Journal of Robotics and Control (JRC)
Volume 4, Issue 6, 2023
ISSN: 2715-5072, DOI: 10.18196/jrc.v4i6.20342

Plant Leaf Disease Detection Using Efficient Image Processing and Machine Learning Algorithms

Kiran S M [1*], Dr. Chandrappa D N [2]
[1] Research Scholar, Dept. of ECE, SJBIT, Bangalore, Affiliated to VTU Belagavi, Karnataka, India-590018
[2] Professor, Dept. of ECE, SJBIT, Bangalore, Affiliated to VTU Belagavi, Karnataka, India-590018
Email: [1] sm.kirana@gmail.com
*Corresponding Author

***Abstract*—India is often described as a country of villages, where a majority of the population depends on agriculture for their livelihood. The landscape of Indian agriculture is approximately 159.7 million hectares. Agriculture plays a pivotal role in India's Gross Domestic Product (GDP), accounting for about 18% of the nation's economic output. Diseases and pests can have detrimental effects on crops, leading to reduced yields. These challenges can include the spread of plant diseases, infestations by insects or other pests, and the overall degradation of crop health. Early detection of diseases in crops is crucial for several reasons. Detecting diseases at an early stage allows for prompt intervention, such as applying appropriate pesticides or taking preventive measures. The main aim of this study is to develop a highly effective method for plant leaf disease detection using computer vision techniques. Here, leaf disease detection comprises histogram equalization, denoising, image color threshold masking, feature descriptors such as Haralick textures, Hu moments, and color histograms to extract the salient features of leaf images. These features are then used to classify the images by training Logistic Regression, Linear Discriminant Analysis, K-nearest neighbor, decision tree, Random Forest, and Support Vector Machine algorithms using K-fold validation. K-fold validation is used to separate the validation samples from the training samples, and the K indicates the number of times this is repeated for the generalization. The training and validation processes are performed in two approaches. The first approach uses default hyperparameters with segmented and non-segmented images. In the second approach, all hyperparameters of the models are optimized to train segmented datasets. The classification accuracy improved by 2.19% by utilizing segmentation and hyperparameter tuning further improved by 0.48%. The highest average classification accuracy of 97.92% is achieved using the Random Forest classifier to classify 40 classes of 10 different plant species. Accurate detection of plant disease leads to the sustained growth of plants throughout the growing span of the plants.**

Keywords—Leaf Diseases; Machine Learning; Support Vector Machine; K-Nearest Neighbor; Random Forest; Decision Tree.

I. INTRODUCTION

Agriculture serves as the primary source of income for over 58% of India's population [1]. As of April 2023, India is home to more than 96 million farmers. The agriculture sector contributes to over 18% of India's GDP [2]. Notable commercial crops in India include potatoes, tomatoes, mangoes, apples, grapes, peppers, soybeans, cotton, jute, tobacco, coffee, tea, and mustard [3].

Potatoes and tomatoes are the prominent crops grown globally. India contributes approximately 11% of the world's tomato production [4]. Exporting a significant portion of its tomatoes to countries such as Pakistan, Bangladesh, the Maldives, the United Arab Emirates (UAE), and the United States [5].

The yield of tomatoes or any crops depends on numerous factors, including soil fertility, environmental conditions, pests, and diseases. Diseases are significant contributors to crop losses, making early detection crucial. Detecting diseases through visual inspection can be challenging, especially when cultivating a variety of crops, even for experienced pathologists [6], [7]. In rural areas, open-eye inspection remains a common method for disease classification [8]. However, the reliance on visual methods can lead to delays in disease identification due to a shortage of experts in rural areas [9]. Disease detection using automated systems allows for early intervention, helping farmers to implement timely and targeted control measures. Early detection enables the implementation of effective strategies to contain or eradicate the disease before it spreads extensively.

Advancements in technology can transform the lives of farmers, providing them with a range of automated systems. Farmers can easily capture images of plant parts using standard digital cameras and upload them to disease detection systems, which provide information about treatment options and recommended pesticides [10]. Bacteria and fungi often cause plant diseases that can affect various plant parts, including leaves, stems, and roots [11], [12]. Since many disease symptoms manifest in the leaves, numerous researchers have focused on leaf disease detection using image processing and computer vision techniques.

Image processing and computer vision techniques are used to extract shape and texture features [13]–[21]. Among these methods, the combination of machine learning algorithms with image texture features is widely applied in plant disease detection. Notable machine learning algorithms viz., Random Forest Classifier (RFC), Logistic Regression Classifier (LRC), Support Vector Machine (SVM), Decision Tree Classifier (DTC), Linear Discriminant Analysis (LDA), and *K*-Nearest Neighbor (K-NN) [22]. Also, deep Convolutional Neural Networks (CNN) play a pivotal role in extracting complicated patterns to identify plant diseases.

In real natural environmental conditions, plant disease detection faces numerous challenges, including issues such as noise and lower contrast in lesion images, as well as small differences between the background and the lesion area [23]. To address these challenges, a novel technique has been proposed, which utilizes efficient image processing and machine learning classification techniques. In the proposed methodology, histogram equalization is used to enhance image quality, and the color denoising technique is used to eliminate noise. Subsequently, the leaf area is separated from the background using threshold masking [24]. Texture and color features of the image are extracted, including Hu moments, Haralick textures, and color histograms [25]. These features are then employed for classification through the application of machine learning algorithms.

The proposed work mainly highlights:

1. Importance of image pre-processing in plant disease detection to improve the classification accuracy.
2. Importance of choosing the optimum hyperparameters for machine learning algorithms for accurate disease classification.

Organization of the manuscript: section II discusses about the related works carried out by researchers globally, focusing on plant leaf disease detection. Section III explains the proposed methodology and the dastasets used in the research emphasizing more on the novelty of the proposed methodology. In Section IV experimental results are discussed, qualitatively and quantitatively, with the evaluation metrics. Finally, the conclusion and future works are given in section V.

II. RELATED WORK

Historically, significant researchers have focused on plant leaf disease detection using image processing techniques. The most recent techniques for plant leaf disease detection are reviewed in [26]–[30]. In recent years, there has been a growing emphasis on the use of machine learning for leaf disease detection.

M. R. Raigonda et al. [31], implemented a preprocessing and image segmentation approach to accurately identify leaf diseases in potato plants. Image sharpening through contrast enhancement is focused initially, and denoising techniques using median and Gaussian filters are applied at a later stage. For highlighting the region of interest, they employed k-means clustering as an image segmentation method. Color, shape, and texture features were subsequently extracted and fed into the classifier, enabling accurate disease detection.

