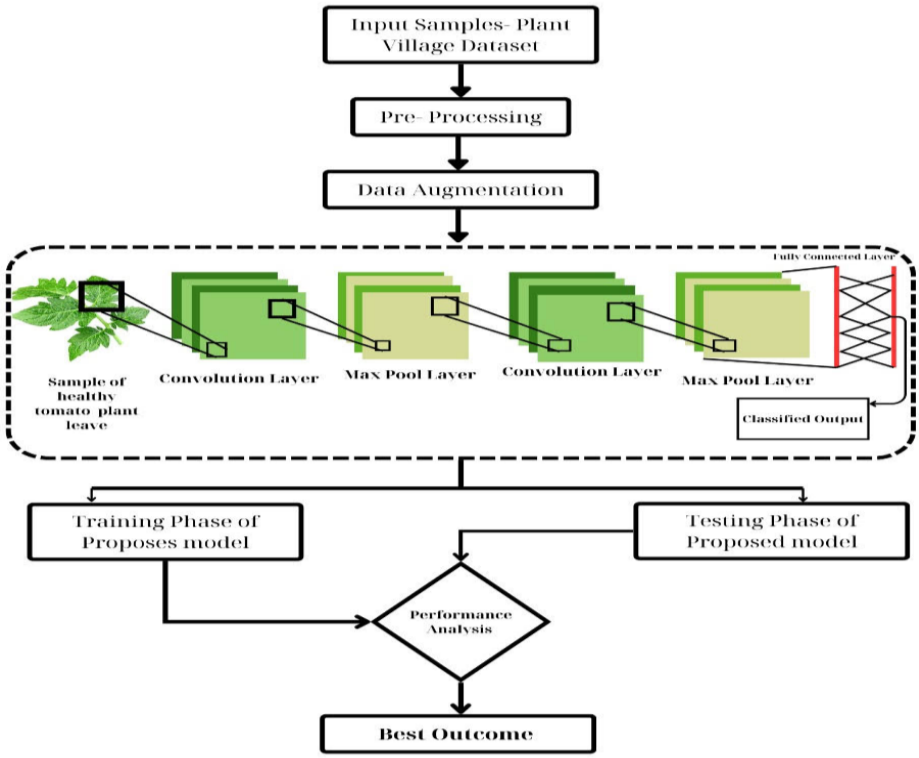

Figure 3 CNN Architecture

4.2 K Nearest Neighbor

K nearest-neighbor is a Machine Learning based algorithm, used specially for classification and regression-based tasks. It belongs to the category of supervised learning and hence depends on labelled datasets to perform classification on any data. KNN is heavily employed in the tasks that require feature recognition and hence it becomes a good fit for tomato plant disease detection model. Working with KNN is widely classified into 3 steps; training phase, prediction phase and choosing the value of 'K'. Now, since KNN falls under the category of a lazy algorithm, it doesn't openly go through the training phase but stores the training data which is further used for making predictions. During the training phase KNN finds the distance (generally Euclidean distance) between the data point as shown in Figure 4 and the pre-existing points in the dataset to find the closest neighbors and proceeds in different ways depending on the requirement i.e regression or classification. In the third step, the algorithm needs to choose a value for 'K', which can also be understood as the number of neighbors to consider, is the most important step as it determines whether the algorithm is going to be noisy (when k is small) or if the algorithm smooths out the detail's way too much (when k is very large).

