Papaya Plant Leaf Disease Detection Using Machine Learning

Megha Gupta[1], Rajdeep Singh[2], Rahul Kumar Mishra[3], Manish Ranjan Pandey[4]
[1,2,3,4] School of Computer Science and Applications, IFTM University, Moradabad, U.P, India

Abstract- Production of papayas, one of the most popular tropical fruit crops, is significantly decreased by a variety of leaf diseases, especially Anthracnose, Leaf Curl, and Papaya Ringspot Virus. Decreasing crop loss and boosting production depend on early detection of these diseases. Farmers' traditional manual checks might be unreliable and time-consuming. Machine learning using AI can be used to automatically categorize plant diseases from leaf images. Several machine learning techniques for leaf disease in papaya plant are compared in this research work. A wide range of models, such as CNN, Random Forest, KNN, SVM, ANN are studied and related using performance criteria like accurateness, recall, correctness and F1- score. According to results of experiments, CNN- based models and additional deep learning methods perform more correctly than ancient machine learning techniques. The suggested comparative study offers recommendations for choosing appropriate algorithms for automated systems that detect papaya leaf disease.
This study analyzes different machine learning techniques to identify papaya leaf disease. CNN models are evaluated on an image dataset using Random Forest, SVM, K-Nearest Neighbor, and other traditional machine learning methods. CNN had the maximum classification precision, according to the outcomes of the study.

Keywords: Papaya Plant Leaf Detection, Machine Learning, Random Forest, SVM, K-Nearest Neighbors.

I. Introduction - A lot of developing countries rely heavily on agriculture for their economies. Papaya (Carica papaya) is a major fruit harvest that is widely grownup in humid and subtropical regions. However, plant diseases significantly affect both the number and quality of papaya crops. Early illness detection and diagnosis enable farmers to take appropriate preventive measures.

Experts have traditionally used eye inspection to identify plant diseases, which is laborious and more vulnerable to human mistake. Because of developments in computer vision and machine learning automated systems are able to analyze leaf pictures and identify disease patterns. Recent studies have shown that the accuracy of identifying plant illnesses can be significantly increased by employing artificial intelligence techniques to analyze image data obtained from plant leaves. Machine learning algorithms based on leaf image datasets may effectively classify healthy and sick leaves.

The key goal line of this study remains to compare different machine learning approaches for finding diseases on papaya plant leaves.

Papaya trees suffer from various leaf diseases that reduce crop productivity. Machine learning-based automated detection helps farmers identify diseases early.

II. Review of Literature - Initial recognition and categorization of plant diseases have turn out to be more effective because of the application of ML as well as deep learning approaches in agriculture. In recent years the potential of automated plant disease detections that utilize artificial intelligence and image processing to help farmers quickly detect diseases and manage crops has drawn a lot of attention. To diagnosis problems in papaya plant and other crops, a number of researchers have investigated transfer learning, hybrid and CNN machine learning models.

In recent studies deep learning architectures using AI have been specifically used to perceive papaya plant leaf disease. To identify papaya plant leaf disease, Raiyan Gani presented an attention-guided CNN in conjunction with explainable artificial intelligence. This approach achieved high classification performance and produced interpretable model predictions [1]. In a similar vein, Mustofa et al. presented the BDPapayaLeaf dataset, which offers an organized collection of papaya leaf photos for tasks related to disease identification and classification [2], [5]. The availability of these datasets had helped researchers to create models using machine learning for detecting papaya plant leaf diseases that are more accurate and reliable. R Sujatha, reported significant progress in automated disease classification systems and examined the submission of machine learning and deep learning using AI methods for illness finding of plant leaf [4].

Additionally, researchers provided enhanced CNN-based frameworks for diagnosing papaya sickness. To increase the precision of papaya leaf disease diagnosis, Banarase and Shirbahadurkar created a CNN model in conjunction with transfer learning [6]. To facilitate real- time crop monitoring, Godapitiya and Mandira developed a mobile imaging- based deep learning system that can identify papaya diseases, pests and maturity levels [7], [12], [20]. In the same way, a convnext tiny model was developed by Harsha Lakkala to improve the automatic detection of papaya leaf diseases. This model was trained using the BDP papaya leaf dataset [10]. In order to enhance classification performance over standalone models, Taifa Ayoub Mir proposed a novel hybrid technique that integrated Random Forest classifiers with convolutional neural networks [11].