Md. R. Mia et al. [32], employed an Artificial Neural Network (ANN) for mango leaf disease detection. They converted the original RGB leaf images to LAB color space and used k-means clustering for segmentation. The cluster representing the disease-affected area was used to extract 13 features, including contrast, energy, correlation, mean, moment, standard deviation, etc., These features were then used to train the machine learning system to recognize leaf disease.

In the study by S. S. Harakannanavar et al. [33], images were resized, and their quality was improved through histogram equalization. Lesion areas were partitioned using k-means clustering, and image boundaries were extracted using contour tracing. Informative features from image samples were extracted using Principle Component Analysis (PCA), Discrete Wavelet Transforms (DWT), and Grey Level Co-occurrence Matrix (GLCM). These features were employed to classify images using machine learning techniques such as KNN, SVM, and CNN.

M. Badiger et al. [34], developed a leaf disease classifier using SVM. The authors standardized the image sizes and applied k-means clustering for image segmentation. The SVM classified diseases using GLCM features. A. S. Deshapande et al. [35], implemented a machine-learning algorithm for disease classification in maize leaves. The authors utilized eighteen histogram features and eight Haar wavelet features with SVM and KNN classifiers. These classifiers achieved an accuracy of 85% for KNN and 88% for SVM. In another study [36], researchers focused on classifying diseased tobacco leaves with 120 leaf images. They implemented a CNN model and compared it with existing models, demonstrating an accuracy of 85.1% for their proposed model.

A. K. Singh et al. [37], introduced two methods for classifying plant leaf diseases using the PlantVillage dataset. In the first method, they employed CNN for image feature extraction, followed by classification using a Bayesian-optimized support vector machine. In the second method, features including the histogram of oriented gradients, color moments, and GLCM were extracted. Feature selection was performed using a binary particle swarm optimizer and the selected features were used for image classification with a random forest classifier.

P. Shetty et. al. [38], focused on classifying diseases in tomato plant leaves using image processing and machine learning classifiers. They aimed to classify four diseases: Leaf mold, Late blight, Bacterial spot, and Early blight using Linear discrimination analysis, Logistic regression, KNN, Decision tree, SVM, Naïve Bayes, and Random Forest. Experimental results revealed that the Random Forest classifier outperformed other classifiers in terms of classification accuracy.

Bijaya Hatuwal et al. [39], in their proposed method, multiple plant leaf diseases were classified using SVM, random forest, k-nearest neighbor, and CNN models. For CNN, images were directly used for training and classification, while the other three models utilized image features. Features like entropy, inverse difference moments, contrast, and correlation were extracted using Haralick textures. Among the models used, RFC, SVM, and KNN achieved classification accuracies of 87.43%, 78.61%, and 76.96%, respectively, while CNN achieved an accuracy of 97.89%.

Transfer learning with pre-trained deep convolutional neural networks was applied in [8]. Experiments were conducted on the popular PlantVillage dataset using DenseNet-121, VGG16, ResNet-50, and InceptionV4. The

results demonstrated that DenseNet-121 achieved the highest accuracy at 99.81%.

B. Vikki et al. [40], classified 38 classes of images from the PlantVillage database using transfer learning with AlexNet, InceptionV3, MobileNet, and a simple sequential model. Their experiments revealed a maximum classification accuracy of 97.52% for the MobileNet model.

Existing methods for plant leaf disease detection face challenges in providing accurate output. The leaf overlap, poor lighting, and randomness in air flow are the major issues while capturing the images in real-time environmental conditions, as the aforementioned conditions can obscure the lesion area, necessitating proper image pre-processing techniques. Additionally, existing algorithms tend to consume substantial time due to their complexity. To address these issues, in the proposed work, efficient image pre-processing techniques like image denoising, image enhancement, and segmentation are used along with machine learning algorithms to produce accurate results with reduced processing time and complexity.

III. Materials and Methods

The block diagram of the proposed leaf disease detection model is shown in Fig. 1. The proposed model is developed using digital image processing and machine learning approaches.

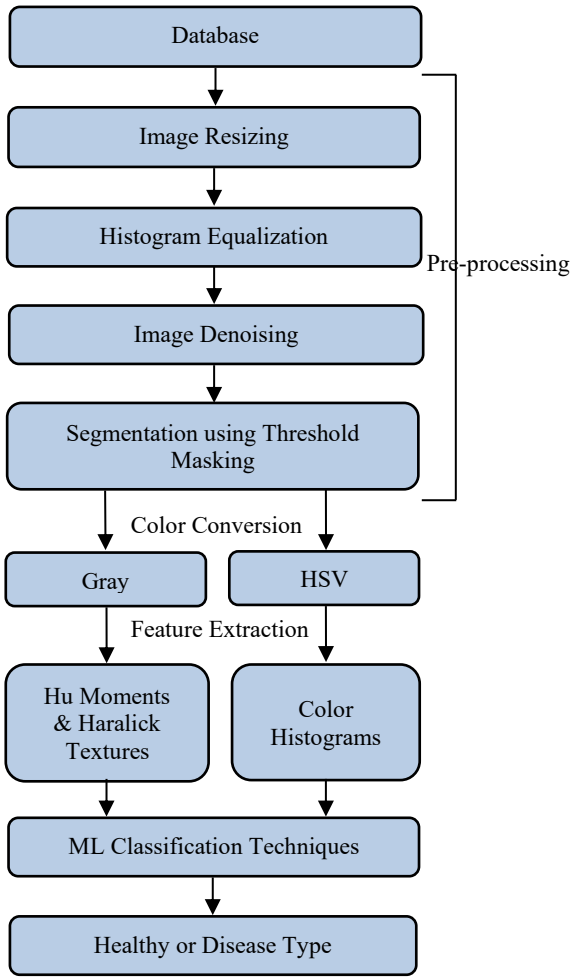


Fig. 1. The proposed leaf disease detection model

The experimentation is performed on the publicly available PlantVillage [41] and MangoLeafDB [42] datasets. Pre-processing techniques like image resizing, histogram equalization, gaussian denoising, and segmentation are applied to all the images in the database to make the feature extraction and classification process accurate. Image texture and color features are extracted from the pre-processed images and used to classify the images as healthy or the disease type using machine learning classification algorithms.

A. Dataset

The PlantVillage dataset comprises 38 classes of leaves from different plants, while the MangoLeafDB dataset consists of 7 unhealthy and 1 healthy class of mango leaves. A total of 41,546 images across 40 classes are chosen from the combined dataset, as shown in Table I. The choice of data split is predominantly influenced by the dataset size, and since the number of images is deemed sufficient for generalization purposes, a ratio of 80:20 has been selected to balance training and testing without encountering overfitting concerns.