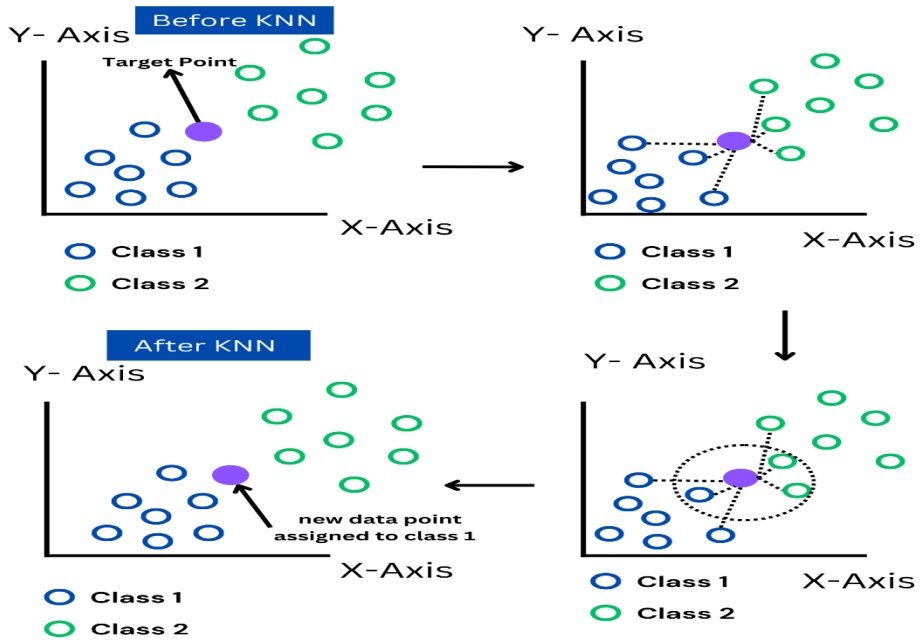

Figure 4 KNN Architecture

4.3 Support Vector Machine

Support vector machine is another Machine Learning algorithm which we have employed in our study as it is one of the best working algorithms for classification and regression type problems as shown in Figure 5. The idea behind the SVM is division of various classes in a dataset into different data points that are separated with the help of a Maximal Marginal Plane. SVM also falls under the category of supervised learning and uses labelled datasets for its functioning. There are various data points that get separated because of the hyperplane and the points lying closest to the decision line are known as support vectors which are used for defining the position and the placement of the hyperplane. The purpose of SVM is to maximize the margin for which it uses the Kernel trick where non-linear data is handled by converting it into higher dimensional space using the kernel function. As the process of finding the best solution progresses with the help of SVM, classification of new data points takes place by determining which side of the hyperplane they lie on. SVM becomes one of the most suitable algorithms for classification tasks as it works better towards avoiding overfitting as compared to other algorithms.

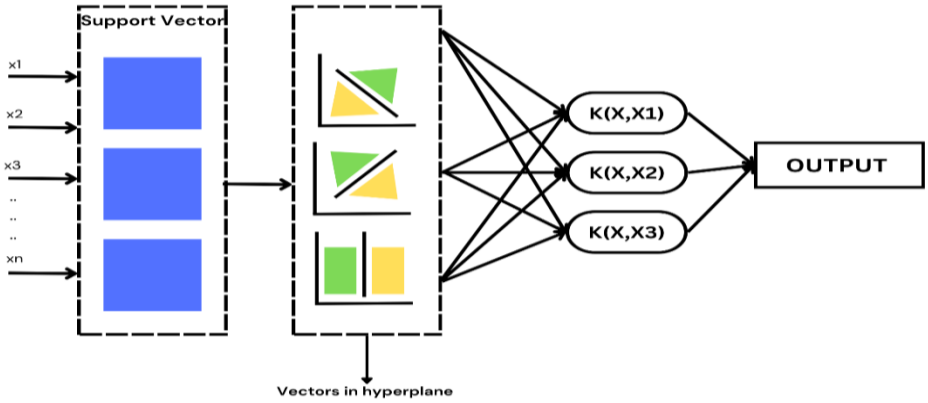

Fig. 5 SVM Architecture

5 Result and Discussion:

This paper shows why the plant disease detection model is essential and how it can be utilized for taking timely measures. This model was trained using Deep Learning and Convolution Neural Network which employed python. 10% images, that comes out to be at 1606 were taken from the dataset and utilized for the purpose of testing the model accuracy. From the 10 different classes available in the dataset, 10% of the images were selected randomly from each class to generate the best possible accuracy. After multiple training and testing runs the accuracy of the model comes out to be at 97.22% while predicting healthy and unhealthy leaves of the tomato plant. The graphs as shown in Figure 6, Figure 7 and Figure 8 describes the Training accuracy as well as the Validation accuracy produced by our model.