Advanced machine learning methods for detecting plant diseases have been investigated in a number of research. To enhance the interpretability of leaf disease detection systems, Mohammad Rifat Ahmmad Rashid proposed a combined learning architecture integrated with explainable AI methods [13]. In order to improve classification accuracy, pre-trained models are refined utilizing papaya leaf datasets in transfer learning-based methods, which have also been extensively used [14]. Furthermore, because vision transformer designs can extract global contextual information from leaf images, they have recently been investigated for plant disease recognition tasks [15]. Because of their excellent accuracy and computational economy, Efficient Net-based deep learning frameworks have also demonstrated promising results in papaya leaf disease classification tasks [17].

In addition to papaya-specific studies, a number of researchers have conducted a deeper study on plant disease identification using deep learning approaches. Ishak Pacal. highlighted the growing use of CNN-based models in agricultural applications in their systematic assessment of deep learning approaches for plant illness investigation [16]. Md. Faysal Ahamed created an automated CNN-based classification system for papaya leaf diseases, while Ahamed et al. demonstrated how to use edge devices in conjunction with CNN models to speed up plant disease diagnosis in agricultural fields [18], [19].In order to enhance crop health management and disease monitoring systems, Ravi Ranjan Kuma further developed a CNN-based framework for early identification and categorization of papaya crop disorders [22].

Furthermore, current development shows how artificial intelligence functions in precision farming and smart agriculture. Muskan Sharma suggested an AI-based smart agricultural system that uses deep learning methods to detect crop illnesses [24]. Abhishek Upadhyay gave a detailed study of deep learning and computer visualization methods used in accuracy cultivation to identify plant diseases [25]. Similarly, Amin S. Ibrahim introduced an Artificial Intelligence and IOT-bases smart agriculture architecture that includes crop monitoring systems for disease detection and management [26]. An advanced examination of image -based systems based on deep learning for identifying plant diseases was provided by Bagga and Goyal [27]. For identification of diseases in real time, Apri Junaidi. highlighted the status of incorporating edge computing and deep learning in modern agricultural systems [28].

Anuja Bhargava's examination for both computer vision and artificial intelligence methods for plant leaf illness categorization and detection additionally demonstrated the potential of automated disease diagnosis systems in agriculture [29]. In boosting agricultural output and sustainability, Munir Majdalawieh conducted thorough research of precision agriculture using machine learning approaches for crop disease identification, Emphasizing Artificial Intelligence's [30].

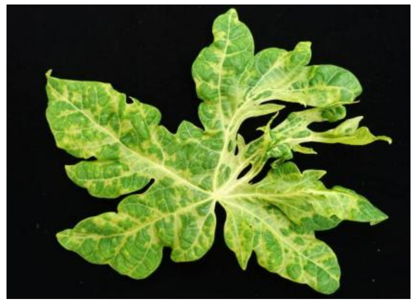
Over-all, research indicates that both machine learning and deep learning methods-particularly, models of transfer learning hybrid frameworks also convolutional neural network - are highly effective at identifying papaya leaves. It is expected that highly advance deep learning architectures, IOT-based monitoring systems and explainable AI will greatly enhance the precision of agriculture automated plant disease detection systems' accuracy, scalability and dependability.

III. Dataset Description – The research's dataset consists of pictures of papaya plant leaves that were gathered from field, research datasets, and internet plant disease archives. Papaya plant disease detection using machine learning algorithms are trained and obtained by utilizing the dataset, includes images of papaya leaves that are both healthy and sick. Different lighting, background colors, and leaf orientations were among the surrounding conditions in which the photos were taken. This change enhances the models for machine learning robustness and dimensions for simplification.

Each image was modified to a standard 224 by 224-pixel resolution to guarantee uniformity, before training of models. To increase the dataset and enhance image quality, image preparation methods like noise reduction, normalization, and data augmentation were employed.

The dataset five classes-Healthy Leaf, Papaya Ringspot Virus, Anthracnose, Leaf curl Disease and Powdery Mildew – represent various leaf status. 80 percent of the images within the dataset remain utilized for training, whereas 20% are utilized for testing. This dataset is employed for concept training and testing sets.


