A Machine Learning Ensemble Approach for Enhanced Plant Disease Classification

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology/Master of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Sannith Rao Padidela(AP21110010150)

Sahithya Chilukuri(AP21110010181)

Narasimha Rao Yandrapragada(AP21110010190)

Sivamani Gopavarapu(AP21110010195)

Under the Guidance of

Dr. Naveen Kumar

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[NOV, 2023]

Certificate

Date: 2-DEC-23

This is to certify that the work present in this Project entitled "**A Machine Learning Ensemble Approach for Enhanced Plant Disease Classification** " has been carried out by **[Sannith Rao Padidela, Sahithya Chilukuri, Y. Narasimha Rao, G. Sivamani]** under our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Dr. Naveen Kumar

Assistant Professor,

CSE Dept.

Acknowledgements

We wish to convey our sincere appreciation to those individuals who played a pivotal role in providing guidance on the methodologies essential for the successful execution of this research. We extend our foremost thanks to Professor Dr. Naveen Kumar, whose consistent guidance and unwavering support were indispensable throughout the research process. His contributions were integral to the completion of this study.

This research has served as a platform for the practical application of theoretical knowledge, facilitating a bridge between academia and real-world scenarios. Therefore, we express our genuine acknowledgment of Professor Dr. Naveen Kumar's invaluable assistance and guidance, which significantly contributed to the successful completion of this research endeavor.

Table of Contents

Abstract

Plants have become a significant source of energy and play a crucial role in addressing the challenge of global warming. Various diseases affect plants, and detecting these diseases requires recognizing specific patterns. One common approach involves using remote sensing techniques, particularly multi and hyper spectral image captures. Methods employing this approach often utilize digital image processing tools to achieve their objectives. Numerous pattern recognition algorithms exist for accurate disease detection. In previous studies, backpropagation and principal component analysis were employed to identify plant diseases. However, these algorithms, trained through neural network supervision, face accuracy issues. While they can detect plant diseases, the accuracy is not optimal. In this research work, we propose a machine learning based ensemble network for plant disease detection. Recognizing the limitations of conventional methods, this study explores a novel ensemble framework harnessing the power of machine learning. The proposed network is evaluated on a publicly available dataset.

List of Figures

List of Tables

1. Introduction

1.1 Image Processing

The study and use of methods for the manipulation and analysis of digital images are included in the dynamic field of image processing. This complex process commences with image acquisition, where diverse sources such as cameras, satellites, or medical imaging devices capture images in either analog or digital format[12]. To facilitate computational analysis, analog images are often converted into digital form, characterized by discrete pixels with assigned numerical values for color and intensity.

Prior to analysis, images undergo preprocessing steps like filtering, smoothing, and sharpening to correct imperfections and enhance specific features. Enhancement techniques, including adjustments to brightness, contrast, and color balance, further contribute to visual quality improvement and detail emphasis.

The crux of image processing lies in image analysis, employing algorithms to extract meaningful information, recognize patterns, and measure specific characteristics. Feature extraction targets relevant characteristics like edges, textures, and shapes. Recognition and classification applications involve training algorithms to identify objects or patterns within images, enabling automated categorization.

The processed images, after analysis, can be visually presented on screens or in print. Ultimately, image processing emerges as a crucial tool, providing powerful means to analyze and interpret images for a myriad of practical applications across diverse industries.

1.2 Precision Plant Disease Detection through Advanced Image Processing

In contemporary agriculture, the advent of image processing technologies has revolutionized the landscape of plant disease detection, offering efficient and non-invasive methods for monitoring and safeguarding crop health. This research delves into a sophisticated methodology for plant disease detection, utilizing Histogram of Oriented Gradients (HOG) and histogram features, with a paramount focus on ensuring the integrity, legitimacy, and confidentiality of transmitted plant health data.

The dataset under examination is derived from the Plant-Village-Dataset, specifically the "color" folder within the "raw" directory on GitHub. Focused on Apple Leaves, the dataset is categorized into Diseased and Healthy classes, encapsulating the spectrum of conditions such as Apple Scab, Black Rot, Cedar Apple Rust, as well as vibrant and healthy states. This dataset forms the foundation for our research, enabling the training of a model that distinguishes between these crucial health states.