TABLE I. Dataset Specifications

Plant Name	Disease Type	Dataset Size
Tomato	Healthy	1590
	Bacterial spot	2127
	Yellow leaf virus	1980
	Septoria leaf spot	1771
	Spider mites	1676
	Leaf mould	1904
	Early blight	1500
	Late blight	1754
	Target spot	1404
Apple	Healthy	1645
	Apple Scab	630
	Cedar apple rust	275
	Black rot	621
Corn	Healthy	1162
	Common rust	1192
	Northern leaf blight	985
	Gray leaf spot	513
Grape	Healthy	423
	Black rot	1180
	Black measles	1383
	Leaf blight	1076
Potato	Healthy	152
	Late blight	1000
	Early blight	1000
Cherry	Healthy	854
	Powdery Mildew	1052
Peach	Healthy	360
	Bacterial spot	2297
Pepper	Healthy	1478
	Bacterial spot	997
Strawberry	Healthy	456
	Leaf Scorch	1109
Mango	Healthy	500
	Gall midge	500
	Bacterial canker	500
	Powdery mildew	500
	Sooty mould	500
	Dieback	500
	Cutting weevil	500
	Anthracnose	500

B. Image Pre-processing

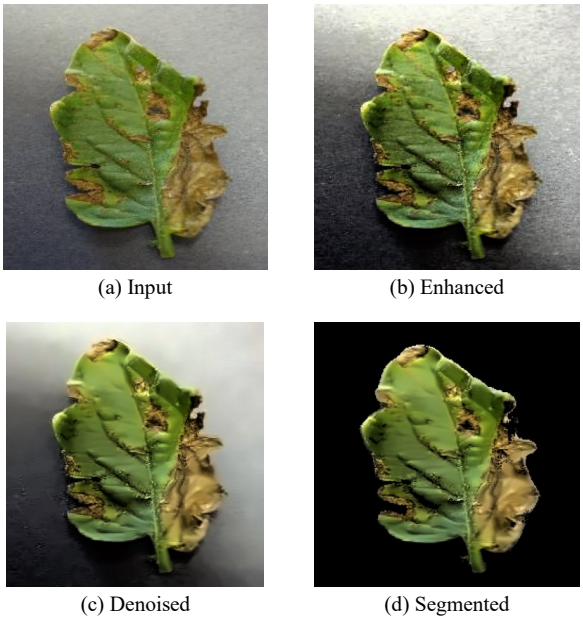
Image pre-processing is a crucial step in computer vision-based image processing systems, as its primary purpose is to enhance the accuracy of image classification [43].

Image Resize: In this study, all the images in the datasets are of size 256×256 pixels. This consistent sizing ensures that the results can be directly compared with existing models. In cases where images deviate from this specified size, they are resized to 256×256 to maintain uniformity and enable fair comparisons.

Image Enhancement: Subsequently, the adaptive histogram equalization [AHE] technique is applied to enhance image contrast [44]. AHE improves the visibility of details in both bright and dark regions by dividing the image into smaller regions and applying histogram equalization independently to each of these regions. This adaptability allows AHE to handle varying illumination conditions within an image.

Image Denoising: In the next step, a color image denoising technique is used to reduce image noise. In this process, RGB images are converted to the CIE LAB color space, and the L and AB components are denoised separately using a Gaussian filter before being converted back to the RGB color space [45].

Image Segmentation: Segmentation is employed to extract the leaf part from the image by suppressing the background pixels. In this study, a threshold-based segmentation method [46] is utilized, in which green and brown masks are individually created with their respective lower and upper threshold values. These threshold values are set based on the background in the image. The final mask is generated by combining the green and brown masks. A logical 'AND' operation is then applied between the input pre-processed images and the final mask to remove the background from the leaf. The output of the pre-processing steps is shown in Fig. 2(a-d).



(a) Input (b) Enhanced

(c) Denoised (d) Segmented

Fig. 2. (a-d) The output of the pre-processing steps for a tomato leaf from the PlantVillage dataset

C. Feature Extraction

Features are extracted from pre-processed images using Hu moments, Haralick textures, and color histogram feature descriptors. Hu moments provide an array of shape descriptors, calculated over a single channel of an image to precisely describe the leaf boundary. Haralick textures are used to differentiate texture features in leaf images. These texture features at pixel positions (I, J) are based on the frequency of pixel I occurring next to pixel J. Common texture features used in image classification problems include energy, entropy, homogeneity, autocorrelation, cross-correlation, dissimilarity, average, sum of squares, and variance. In leaf disease classification these features describe the shape and textures of disease affected area. To compute Hu moments and Haralick features, the segmented RGB images should first be converted to grayscale.

The detailed representation of colors in the image is obtained by calculating the color histogram. These color histograms help to differentiate the color changes in the disease affected region with respect to different disease classes. Since the HSV model closely aligns with the human eye's ability to perceive colors [47], input RGB images are converted to the HSV color space, and then the histogram is calculated over the HSV color space. This histogram plot provides information about the number of pixels that represent a given color range. All the features, including Hu moments, Haralick textures, and the color histogram are combined into a feature vector. This feature vector serves as input to the classifiers for recognizing the image class.

D. Classification

The extracted features are normalized and then used for training the classifier. Training is performed using machine learning algorithms such as logistic regression, linear discriminant analysis (LDA), K-nearest neighbor (K-NN), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and support vector machine (SVM).

The logistic regression model converts the continuous output of the linear regression function into categorical values by applying a sigmoid function. This sigmoid function maps any set of real-valued independent variables as input to a value ranging from 0 to 1 [48]. Additionally, extensions such as one-vs-rest enable logistic regression to handle multi-class classification problems. This model produces coefficients for each feature, indicating the strength and direction of their influence on the predicted outcome.

LDA operates by reducing the dimensionality of the data while enhancing class separation. This is achieved by identifying a set of linear discriminants that maximize the ratio of between-class variance to within-class variance. In simpler terms, LDA identifies the optimal directions in the feature space to effectively distinguish between various data classes [49].

KNN is a supervised learning algorithm that assumes samples of the same class have similarities in the feature space. To identify the class for any sample, this algorithm considers the k closest neighbors of the sample and then applies simple rules for classification [50]. KNN does not assume linear relationships between features. It can capture

complex decision boundaries, making it suitable for tasks where classes are not easily separable by linear boundaries.