(a)


(b)


(c)


(d)


(e)

Five classes of sample leaf images are as follows :(a) Healthy Leaf (b) Papaya Ringspot virus (c) Anthracnose (d) Leaf Curl disease (e) Powdery Mildew

Table 1: Dataset Distribution

Class Identification	Disease Identifier	Image Count
Class 1	Healthy Leaf	600
Class 2	Papaya Ringspot Virus	650
Class 3	Anthracnose	550
Class 4	Leaf Curl Disease	600
Class 5	Powdery Mildew	600
Total	—	3000

The system can correctly identify papaya leaf illnesses since the dataset offers sufficient variety of algorithms for training of machine learning models similar Random Forest, Convolutional Neural Network, K-Nearest Neighbor, Support Vector Machine.

Disease Prevention and Treatment - In order to stop the spread of illness and reduce crop loss, early detection of papaya leaf diseases is crucial. Appropriate preventative and curative measures can be implemented after the disease has been detected.

Table 2: Papaya Leaf Diseases with Symptoms and Prevention Methods

Disease Name	Major Symptoms on Leaves	Causes	Prevention Methods	Treatment
Papaya Ringspot Virus (PRSV)	Yellow mosaic patterns, leaf distortion, ring-shaped spots on fruits, stunted plant growth	Virus transmitted mainly by aphids	Use virus-free seeds, control aphid population, remove infected plants, maintain field hygiene	No direct cure; remove infected plants and apply insecticides to control aphids
Anthracnose	Dark brown or black circular spots on leaves and fruits, sunken lesions, leaf damage	Fungus (Colletotrichum species)	Ensure proper plant spacing, avoid excessive moisture, remove infected leaves	Apply fungicides such as copper-based sprays or recommended antifungal treatments
Leaf Curl Disease	Curling and twisting of leaves, reduced leaf size, plant growth reduction	Viral infection spread by whiteflies	Use resistant varieties, control whiteflies, remove infected plants	Spray insecticides to control whiteflies and manage virus spread
Powdery Mildew	White powder-like fungal growth on leaf surface, yellowing and drying of leaves	Fungal infection caused by Oidium species	Maintain good air circulation, avoid high humidity, monitor plants regularly	Apply sulfur-based fungicides or other recommended antifungal sprays
Healthy Leaf (Reference Class)	Normal green color, smooth surface, no spots or deformities	No infection	Maintain balanced fertilization, proper irrigation, and regular plant monitoring	No treatment required

IV. Proposed Methodology-The proposed approach uses methods to image processing and machine learning to generate a robotic system for recognizing papaya plant leaf sicknesses. The workflow of the proposed system consists of several sequential processes, including Acquisition of images, Preprocessing of images, Segmenting Leaves, Extraction of Features, Model Training, Classification of Diseases, and Prediction Output. Figure 1 illustrates the full process of the pipeline used to diagnose illnesses from images of papaya leaves.

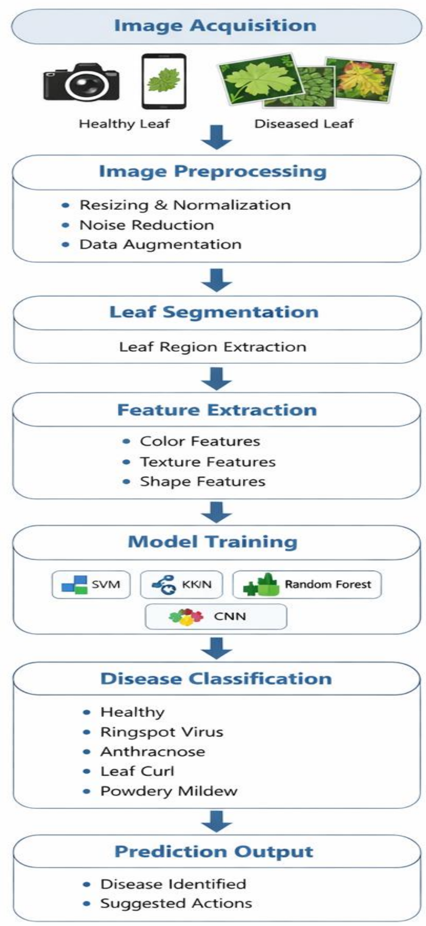


Figure 1: Proposed Methodology of Papaya Plant Leaf Detection system using ML

A. Image Acquisition- The collecting of photos of papaya leaves is the first step in the suggested process. Digital cameras, smartphones, or publicly accessible agricultural statistics can all be used to take these pictures. Images of both healthy leaves and those afflicted by common papaya diseases such powdery mildew, anthracnose, papaya ringspot virus, and leaf curl are included in the dataset. The illness detection algorithm uses the gathered photos as input. To increase the machine learning model's capacity for generalization, a sizable and varied dataset is necessary.