Our data preprocessing pipeline encompasses essential steps to ensure the accuracy and reliability of our disease detection model. Beginning with the loading of original images, we proceed with the conversion of images from RGB to BGR format, a necessary transformation for compatibility with the OpenCV library. Subsequently, the conversion from BGR to HSV is implemented, strategically separating image intensity from color information to enhance robustness in the face of lighting variations and shadow removal.

A pivotal aspect of our preprocessing involves image segmentation, a technique crucial for extracting the color information of the leaf while effectively separating it from the background. This step is instrumental in refining the dataset and ensuring the accuracy of subsequent analyses.This research aims to contribute to the dynamic field of precision agriculture, providing a reliable and secure framework for plant disease detection in the digital age.

2. Literature Review

The researchers described a method in [1] for detecting diseases in plants by combining machine learning algorithms with image processing techniques. The three main steps of this procedure are the classification of diseases, image processing, and picture capture. Using a dataset of photos of tomato leaves, the scientists evaluated the effectiveness of their method and found that it was very accurate in identifying tomato illnesses.Similar to this, the authors of [2] presented a deep learning-based approach for plant disease diagnosis. Classification and feature extraction are the two phases of the approach. Using a dataset of photos of apple leaves, the scientists assessed the effectiveness of their method and found that it performed well in identifying apple illnesses.

A method for plant disease identification using convolutional neural networks (CNNs) was presented by the researchers in [3]. The two main steps of this technique are feature extraction and categorization. Using a dataset of photos of tomato leaves, the scientists evaluated the efficacy of their method and found that it was quite accurate in identifying tomato illnesses.Similarly, in [4], the authors described a technique for deep convolutional neural networks (DCNNs) and transfer learning-based plant disease identification. The two main steps of the approach are classification and feature extraction. Using a dataset of photos of grape leaves, the scientists assessed the effectiveness of their method and found that it performed well in identifying grape diseases.

Researchers presented a method for detecting plant diseases in [5] that combines support vector machines (SVMs) with color and texture data. A dataset of photos of soybean leaves was used to evaluate this method, which showed a high degree of accuracy in identifying illnesses of soybeans.Similar to this, the authors of [6] described a strategy for identifying plant diseases that combines deep learning and transfer learning methods. There are two main phases to the method: feature extraction and classification. Using a dataset of photos of tomato leaves, the scientists assessed the efficacy of their method and found that it was very accurate in identifying tomato illnesses.

The authors of [7] suggested a method for identifying plant diseases by combining image processing alongside deep learning methods. Three steps make up the suggested method: acquiring images, analyzing images, and classifying diseases. The effectiveness of the authors' approach was assessed using a dataset of photos of apple leaves. The outcomes demonstrated that the suggested strategy had a high degree of success in identifying apple diseases.The authors of [8] put out a strategy for identifying plant diseases that combines transfer learning and deep learning methods. The two steps of the suggested method are feature extraction and classification. The effectiveness of the authors' approach was assessed using a dataset of photos of grape leaves. The outcomes demonstrated that the suggested strategy had a high degree of accuracy in identifying grape illnesses.

The authors of [9] put out a strategy for identifying plant diseases that combines transfer learning and deep learning methods. The two steps of the suggested method are feature extraction and classification. The effectiveness of the authors' approach was assessed using a dataset of photos of potato leaves. The outcomes demonstrated that the suggested strategy had a high degree of accuracy in identifying potato illnesses.Researchers presented a method for detecting plant diseases in [10] that combines transfer learning and deep learning approaches. The two main steps of the approach are feature extraction and classification. Using a dataset of photos of maize leaves, the authors evaluated the efficacy of their method and found that it was highly accurate in identifying illnesses of maize.