The Decision Tree algorithm involves predefined target variables and constructs a tree-like structure consisting of multiple branches and leaf nodes. Each leaf node represents a specific decision, while each branch node signifies a choice among various alternatives. The decision tree outputs a Yes/No decision based on an input object that describes a set of properties [51]. Decision Trees can model non-linear relationships between features and the target variable. This flexibility allows them to capture complex patterns in image data, which might be challenging for linear models.

The Random Forest Classifier consists of a number of decision trees. The final output of this classifier depends on the outcomes of the individual decision trees. This algorithm is used for both regression and classification. It outputs the mean prediction in regression problems and the class in classification [52]. By combining the predictions of multiple trees, the model tends to generalize well to unseen data and reduces the risk of overfitting.

Support Vector Machine is indeed one of the most widely used machine learning algorithms, especially for classification tasks. SVM classifies a number of classes in one-dimensional feature space by drawing straight lines called hyperplanes between the classes [53]. This means that the features on one side of the line represent one class, while those on the other side represent another class. SVM has the capability to fit complex datasets and exhibits good generalization properties [54].

The proposed approach aims to improve leaf image analysis through the implementation of efficient noise-removal methods and background-removal techniques. The primary objective is to ensure image clarity by eliminating noise and removing the background without affecting the lesion area. The methodology underscores the utilization of uncomplicated machine learning algorithms to maintain a minimal model complexity while still achieving high classification accuracy.

IV. EXPERIMENTAL RESULTS

This section presents the simulation outcomes of the proposed model. During the experimentation process, image pre-processing, feature extraction, and image classification were performed using Jupyter Notebook with Python 3.11, along with libraries such as OpenCV, Keras, OS module, Globe module, and GridSearchCV.

The hardware setup consisted of an Intel(R) Core(TM) i5-4200U CPU running at 1.60GHz with a maximum turbo frequency of 2.30 GHz and 4GB of RAM. This configuration was utilized for training the classifiers and evaluating the performance of the proposed model.

The classifiers were trained using features obtained from each image. Six machine learning classifiers were trained and validated using the K-fold cross-validation technique. K-fold validation is the most popular technique for validating machine learning algorithms. In this technique, the available test data is split into K sample planes, and K iterations are performed for validation. In each iteration, one sample plane is used for validation and all other $K-1$ sample planes are used for training. This process continues K times and all the sample planes are used as test samples at least once. The final accuracy is calculated as the average accuracy of all K iterations.

The experimentation is conducted in two approaches using nine classes of tomato leaf images. In the first approach, the classifiers performance on segmented and non-segmented images is evaluated. In the second approach, the classifier performance is evaluated by choosing optimal hyperparameters. later the optimized model is used to classify all other images given in the dataset.

The model is evaluated using performance evaluation metrics like accuracy, precision, recall, and F1-score. Accuracy is a measure of overall correctness in a model, representing the ratio of correctly predicted instances to the total instances. Precision gauges the accuracy of positive predictions by calculating the ratio of true positives to the sum of true positives and false positives. Recall, or sensitivity, measures a model's ability to identify all relevant instances by calculating the ratio of true positives to the sum of true positives and false negatives. The F1 score, a harmonic mean of precision and recall, offers a balanced evaluation of a model's performance.

A. Performance of Classifiers on Segmented and Non-Segmented Images

Initially, the features of tomato leaf images are directly extracted from denoised images without segmentation. These features are then used to train classifiers employing the K-fold validation technique. The classification accuracy is observed to vary with the choice of the parameter K. Table II presents the classification results achieved without image segmentation. It is evident from the table that the Random Forest classifier outperforms other classifiers in terms of accuracy.

TABLE II. CLASSIFIERS' PERFORMANCE ON NON-SEGMENTED IMAGES

Classifier	Classification accuracy (%)			
	K=10	K=20	K=30	K=40
Logistic Regression	80.71	80.91	80.72	81.03
Linear Discriminant Analysis	78.74	79.78	79.77	79.77
K-Nearest Neighbor	84.71	84.81	84.85	84.98
Decision Tree Classifier	79.78	80.10	81.02	80.65
Random Forest Clasifier	94.20	94.23	94.43	94.17
Support Vector Machine	82.02	82.18	82.29	82.32

Subsequently, in the image pre-processing stage, image segmentation was carried out to eliminate the background from the leaf images. Features are then extracted from these segmented images. Table III displays the classifier performance on these segmented images for various values of K, enabling a comparison of their effectiveness in this segmented context.

The comparison between Table II and Table III reveals that the introduction of image segmentation has a beneficial impact on image classification performance. Furthermore, it's noteworthy that the Random Forest classifier consistently achieves the highest accuracy in both approaches across all values of K. The accuracy tends to increase initially with an

increase in the number of cross-validation folds (K) and reaches its peak at $K = 30$.

TABLE III. CLASSIFIERS' PERFORMANCE ON SEGMENTED IMAGES

Classifier	Classification accuracy (%)			
	***K*=10**	***K*=20**	***K*=30**	***K*=40**
Logistic Regression	90.36	90.39	90.45	90.45
Linear Discriminant Analysis	88.56	88.66	88.77	88.75
K-Nearest Neighbor	92.77	93.15	93.20	93.18
Decision Tree Classifier	84.43	85.75	85.77	85.76
Random Forest Classifier	96.35	96.54	96.62	96.62
Support Vector Machine	92.72	92.22	92.98	92.97

Specifically, the Random Forest classifier achieves a maximum accuracy of 94.43% without image segmentation and an improved accuracy of 96.62% with image segmentation both occurring at $K = 30$. Fig. 3 visually depicts the comparative accuracy of all the classifiers, showcasing the advantage of image segmentation in enhancing classification results, especially at $K = 30$.

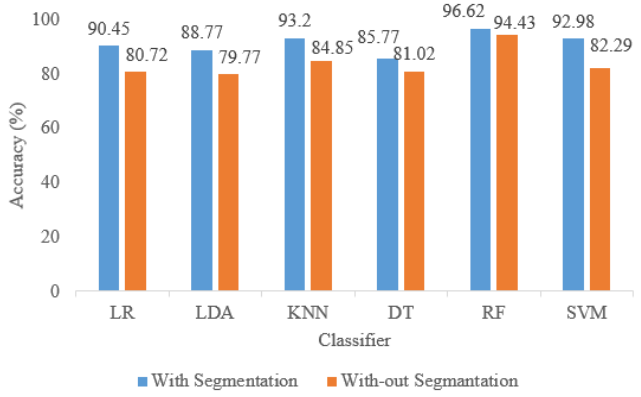


Fig. 3. Effect of image segmentation on classifier's performance

In the segmentation process the leaf background is removed, thereby eliminating the unwanted information in the image which leads to improved accuracy.