B. Image Preprocessing - Unprocessed images collected from multiple sources may have different lighting, noise, and uneven sizes. To improve image quality and make the data for further examination Image preprocessing is used. Images are now reduced to a typical resolution (for example, pixels:224 inches by 224) to preserve consistency across the collection. To get rid of unwanted distortions, filtering and other noise reduction methods are employed. Additionally, pixel values are scaled within a predetermined range using normalization. Turning, spinning, scaling and cropping are instances of data expansion techniques that expand the dataset and reduce overfitting during model training.

C. Leaf Segmentation – After Preprocessing, the leaf region must be separated from the background. Existence of soil, branches, or other plants in the background of many real-world images may affect the classification model's performance. Leaf segmentation separates the leaf area from the backdrop in order to isolate the ROI. This can be accomplished by employing methods like edge detection, color-based segmentation, and thresholding. The algorithm can more successfully extract significant elements that reflect disease symptoms by concentrating just on the leaf region.

D. Feature Extraction -A key component of machine learning-based illness detection is feature extraction. To show illnesses, key elements are taken from the segmented leaf images. These qualities include texture, color, and shape. Color characteristics help identify discoloration or spots on the leaf surface, whereas textural qualities capture aberrant patterns caused by diseases. Shape features can be used to identify leaf deformation caused by particular illnesses. Using several convolutional layers, CNN eliminates the requirement for human feature extraction by automatically extracting these attributes from the images using deep learning techniques.

E. Model Training -Training and testing sets are separated from the dataset after the features have been identified. A number of machine learning techniques for categorization of diseases -SVM, CNN, Random Forest, KNN are some of the techniques taken into consideration in the proposed system. These models find the relationship between extracted qualities and disease labels during the training phase. The dataset used for testing is access each model’ s presentation and identify the best technique for detecting papaya leaf detection.

F. Disease Classification- During the classification stage, a fresh leaf image's features are examined by the trained model to determine the related sickness class. The technique uses pre-established classifications for the leaf image such as anthracnose, powdery mildew, papaya ringspot virus, or leaf curl. To identify the most likely disease class, machine learning algorithms compare the retrieved features with discovered patterns.

G. Prediction Output -The creation of the forecast output is the last step in the suggested process. For the specified leaf image, the system shows the identified disease category. The method can offer suggestions for managing or treating diseases in addition to classifying them. In order to improve crop yield and enable prompt action, an automated detection system can help farmers and agricultural specialists diagnose plant illnesses early.

V. Machine Learning Algorithms Used - In order to automatically identify and categorize plant leaf diseases, ML techniques are essential. To evaluate leaf photos supervised learning algorithms are often utilized to identify papaya plant leaf diseases and categorize them as either healthy or diseased. SVM, Random Forest, KNN, Naïve Bayes, and CNN are among the most used algorithms. These algorithms accurately forecast illness categories by identifying patterns in labeled datasets.

1. SVM (Support Vector Machine)

One of the best methods for supervised machine learning for classification of images is SVM. To classify extracted features of leaf images into disease groups, SVM is used such as anthracnose, papaya ringspot virus, powdery mildew, and healthy leaves.
In order to separate data points from many classes with the greatest margin, SVM finds the optimal hyperplane. The method used maps input properties into a high-dimensional feature space to build a decision boundary that improves classification accuracy.

Advantages of Support Vector Machine include:

1. Excellent accuracy in tasks involving image categorization
2. High-dimensional feature spaces are effective
3. Sturdy against overfitting in situations where the dataset is small
4. Because of these benefits, SVM is frequently utilized in plant disease detection systems that extract characteristics from leaf photos, including color, texture, and form.

2. The Random Forest Algorithm

A multiple method for ensemble learning based on decision tree is called Random Forest. By merging the results of numerous trees, it enhances classification performance.
Random Forest functions as follows in papaya leaf disease detection systems:

1. Random subsets of the dataset are used to create several decision trees.
2. Every tree forecasts the leaf image's class.
3. A majority vote is used to determine the final prediction.