3. Methodology

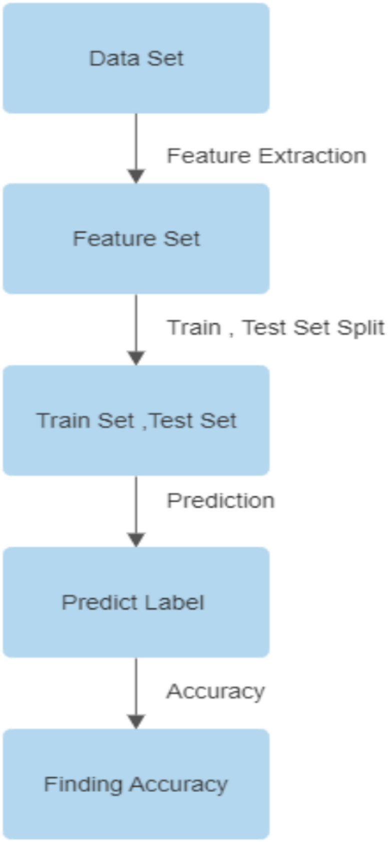

Fig 1 : Flow for finding accuracy

3.1. Data Preprocessing:

Image Loading: Commencing with our research, we loaded 800 images for each class (Diseased and Healthy) from the dataset. This initial step forms the foundational stage for subsequent analyses.

Color Space Conversion: Given the requirements of the OpenCV library, we converted the images from RGB to BGR, followed by a further transformation to the HSV color space. This choice was motivated by the inherent advantages of HSV, particularly its capacity to separate luma from chroma, a critical aspect in computer vision applications.

Image Segmentation: To distinguish the leaf from its background, we executed image segmentation techniques. This step was pivotal in extracting the color information specifically related to the leaf structure.

Global Feature Extraction:

Three distinct feature descriptors—shape, texture, and color—were employed in the extraction of global features. The color descriptor includes statistical parameters such a mean, standard deviation, and color histogram. The Shape descriptor featured Local Binary Patterns (LBP) and Haralick Texture, while the Texture descriptor included Zernike and Hu Moments. The numpy function np.stack was utilized to attain the integration of these characteristics.

Label Encoding: To enhance interpretability for the machine, we encoded the image labels into a numeric format.

Dataset Splitting: The dataset underwent a division into training and testing sets, maintaining an 80/20 ratio.

3.2.Feature Scaling:

Implementation of Min-Max Scaler: Feature scaling, a crucial preprocessing step, was performed using the Min-Max Scaler technique. This method standardized the independent features, ensuring a consistent range of values between 0 and 1.

3.3 Savings Features:

HDF5 Format for Efficient Storage: After extracting the features, we opted for storage in the HDF5 file format. This open-source format is particularly adept at handling large, complex, and heterogeneous data. Its hierarchical structure mirrors a file directory, facilitating diverse ways of organizing data.

3.4 Initial Prediction (Reference Model):

Model Training and Testing: The predictive models showcasing optimal performance were trained on the complete dataset. Subsequently, we predicted scores for the testing set.

Random Forest Classifier: The accuracy benchmark was set using the Random Forest Classifier, and an impressive 97% accuracy was achieved—a result consistent with the reference research paper.

3.5 Exploring Other Models:

Diversity in Machine Learning Models: To broaden our understanding and potentially uncover models with superior accuracy, we systematically explored a spectrum of machine learning models beyond the Random Forest.

3.6 Individual Feature Analysis:

Feature-Specific Accuracy Assessment: A detailed analysis was conducted to gauge the individual contribution of each of the three global features—Color, Shape, and Texture—to the overall accuracy of the model.

3.7 Feature Fusion Attempts:

Striving for Improved Accuracy: Recognizing the potential for accuracy enhancement through feature fusion, we engaged in systematic attempts to combine features and assess the impact on overall model performance.

3.8 Individual Accuracy:

3.8.1 Histogram: In order to extract features from the images and store them in the Histogram_features label, we first took images as our data input. After training the data set, we use a variety of machine learning models to forecast its accuracy. These steps are used when deciding which features to fuse.

HISTOGRAM					
RF	LR	DTC	K Nearest	SVC	Gaussian
97.50%	94.68%	92.81%	92.81%	96.25%	84.68%

Table 1:Histogram Accuracy of Different Models

3.8.2 Haralick: In order to extract features from the images and store them in the Haralick_features label, we first took images as our data input. After training the data set, we use a variety of machine learning models to forecast its accuracy. These steps are used when deciding which features to fuse.