B. Performance of Classifiers on Tuning Hyperparameters

The classification accuracy of a classifier is influenced by variety of training parameters, including the number of trees, kernel size, penalty parameter, class weights etc.,.

To investigate the impact of the number of trees in the Random Forest classifier, we conducted an ablation study. The Random Forest classifier consists of multiple decision trees built on different subsets of the dataset. It aggregates predictions from each tree and makes a final prediction based on majority votes. Fig. 4 illustrates the performance of the Random Forest classifier on the tomato leaf dataset. This visualization allows us to assess how the number of trees affects the classifier accuracy.

In Fig 4, initially, the classification accuracy increases with an increase in the number of trees. The maximum test accuracy of 97.14% is recorded for 300 trees. Increasing the number of trees increases accuracy and increases computational complexity thereby, increasing the training time. Also, with a huge number of trees the model may overfit the dataset that was trained on and cause the reduction of model accuracy as observed after 300 trees.

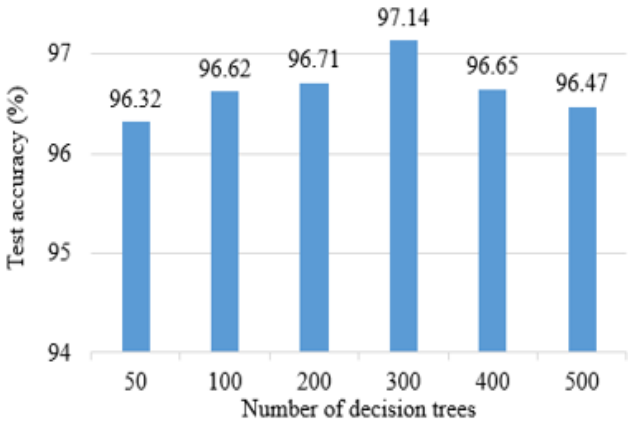


Fig. 4. Performance of RFC with the number of trees

Every machine learning algorithm is a mathematical model defined using the number of parameters that need to be trained from data. The kind of model parameters are called "hyperparameters" that cannot train directly from a regular training process. Each model and a dataset needs different set of hyperparameters. One way to determine the correct value for the hyperparameters is through multiple experiments, where each time pick a set of hyperparameters and train the model, this is called hyperparameter tuning. After multiple sets of experiments choose the best set of hyperparameters by evaluating the accuracy or loss. There are several automated methods available for this process: Bayesian optimization, Grid search, and Random search. These techniques train the model by choosing every possible set of hyperparameters and evaluating the model performance for each set. From this the best set of hyperparameters can be chosen, this process is called hyperparameter optimization.

In this research, the hyperparameter optimization is performed using GridSearchCV. With multiple values of each hyperparameter for all the models. Some of the optimized values are: "penalty"=none, "solver"=lbfgs and "max_iret"=200 for logistic regression, "solver"=svd for linear discrimination analysis, "n_neighbors"=1 for KNN, "criterion"=gini and "random_state"=500 for decision tree, "kernel"=rbf, "C"=150 and "degree"=1 for support vector machine, "n_estimators"= 500 and "bootstrap"=False for random forest classifier.

The performance metrics for all six classifiers on the tomato leaf dataset using hyperparameter optimization are shown in Table IV. Compared to Table III and Table IV the classification results are better with hyperparameter optimization as shown in Table IV. During the hyperparameter optimization, GridSearchCV evaluates the model performance with multiple combinations of parameters and automatically chooses the optimum parameter for better classification accuracy.

Table V shows the classification results for all the six classifiers on all 40 classes of images. It is noticed from Table V, that the random forest classifier outperformed in classifying all the ten types of plant classes with an average accuracy of 97.92% followed by the SVM classifier which has an average classification accuracy of 96.91%.

TABLE IV. CLASSIFIERS' PERFORMANCE ON TOMATO LEAF DATASET WITH HYPERPARAMETER OPTIMIZATION

Classifier	Performance Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Logistic Regression	94.72	95.00	94.86	95.00
Linear Discriminant Analysis	89.68	89.92	90.10	89.78
K-Nearest Neighbor	94.44	95.00	94.70	94.97
Decision Tree Classieier	85.86	86.00	86.05	85.00
Random Forest Classifier	98.02	98.00	98.71	98.23
Support Vector Machine	97.10	97.08	97.47	97.60

TABLE V. CLASSIFICATION RESULTS WITH OPTIMIZED HYPERPARAMETERS

Plant	Disease classes	Classification accuracy (%) for different classifiers					
		LRC	LDA	KNN	DTC	SVM	RFC
Apple	4	94.36	93.55	94.76	92.54	97.98	98.79
Cherry	2	99.08	99.39	99.08	97.56	99.08	99.69
Corn	4	86.17	84.84	90.69	90.56	90.29	94.02
Grape	4	94.01	92.95	92.82	89.62	95.88	96.14
Peach	2	99.49	98.23	98.48	98.48	98.73	99.49
Pepper	2	87.72	85.21	88.97	87.22	93.74	94.99
Potato	3	97.62	93.96	98.12	96.52	99.08	99.26
Strawberry	2	93.68	89.56	95.6	97.25	98.63	99.73
Tomato	9	94.72	89.68	94.44	85.86	97.10	98.02
Mango	8	97.12	93.62	97.25	93.25	98.63	99.00

The use of machine learning algorithms for plant leaf disease detection is the best idea for the early detection of diseases before they spread over the farm. From Table III and Table IV, it is clear that image background removal using the segmentation technique improves the accuracy of image classification and also, the random forest classifier performed best among the machine learning algorithms. The performance of the algorithms varied with the value of K in K-fold validation. The classification models have their own training parameters to be tuned for accurate training and classification. The results tabulated in Table V are evidence for having the best classification results with optimized hyperparameters.

The use of preprocessing techniques like image enhancement, denoising, and threshold-based segmentation helped to identify the disease parts easily and this led to improved classification accuracy in the proposed model compared to other state-of-art methods as given in Table VI.

The proposed model can identify the plant disease with less computational complexity and the accuracy of classifying the leaf disease is more compared to even CNN models that are computationally very heavy. The proposed model can be made as a mobile application, farmers can upload the images of the leaf. The proposed model can be made to provide recommendations on the health of the leaf and the possible pesticide to use to eradicate the disease thereby increasing the crop yield.

The proposed algorithm presents various benefits, especially in terms of its size and resource demands. The algorithm is less computationally intricate, enhancing its suitability for a broader array of hardware.