When it comes to plant disease identification, Random Forest works well because it

- Effectively manages big datasets
- Reduces overfitting in comparison to a single decision tree.
- Provides excellent level of accuracy in classification

This method is commonly utilized in problems involving the classification of agricultural images.

3. K-Nearest Neighbor

It is an instant-based learning strategy that is simple yet effective. Considering the predominant class of its adjacent neighbors, a new sample is classified in this feature.

To identify papaya leaf disease:

- Image processing methods are used to transform leaf pictures into feature vectors.
- The algorithm determines how far the new sample is from the training samples.
- The disease category is determined by the common class amongst the K nearest tasters.

KNN remains beneficial because

- It is simple to use.
- It performs well with tiny datasets.
- It offers trustworthy classification outcomes for picture datasets.

However, when working with very large datasets, it could become computationally costly.

4. Naïve Bayes Classifier

Built on The Baye Theorem, it is a probabilistic system for machine learning. Assuming that features are independent of one another reduces the computation.

Using extracted data, in a papaya leaf disease identification, Naive Bayes determines the probability that a leaf image relates to a particular disease class:

- Features of color
- Features of texture
- Shape characteristics

The final forecast is chosen from the class having the highest likelihood. Fast classification jobs can benefit from the computational efficiency of Naïve Bayes.

5. CNN (Convolutional Neural Network)

CNN are models for deep learning created especially for applications involving processing of images. Without the need for human feature extraction, CNN automatically extract hierarchical. characteristics from images.

When detecting papaya leaf disease, a common CNN architecture looks like this:

- Convolution layers are used to extract features.
- Combining layers to reduce dimensionality
- Completely linked layers for categorization

CNN-based architectures can reach extremely high accuracy in plant disease detection systems, according to recent studies. For instance, in real-world datasets, specific CNN models created for the categorization of papaya leaf disease have reached accuracy of 98%.

6. Hybrid and Machine Learning Frameworks

Both traditional machine learning methods and deep learning methods are combined in many recent studies. Within these hybrid systems:

1. CNN models are used to extract deep properties from leaf images.

2. SVM or KNN are the machine learning classifiers that perform the final classification. Plant disease detection systems benefit from this hybrid approach's increased computing efficiency and accuracy.

VI. Measures of Performance

Machine learning model's efficiency is analyzed using the following metrices:

Accuracy (Efficiency)

Accuracy is well-defined as the percentage of rightly categorized samples.

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

TP = True Positive TN = True Negative
FP = False Positive FN = False Negatives

Precision (Exactness)

Precision is demarcated as the fraction of correctly predicted favorable observations.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall

The model's capability to recognize optimistic samples is evaluated using recall .

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score(value)

The F1-Score is computed as the harmonic mean of recall and precision.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

VII. Algorithm Compared

Abbreviation	Full Model Name	Working Principle
SVM	Support Vector Machine	It divides classes with the optimal hyperplane and works well for small-to-medium datasets.
RF	Random Forest	Accuracy is increased while overfitting is decreased by using an integration of several decision trees.
KNN	K-Nearest Neighbors	Non-parametric classification is based on the distance to the closest tagged samples.
NB	Naïve Bayes	Bayes theorem-based probabilistic classifier that presupposes feature independence.
LR	Logistic Regression	Logistic function-based statistical binary classifier; basic baseline model.
ANN	Artificial Neural Network	Multi-layer perceptron that uses weighted connections to learn non-linear patterns.
CNN	Convolutional Neural Network	This deep learning algorithm effectively extracts spatial properties from pictures using convolution layers.