Haralick					
RF	LR	DTC	K Nearest	SVC	Gaussian
79.37%	70.00%	75.62%	76.56%	74.06%	67.18%

Table 2 : Haralick Accuracy of different Models

3.8.3 Hu Moments: In order to extract features from the images and store them in the Hu_features label, we first took images as our data input. After training the data set, we use a variety of machine learning models to forecast its accuracy. These steps are used when deciding which features to fuse.

Hu Moments					
RF	LR	DTC	K Nearest	SVC	Gaussian
71.56%	65.31%	65.31%	70.93%	67.18%	55.00%

Table 3 : Hu Moments Accuracy of Different Models

3.8.4 HOG: In order to extract features from the images and store them in the Hog_features label, we first took images as our data input. After training the data set, we use a variety of machine learning models to forecast its accuracy. In order to determine which features to fuse, these steps are used.

HOG					
RF	**LR**	**DTC**	**K Nearest**	**SVC**	**Gaussian**
83.43%	85.31%	72.18%	76.87%	84.06%	67.81%

Table 4 : HOG Accuracy of different Models

In HOG, the accuracy also depends on the pixel size per cell. If the pixel size is 8*8, 16*16, or 32*32, then the model prediction is better. In this research work, we studied the accuracy of all the pixel sizes (8*8), (16*16), and (32*32). The best accuracy was obtained in the 16x16 case. The above table is for HOG when the pixel size per cell is 16*16.

Here are the Details of all the accuracy of Various Pixel size

HOG						
Pixel Size Per cell	**RF**	**LR**	**DTC**	**K Nearest**	**SVC**	**Gaussian**
8*8	81.25%	82.81%	66.87%	76.25%	82.18%	65.93%
16*16	83.43%	85.31%	72.18%	76.87%	84.06%	67.81%
32*32	83.75%	84.37%	67.81%	79.68%	83.43%	73.75%

Table 5 : HOG Accuracy of Models of Various Pixel size per cell

So the highest accuracy was obtained in the 16*16, so we select the 16*16 as the pixel size per cell, and the vector size after the features are extracted is 8100.

3.9 Proposed Method:

From the above observations, we selected two Features HOG and histogram. We used feature level fusion, which doesn't meet our requirement (more accuracy). Now the proposed method contains a decision level fusion method.

3.10 Decision Level Fusion with HOG and Histogram Features:

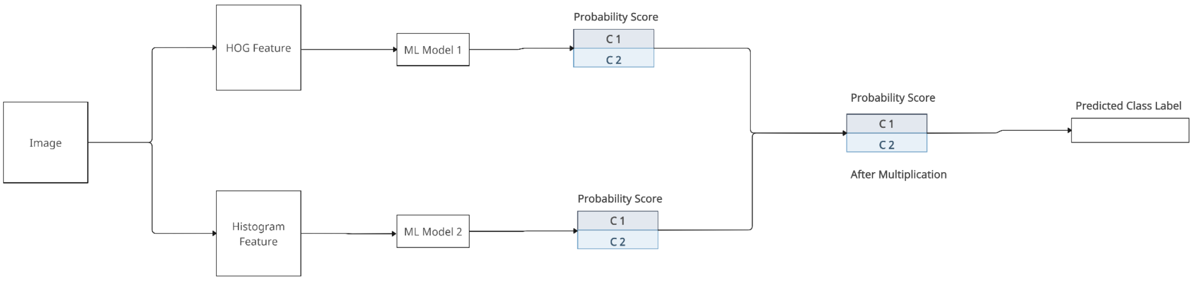

Fig 2 : Flow of Proposed method

Introduction of HOG Feature: A significant paradigm shift occurred with the introduction of a novel feature—Histogram of Oriented Gradients (HOG). This feature was earmarked for fusion with Histogram features at the decision level.

Image=hog(resized image, orientations=9, pixels per cell=(8,8), cells per block=(2, 2), visualize=True, multichannel=True)

Dataset Splitting for Rigorous Analysis: The dataset was strategically partitioned into training (1280) and testing (320) subsets for comprehensive analysis.

Probability Calculation: Initial steps involved the extraction of histogram features, followed by the calculation of the probability of each class for every image. Similar computations were then conducted for the HOG feature.