TABLE VI. COMPARISON OF THE PROPOSED METHOD WITH STATE OF ART METHODS

Reference	Dataset	Technique used	Accuracy (%)	Proposed accuracy with RF Classifier (%)
S. S. Harakannanavar et al. [33]	Tomato	K-means clustering, PCA, GLCM, DWT, SVM, KNN	97.00	98.02
M. Badiger et al. [34]	Tomato	K-means clustering, GLCM & SVM	96.00	98.02
H. Bijaya et al. [39]	Grape	Haralick Textures, SVM, KNN, and Random forest classifiers	87.43	96.14
B. Vikki et al. [40]	Tomato	InceptionV3, MobileNet, AlexNet, and Sequential CNN	97.52	98.02
S. Jana et al. [55]	Plant Village	Slice image fragmentation and DenseNet	97.30	97.71
J. Basavaiah et al. [56]	Tomato	Color histogram, Haralick texture, Hu moments, local binary patterns, RF	94.00	98.20
P. Bansal et al. [57]	Apple	Transfer learning of DenseNet121, EfficientNetB7	96.25	98.79
S. Nandhini et al. [58]	Tomato, Corn, Apple	CNN	96-98	94-99
R. Gajjar et al. [59]	PlantVillage	Deep CNN & Single shot detector	96.88	97.71
T. S. Xian al. [60]	Tomato	Haralick texture features & Extreme learning machine classifier	84.94	98.02

Nevertheless, it is crucial to acknowledge certain limitations in this study. The images in the database adhere to a standardized size and are captured under controlled conditions. Consequently, the effectiveness of the proposed

algorithm has not been evaluated on open-field datasets or real-time images.

To further enhance the model's performance in future work, potential avenues for improvement include the incorporation of fusion techniques for image feature extraction and the inclusion of diverse plant leaf datasets to increase the model's robustness and generalizability.

V. CONCLUSION

The objective of the proposed work is to classify leaf diseases in agricultural crops using efficient image processing and machine learning algorithms. The use of computer vision techniques helps to detect diseases in early stages with minimal time and avoid the spread of disease over the fields, this leads to improved crop yields. To achieve this, the proposed model employs image processing techniques, which encompass various steps such as image resizing, enhancement, denoising, and threshold-based segmentation. Moreover, the machine learning algorithm utilizes multiple feature descriptors including Haralick textures, Hu moments, and color histograms to capture both texture and color characteristics from leaf images for disease classification.

The use of image segmentation and hyperparameter optimization enhances the classification accuracy by 3.82% with a random forest classifier (RFC). It is observed that, the random forest classifier stands out as a particularly suitable choice for leaf disease classification. Superior classification accuracy is achieved by RFC compared to the other classifiers. RFC combines the predictions of multiple trees, tends to generalize well to unseen data, and reduces the risk of overfitting. Notably, the proposed model with RFC has classified the tomato leaf dataset with 98.02% accuracy, and the near competitor [40] has obtained 97.52%. Compared to this model the proposed model achieved approximately 0.5% improvement in accuracy and even outperforming other state-of-the-art methods. The database used in this research is captured under controlled conditions and also the effectiveness of the algorithm has not been evaluated on the open-field dataset. In the future, the use of image pre-processing techniques with Deep CNN models can improve the disease classification accuracy.

The proposed accurate model can be implemented as a standalone application to effectively classify the diseased leaf images in the early stages. This aids the farmers in proper crop management. Detection of diseases in early stages prevents the disease spreading over the crops and improves the crop yield this leads to global improvement in food production.

REFERENCES

[1] A. Gulati and R. Juneja, "Transforming Indian Agriculture," in *Indian Agriculture Towards 2030*, pp. 9–37, 2022, doi: 10.1007/978-981-19-0763-0_2.

[2] B. Sumanta and C. Timir Baran, "Contribution of the Agriculture sector to the economic growth of India with both being interdependent on each other," *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 10, pp. 103–110, 2021.

[3] K. Kutty, "Growth Trends of Commercial Crops Production, Area, and Yield in India: An Appraisement of the Structural Stability Regression Model," *Stud. Appl. Econ.*, vol. 41, no. 1, 2022, doi: 10.25115/sae.v41i1.8625.

[4] J. Zheng *et al.*, "Quality Improvement of Tomato Fruits by Preharvest Application of Chitosan Oligosaccharide," *Horticulturae*, vol. 9, no. 3, p. 300, 2023, doi: 10.3390/horticulturae9030300.

[5] A. Sharma, L. M. Kathuria, and T. Kaur, "Analyzing relative export competitiveness of Indian agricultural food products: a study of fresh and processed fruits and vegetables," *Compet. Rev. Int. Bus. J.*, vol. 33, no. 6, pp. 1090–1117, 2023, doi: 10.1108/CR-03-2022-0039.

[6] D. Vu, T. Nguyen, T. V. Nguyen, T. N. Nguyen, F. Massacci, and P. H. Phung, "HIT4Mal: Hybrid image transformation for malware classification," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 11, p. e3789, 2020, doi: 10.1002/ett.3789.

[7] E. David *et al.*, "Global Wheat Head Detection (GWHD) Dataset: A Large and Diverse Dataset of High-Resolution RGB-Labelled Images to Develop and Benchmark Wheat Head Detection Methods," *Plant Phenomics*, vol. 2020, pp. 1-12, 2020, doi: 10.34133/2020/3521852.

[8] J. Andrew, J. Eunice, D. E. Popescu, M. K. Chowdary, and J. Hemanth, "Deep Learning-Based Leaf Disease Detection in Crops Using Images for Agricultural Applications," *Agronomy*, vol. 12, no. 10, p. 2395, 2022, doi: 10.3390/agronomy12102395.

[9] X. Bai *et al.*, "Rice heading stage automatic observation by multi-classifier cascade based rice spike detection method," *Agric. For. Meteorol.*, vol. 259, pp. 260–270, 2018, doi: 10.1016/j.agrformet.2018.05.001.

[10] A. V. Panchal, S. C. Patel, K. Bagyalakshmi, P. Kumar, I. R. Khan, and M. Soni, "Image-based Plant Diseases Detection using Deep Learning," *Mater. Today Proc.*, vol. 80, pp. 3500–3506, 2023, doi: 10.1016/j.matpr.2021.07.281.

[11] M. Halder, A. Sarkar, and H. Bahar, "Plant Disease Detection By Image Processing: A Literature Review," *SDRP J. Food Sci. Technol.*, vol. 3, no. 6, pp. 534–538, 2018, doi: 10.25177/JFST.3.6.6.