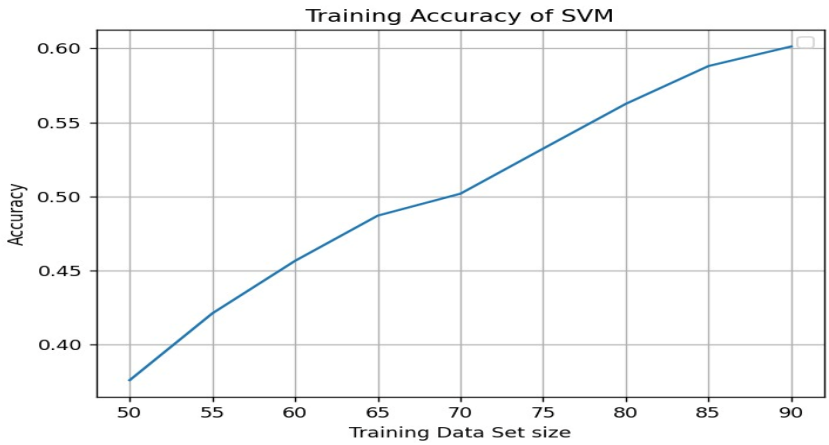

Fig. 6 Training Accuracy Curve for SVM

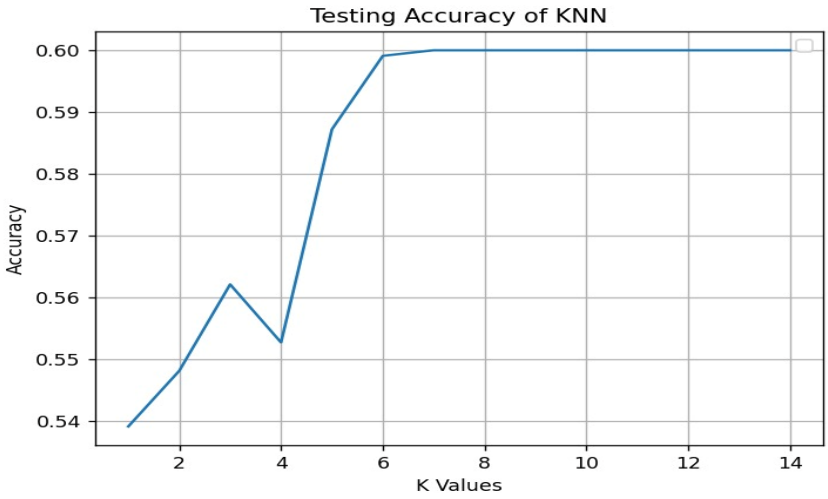

Fig. 7 KNN Model Testing Accuracy Graph

Table III Training Parameters for CNN Model

PARAMETER	VALUES
Batch Size	32
Activation in middle layer	Relu
Activation in final layer	SoftMax
Epochs	20
Channels	3

Fig. 8 Training Vs Validation Accuracy Graph for CNN

The Table IV given below compares the results of various classification techniques used, under different metrics like recall, F1 score and accuracy.

Table IV Performance Metrices of Different Classifiers

Model	Accuracy	Recall	F1 Score
CNN	0.9722	0.993392	0.989495
KNNS	0.60	0.626428	0.610674
SVM	0.5625	0.574761	0.593857

We have plotted a line chart for the performance comparison of various ML and DL classification techniques used with metrics (accuracy, F1 score and recall) on y-axis and different models on x-axis as shown in Figure 9. These multiple tables and graphs conclude that we have achieved the best possible results using the CNN technique followed by the other classifiers.