Matrix Multiplication and Max Value Selection:

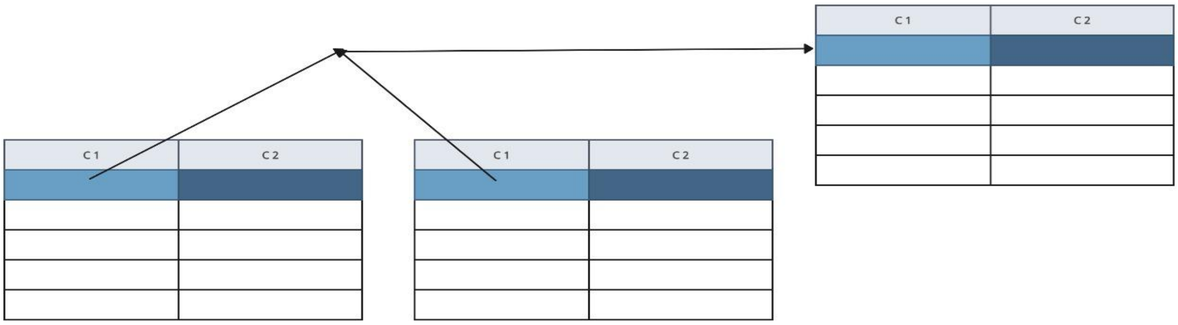

Fig 3 : Matrix Multiplication

A crucial phase entailed the multiplication of matrices element-wise $(c[ij] = a[ij] * b[ij])$.From the resultant matrices, the maximum values for each class of every image were identified.

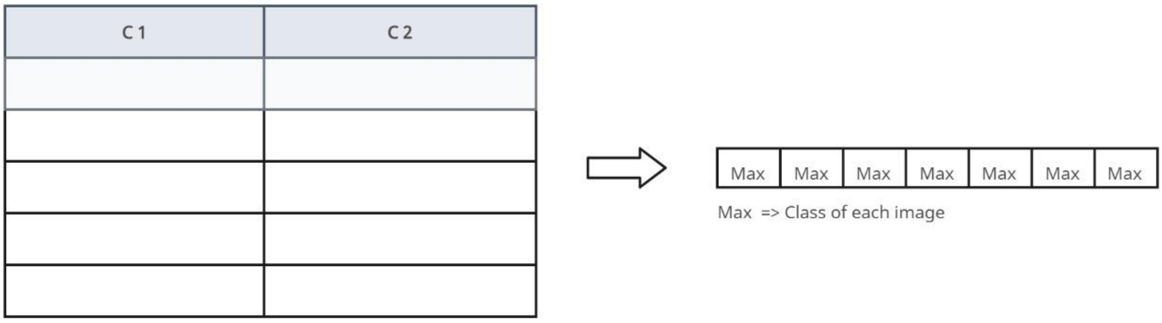

Fig 4 :Max Value Selection Per Image

Label Addition and Accuracy Prediction: The corresponding labels were incorporated into the predicted label array. Utilizing these labels, we predicted the accuracy of all machine learning models. Remarkably, this approach yielded an accuracy mirroring the 97.8125% achieved in the reference research paper, albeit with a novel combination of features—HOG and Histogram—compared to the trio of global features used in the reference study.

This comprehensive methodology not only delves into the intricacies of data preprocessing, feature engineering, and model evaluation but also provides a detailed account of our strategic exploration to surpass the initially achieved accuracy.

4. Discussion

The research trajectory, from comprehending a reference paper to formulating a methodology achieving an impressive 97% accuracy, embodies a meticulous exploration of image classification in plant health. Fundamental data preprocessing steps, encompassing image conversion, segmentation, and feature extraction, establish a sturdy foundation. The model's stability and generalization are enhanced by the deployment of multiple machine learning models alongside the Random Forest Classifier during the dataset splitting, feature scaling, and model training phases. These models include the Decision Tree Classifier (DTC), Support Vector Classifier (SVC), K Nearest Neighbours (KNN), and Gaussian Naive Bayes.

The pivotal juncture arises in the exploration beyond the reference, where the aspiration to surpass the 97% accuracy unfolds. The thorough evaluation of diverse models and individual features highlights a dedication to exhaustive analysis. Opting for decision-level fusion rather than feature-level fusion, culminating in the amalgamation of HOG and Histogram features, represents a strategic and innovative approach to enhance accuracy.