[12] A. S. Zamani *et al.*, "Performance of Machine Learning and Image Processing in Plant Leaf Disease Detection," *J. Food Qual.*, vol. 2022, pp. 1–7, 2022, doi: 10.1155/2022/1598796.

[13] V. K. Vishnoi, K. Kumar, and B. Kumar, "Plant disease detection using computational intelligence and image processing," *J. Plant Dis. Prot.*, vol. 128, no. 1, pp. 19–53, 2021, doi: 10.1007/s41348-020-00368-0.

[14] C. Jackulin and S. Murugavalli, "A comprehensive review on detection of plant disease using machine learning and deep learning approaches," *Meas. Sens.*, vol. 24, p. 100441, 2022, doi: 10.1016/j.measen.2022.100441.

[15] R. Panchami and S. Vinod Chandra, "Rice Leaf Disease Detection and Diagnosis Using Convolution Neural Network," In *Review, preprint*, 2022. doi: 10.21203/rs.3.rs-1812823/v1.

[16] M. Shoaib *et al.*, "An advanced deep learning models-based plant disease detection: A review of recent research," *Front. Plant Sci.*, vol. 14, p. 1158933, 2023, doi: 10.3389/fpls.2023.1158933.

[17] L. Goyal, C. M. Sharma, A. Singh, and P. K. Singh, "Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture," *Inform. Med. Unlocked*, vol. 25, p. 100642, 2021, doi: 10.1016/j.imu.2021.100642.

[18] M. H. Saleem, S. Khanchi, J. Potgieter, and K. M. Arif, "Image-Based Plant Disease Identification by Deep Learning Meta-Architectures," *Plants*, vol. 9, no. 11, p. 1451, 2020, doi: 10.3390/plants9111451.

[19] M. A. Jasim and J. M. AL-Tuwaijari, "Plant Leaf Diseases Detection and Classification Using Image Processing and Deep Learning Techniques," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, pp. 259–265, 2020, doi: 10.1109/CSASE48920.2020.9142097.

[20] R. Deshpande and H. Patidar, "Detection of Plant Leaf Disease by Generative Adversarial and Deep Convolutional Neural Network," *J. Inst. Eng. India Ser. B*, vol. 104, no. 5, pp. 1043–1052, 2023, doi: 10.1007/s40031-023-00907-x.

[21] Y. Chen *et al.*, "DFCANet: A Novel Lightweight Convolutional Neural Network Model for Corn Disease Identification," *Agriculture*, vol. 12, no. 12, p. 2047, 2022, doi: 10.3390/agriculture12122047.

[22] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, 2021, doi: 10.1007/s42979-021-00592-x.

[23] J. Liu and X. Wang, "Plant diseases and pests detection based on deep learning: a review," *Plant Methods*, vol. 17, no. 1, p. 22, 2021, doi: 10.1186/s13007-021-00722-9.

[24] N. Kulkarni, "Color Thresholding Method for Image Segmentation of Natural Images," *Int. J. Image Graph. Signal Process.*, vol. 4, no. 1, pp. 28–34, 2012, doi: 10.5815/ijigsp.2012.01.04.

[25] A. O. Salau and S. Jain, "Feature Extraction: A Survey of the Types, Techniques, Applications," in *2019 International Conference on Signal Processing and Communication (ICSC)*, pp. 158–164, 2019, doi: 10.1109/ICSC45622.2019.8938371.

[26] M. S. P. Ngongoma, M. Kabeya, and K. Moloi, "A Review of Plant Disease Detection Systems for Farming Applications," *Appl. Sci.*, vol. 13, no. 10, p. 5982, 2023, doi: 10.3390/app13105982.

[27] S. M. Kiran and Dr. D. N. Chandrappa, "Current trends in plant disease detection," *Int. J. Sci. Technol. Res.*, vol. 8, no. 12, pp. 3055–3058, 2019.

[28] R. Patel *et al.*, "A review of recent advances in plant-pathogen detection systems," *Heliyon*, vol. 8, no. 12, p. e11855, 2022, doi: 10.1016/j.heliyon.2022.e11855.

[29] M. Nagaraju and P. Chawla, "Systematic review of deep learning techniques in plant disease detection," *Int. J. Syst. Assur. Eng. Manag.*, vol. 11, no. 3, pp. 547–560, 2020, doi: 10.1007/s13198-020-00972-1.

[30] M. H. Saleem, J. Potgieter, and K. M. Arif, "Plant Disease Detection and Classification by Deep Learning," *Plants*, vol. 8, no. 11, p. 468, 2019, doi: 10.3390/plants8110468.

[31] M. R. Raigonda and S. P. Terdal, "Design Engineering A Review on the Disease Identification on the Potato Foliar and Tuber," *Design Engineering*, pp. 13038-13056, 2022, doi: 10.13140/RG.2.2.13561.65126.

[32] Md. R. Mia, S. Roy, S. K. Das, and Md. A. Rahman, "Mango leaf disease recognition using neural network and support vector machine," *Iran J. Comput. Sci.*, vol. 3, no. 3, pp. 185–193, 2020, doi: 10.1007/s42044-020-00057-z.

[33] S. S. Harakannanavar, J. M. Rudagi, V. I. Puranikmath, A. Siddiqua, and R. Pramodhini, "Plant leaf disease detection using computer vision and machine learning algorithms," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 305–310, 2022, doi: 10.1016/j.gltp.2022.03.016.

[34] M. Badiger, V. Kumara, S. C. N. Shetty, and S. Poojary, "Leaf and skin disease detection using image processing," *Glob. Transit. Proc.*, vol. 3, no. 1, pp. 272–278, 2022, doi: 10.1016/j.gltp.2022.03.010.

[35] A. S. Deshapande, S. G. Giraddi, K. G. Karibasappa, and S. D. Desai, "Fungal Disease Detection in Maize Leaves Using Haar Wavelet Features," in *Information and Communication Technology for Intelligent Systems*, vol. 106, pp. 275–286, 2019, doi: 10.1007/978-981-13-1742-2_27.

[36] S. K. Dasari and V. Prasad, "A novel and proposed comprehensive methodology using deep convolutional neural networks for flue cured tobacco leaves classification," *Int. J. Inf. Technol.*, vol. 11, no. 1, pp. 107–117, 2019, doi: 10.1007/s41870-018-0174-4.

[37] A. K. Singh, S. Sreenivasu, Mahalaxmi, H. Sharma, D. D. Patil, and E. Asenso, "Hybrid Feature-Based Disease Detection in Plant Leaf Using Convolutional Neural Network, Bayesian Optimized SVM, and Random Forest Classifier," *J. Food Qual.*, vol. 2022, pp. 1–16, 2022, doi: 10.1155/2022/2845320.