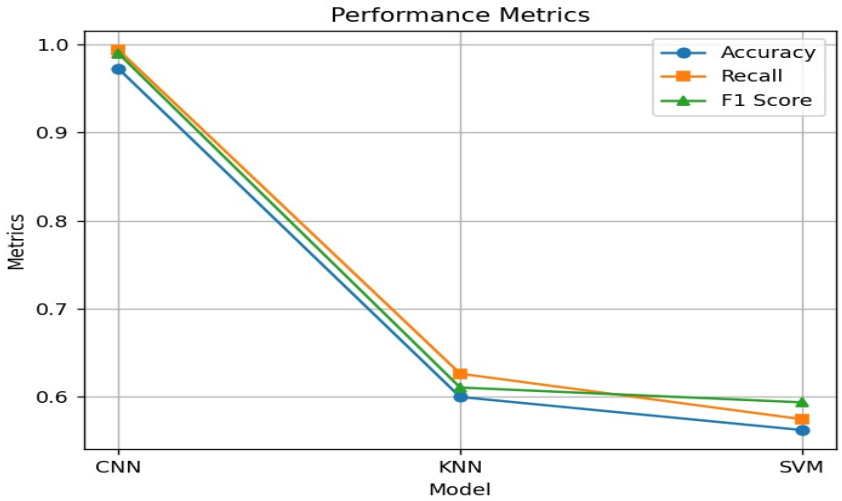

Fig. 9 Performance Comparison of Various ML and DL Models

6 Conclusion and Future Direction:

The overall study was centered around Deep Learning and revolved around it to de-sign a framework that can automatically predict whether the plant is diseased or not. The model works on a simple mechanism that uses the function of feature extraction within the convolution neural network developed. CNN utilizes its multi-layered architecture for the purpose of classification and prediction. For this research a publicly available dataset of 16066 images was used, that contains images of various unhealthy and healthy tomato plant leaves under 10 different classes. The model has produced a general accuracy of 97.22% on the dataset used. Multiple classification algorithms such as SVM, KNN and CNN were utilized and the study shows that CNN is the most accurate in detecting plant diseases. This model can further be trained for multiple plants that will increase the dimensions of its application and usage. With further industrial advancement we can directly deploy this model in the fields that would help in real-time monitoring and early detection of plant diseases. Advanced research and improved studies can help in development of disease fore-casting models that would analyze historical data and environmental factors to fore-cast disease outbreaks. Automated disease detection, classification and treatment can also be achieved by integrating this technology with robotics. This model can be made readily available to all the farmers as well as others by developing it into mobile applications.

7 References:

[1] S Sladojevic, M. Arsenovic, A. Andrela, D. Culibrk and D. Stefanovic, “Deep Neural Networks based recognition of plant diseases by leaf image classification,” 2016.

[2] K. P. Ferentinos, “Deep Learning models for plant disease detection and diagnosis,” 2018.

[3] S. Mohanty, D. Hughes and s. Marcel, “Using deep learning for image-based plant disease detection,” 2016.

[4] F. Qin, D. Liu, B. Sun, L. Ruan, Z. Ma and H. Wang, “Identification of alfalfa leaf diseases using image,” 2016.

[5] A.L. Chai, B.J. Li, Y.X. Shi, Z.X. Cen, H.Y. Huang, J. Liu, “Recognition of tomato foliage disease based on computer vision technology,” 2010.

[6] Vishnu. S, A. Ranjith Ram, “Plant disease detection using leaf pattern: A review,” 2015.

[7] Murk Chohan, Adil Khan, Rozina Chohan, Saif Hassan Katpar, Muhammad Saleem Mahar, “Plant disease detection using deep learning,” 2020.

[8] Anshul Bhatia, Anuradha Chugh, Amit Prakash Singh, “Application of extreme learning machine in plant disease prediction for highly imbalanced dataset,” 2020.

[9] LiLi Li, Shujuan Zhang, Bin Wang, “Plant disease detection and classification by deep learning- a review,” 2021.

[10] Muhammad Hammad Saleem, Johan Potgieter, Khalid Mahmood Arif, “Plant disease detection and Classification by deep learning,” 2019.