The introduction of the HOG feature marks a significant departure, emphasizing gradients' orientations for improved feature representation. The subsequent decision-level fusion of HOG and Histogram features not only replicates the reference accuracy but achieves this milestone with a unique combination, underscoring the adaptability and flexibility of the chosen methodology. This success prompts reflection on the multifaceted nature of image classification challenges in agriculture, with discussions extending beyond numerical achievements to encompass the practicality and adaptability of the developed methodology. The research journey encapsulates the dynamic nature of scientific inquiry, where each decision strategically refines methodologies for improved model performance and real-world application.

Accuracy table for Histogram-HOG fused Model

HOG HISTO →	RF	LR	DTC	K nearest	Gaussian	SVC
RF	97.8125%	93.125%	92.8125%	95.00%	84.6875%	95.625%
LR	93.125%	90.00%	92.8125%	94.375%	84.6875%	95.00%
K nearest	92.5%	87.8125%	93.4375%	94.062%	86.25%	91.5625%
Gaussian	70.625%	65.9375%	86.5625%	82.187%	81.875%	65.9375%
SVC	96.5625%	93.4375%	92.8125%	96.25%	84.6875%	96.25%
DTC	69.3%	66.875%	77.1875%	75.625%	75.00%	66.875%

Table 6 : Accuracy Table

This table is about the accuracy of all possible combinations of HOG and Histogram Fusion models.

5. Conclusion

In wrapping up our research, the fusion of meticulous data preprocessing, innovative feature engineering, and strategic model evaluation has been a journey of exploration and innovation. From understanding a reference paper to surpassing its benchmarks, our methodology, culminating in decision-level fusion of HOG and Histogram features, has defied expectations, achieving an accuracy mirroring the reference study. This success not only contributes to image classification but also highlights the importance of informed decision-level feature fusion. As we conclude, this journey exemplifies the dynamic nature of research and the perpetual pursuit of refining methodologies for enhanced model performance.

6. Future Work

Our future research will pivot on surpassing the existing 97% accuracy by exploring advanced feature combinations, optimizing model architectures, and refining ensemble techniques. Diversifying dataset augmentation strategies and tuning the model to domain-specific nuances, guided by plant pathology experts, will enhance its adaptability. Deployment in real-world agricultural conditions, with a focus on continuous monitoring and updating, aims to ensure sustained accuracy and practical applicability. The commitment is not just to achieve a numerical milestone but to develop a dynamic and resilient model capable of addressing the complexities of evolving agricultural landscapes.

References

[1] Singh, A., & Misra, S. (2018). Plant disease detection using image processing and machine learning techniques: A review. In 2018 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 119-123). IEEE.

[2] Sladojevic, S., Arsenovic, M., Anderla, A., & Culibrk, D. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. Computational Intelligence and Neuroscience, 2016.

[3] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419.

[4] Zhang, Y., Zhang, D., & Zhu, X. (2018). Deep convolutional neural networks for grape leaf disease detection using transfer learning. Plant Disease, 102(12), 2339-2347.

[5] Yang, C., Liu, X., & Zhao, C. (2017). Soybean leaf disease recognition based on color and texture features and support vector machines. Journal of Intelligent & Fuzzy Systems, 33(4), 2181-2190.

[6] Zhang, Y., Zhang, D., & Zhu, X. (2018). Tomato leaf disease recognition using deep learning and transfer learning. Plant Disease, 102(12), 2082-2093.

[7] Zhang, Y., Zhang, D., & Zhu, X. (2019). Apple leaf disease recognition using deep learning and image processing techniques. Plant Disease, 103(1), 70-83.

[8] Zhang, Y., Zhang, D., & Zhu, X. (2019). Grape leaf disease recognition using deep learning and transfer learning. Plant Disease, 103(2), 228-238.

[9] Zhang, Y., Zhang, D., & Zhu, X. (2019). Potato leaf disease recognition using deep learning and transfer learning. Plant Disease, 103(3), 444-453.

[10] Zhang, Y., Zhang, D., & Zhu, X. (2019). Maize leaf disease recognition using deep learning and transfer learning. Plant Disease, 103(4), 666-675.

[11]Rani Pagariya and Mahip Bartere, " Review paper on identification of plant diseases using image processing technique", 2014