[38] P. Shetty, A. Kumar, B. S. Rajesh, M. Balipa, G. Kanchan, and C. S. Kamath, "Tomato Leaf Disease Detection Using Multiple Classifier System," in *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 316–321, 2022, doi: 10.1109/AIDE57180.2022.10059795.

[39] H. Bijaya, S. Aman, and J. Basanta., "Plant Leaf Disease Recognition Using Random Forest, KNN, SVM and CNN.," *Polibits.*, vol. 62, pp. 13–19, 2021, doi: 10.17562/PB-62-2.

[40] B. Vikki and S. Sanjeev, "Plant Leaf Diseases Detection Using Deep Learning Algorithms," *Int. Conf. Mach. Learn. Image Process. Netw. Secur. Data Sci. Springer*, pp. 217–22, 2021.

[41] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," *Front. Plant Sci.*, vol. 7, p. 1419, 2016, doi: 10.3389/fpls.2016.01419.

[42] S. I. Ahmed *et al.*, "MangoLeafBD: A comprehensive image dataset to classify diseased and healthy mango leaves," *Data Brief*, vol. 47, p. 108941, 2023, doi: 10.1016/j.dib.2023.108941.

[43] K. Park, M. Chae, and J. H. Cho, "Image Pre-Processing Method of Machine Learning for Edge Detection with Image Signal Processor Enhancement," *Micromachines*, vol. 12, no. 1, p. 73, 2021, doi: 10.3390/mi12010073.

[44] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis. Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, 1987, doi: 10.1016/S0734-189X(87)80186-X.

[45] K.-L. Chung, W.-N. Yang, Y.-R. Lai, and L.-C. Lin, "Novel peer group filtering method based on the CIELab color space for impulse noise reduction," *Signal Image Video Process.*, vol. 8, no. 8, pp. 1691–1713, 2014, doi: 10.1007/s11760-012-0403-4.

[46] S. Jardim, J. António, and C. Mora, "Image thresholding approaches for medical image segmentation - short literature review," *Procedia Comput. Sci.*, vol. 219, pp. 1485–1492, 2023, doi: 10.1016/j.procs.2023.01.439.

[47] M. E. Moumene, K. Benkedadra, and F. Z. Berras, "Real Time Skin Color Detection Based on Adaptive HSV Thresholding," *Journal of Mobile Multimedia*, pp. 1617-1632, 2022, doi: 10.13052/jmm1550-4646.1867.

[48] M. Tadelo, H. Shifa, and A. Assefa, "Application of logistic regression model for predicting the association of climate change resilient cultural practices with early blight of tomato (*Alternaria solani*) epidemics in the East Shewa, Central Ethiopia," *J. Plant Interact.*, vol. 17, no. 1, pp. 43–49, 2022, doi: 10.1080/17429145.2021.2009581.

[49] S. Ali, M. Hassan, J. Y. Kim, M. I. Farid, M. Sanaullah, and H. Mufti, "FF-PCA-LDA: Intelligent Feature Fusion Based PCA-LDA Classification System for Plant Leaf Diseases," *Appl. Sci.*, vol. 12, no. 7, p. 3514, 2022, doi: 10.3390/app12073514.

[50] V. Gurunathan, T. Sathiya Priya, J. Dhanasekar, M. . *Niranjana*, and S. Suganya, "Plant Leaf Diseases Detection Using KNN Classifier," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 2157–2162, 2023, doi: 10.1109/ICACCS57279.2023.10112901.

[51] D. Nithyashree, B. Ramya, and R. Kumar, "Plant Disease Detection using Decision Tree Algorithm and Automated Disease Cure Ramya.B2, Rohith Kumar.V3, Birundha.R4," *Int. Res. J. Eng. Technol. IRJET*, vol. 7, no. 3, pp. 1834–1838, 2020.

[52] A. Wójtowicz, J. Piekarczyk, B. Czernecki, and H. Ratajkiewicz, "A random forest model for the classification of wheat and rye leaf rust symptoms based on pure spectra at leaf scale," *J. Photochem. Photobiol. B*, vol. 223, p. 112278, 2021, doi: 10.1016/j.jphotobiol.2021.112278.

[53] S. Iniyan, R. Jebakumar, P. Mangalraj, M. Mohit, and A. Nanda, "Plant Disease Identification and Detection Using Support Vector Machines and Artificial Neural Networks," in *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, vol. 1056, pp. 15–27, 2020, doi: 10.1007/978-981-15-0199-9_2.

[54] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognit.*, vol. 103, p. 107298, 2020, doi: 10.1016/j.patcog.2020.107298.

[55] S. Jana, S. D. Thilagavathy, and S. T. Shenbagavalli, "Plant Leaf Disease Prediction Using Deep Dense Net Slice Fragmentation and Segmentation Feature Selection Using Convolution Neural Network," *Nternational J. Intell. Syst. Appl. Eng.*, vol. 11, no. 6S, pp. 76–85, 2023.

[56] J. Basavaiah and A. Arlene Anthony, "Tomato Leaf Disease Classification using Multiple Feature Extraction Techniques," *Wirel. Pers. Commun.*, vol. 115, no. 1, pp. 633–651, 2020, doi: 10.1007/s11277-020-07590-x.

[57] P. Bansal, R. Kumar, and S. Kumar, "Disease Detection in Apple Leaves Using Deep Convolutional Neural Network," *Agriculture*, vol. 11, no. 7, p. 617, 2021, doi: 10.3390/agriculture11070617.

[58] S. Nandhini, R. Suganya, K. Nandhana, S. Varsha, S. Deivalakshmi, and S. K. Thangavel, "Automatic Detection of Leaf Disease Using CNN Algorithm," in *Machine Learning for Predictive Analysis*, vol. 141, pp. 237–244, 2021, doi: 10.1007/978-981-15-7106-0_24.

[59] R. Gajjar, N. Gajjar, V. J. Thakor, N. P. Patel, and S. Ruparelia, "Real-time detection and identification of plant leaf diseases using convolutional neural networks on an embedded platform," *Vis. Comput.*, vol. 38, no. 8, pp. 2923–2938, 2022, doi: 10.1007/s00371-021-02164-9.

[60] T. S. Xian and R. Ngadiran, "Plant Diseases Classification using Machine Learning," *J. Phys. Conf. Ser.*, vol. 1962, no. 1, p. 012024, 2021, doi: 10.1088/1742-6596/1962/1/012024.