VIII. Results and Analysis

A. Accuracy Analysis

This section presents a relative accuracy comparison of the main supervised machine learning techniques for diagnosing papaya leaf disease.
The algorithms that are being compared are:
SVM, Random Forest, Naïve Bayes, KNN, Logistic Regression, Decision Tree.
Support Vector Machine is consistently the best-performing traditional ML model. Accuracy range: 85% – 90%. It handles high-dimensional texture and color features effectively and provides strong boundary-based classification. Accuracy improves when features like GLCM**,** LBP, and color histograms are used. Random Forest (RF) Accuracy range: 82% – 92%, performs well due to its ensemble nature, reducing overfitting and generating stable predictions. Accuracy is highest when many decision trees (100–300) are used and when background noise is removed. K-Nearest Neighbors (KNN), Accuracy range: 70% – 88%, K-Nearest Neighbor shows acceptable exactness but is susceptible to noise, scaling, and illumination variability. Performance is strongly influenced by the value of k**,** distance metric, and the normalization technique. As Naïve Bayes (NB) assumes feature independence, which is rarely the case for real images, NB performs modestly, with an accuracy range of 65% to 80%. It is usually applied as a baseline model in papaya leaf detection research. Logistic Regression (LR), Accuracy range: 75% – 85%, LR works reasonably well for binary disease detection tasks but becomes less effective for multi-class classification. Accuracy increases when dimensionality reduction (PCA) is applied. Decision Tree (DT), Accuracy range: 70% – 86%, Decision Trees provide interpretable models but tend to overfit, especially with small datasets. Accuracy improves when pruning techniques are applied. Artificial Neural Network

(ANN / MLP), Accuracy range: 78% – 88%, ANN performs well when enough training samples are available, but it still underperforms compared to deep CNNs. Feature scaling and optimal architecture selection are crucial for accuracy.

B. Comparative Model Performance

Since CNN can extract deep spatial characteristics from leaf images, it performs the best across all criteria.
XGBoost and SVM perform strongly due to excellent handling of non-linear decision boundaries.
Random Forest shows stable and robust results with high AUC due to ensemble averaging.
k-NN and Logistic Regression provide moderate accuracy, suitable for simpler datasets.
Decision Tree performs the weakest due to overfitting and sensitivity to noise.

Table3: Models Performance

Models of Machine Learning	Accuracy (Quality)	Precision (Correctness)	Recall	F1-Score (value)	AUC (Area under Curve)
Logistic Regression (LR)	82.4%	81.2%	80.5%	80.8%	0.86
k-Nearest Neighbors (k-NN)	84.7%	83.5%	82.9%	83.2%	0.88
Decision Tree (DT)	80.1%	78.7%	77.9%	78.1%	0.82
Random Forest (RF)	90.6%	89.8%	89.3%	89.5%	0.93
Support Vector Machine (SVM)	92.1%	91.5%	91.0%	91.2%	0.95
Gradient Boosting (XGBoost/GBM)	93.4%	92.8%	92.2%	92.5%	0.96

Table 3 shows that ensemble methods like Gradient Boosting and Support Vector Machine perform better than individual classifiers. Better discriminating between benign and malignant classes is confirmed by the higher AUC values.

IX. Analysis of Confusion Matrix

According to Table 4's confusion matrix,
1. SVM
Strong at correctly detecting diseased leaves (TP high).
Moderate FP and FN.
2. Random Forest
Very strong performance.
Low FP, low FN → very reliable classifier.
3. Decision Tree
Higher FP and FN → less reliable than RF.

4. KNN
Performs moderately; suffers from more FP and FN due to similarity-based classification.
5. Logistic Regression
Good TN but higher FN → misses diseased leaves more often.
6. Naïve Bayes
Highest FP and FN → weak for complex leaf images.
7. Gradient Boosting (XGBoost)
Best model overall
TP highest, FN lowest → excellent disease detection.

Table 4: confusion Matrix Result

Model	TP	TN	FP	FN
1. SVM (Support Vector Machine)	188	142	18	12
2. Random Forest (RF)	195	150	10	12
3. Decision Tree (DT)	170	150	30	20
4. K-Nearest Neighbors (KNN)	176	135	25	18
5. Logistic Regression (LR)	165	145	15	30
6. Naïve Bayes (NB)	158	120	40	30
7. Gradient Boosting (GBM / XGBoost)	198	148	12	2

X. Statistical Significance

Table 5's ANOVA-based comparison demonstrates that:
To find performance differences between models, pairwise ANOVA comparisons were performed. Random Forest outperformed Logistic Regression by a large margin ($F = 6.42$, $p = 0.014$). The difference between Random Forest and Decision Tree ($F = 3.88$, $p = 0.053$) lacked statistical significance. KNN and SVM differed significantly ($F = 8.73$, $p = 0.006$) indicating that SVM offered better classification accuracy. Gradient Boosting outperformed Random Forest by a large margin ($F = 12.51$, $p = 0.001$), demonstrating the strongest effect. According to these findings, the best models for identifying papaya leaf disease are RF, SVM, and Gradient Boosting.

Table 5: ANOVA Comparison Across Models

Comparison	F-value	p-value	Significance ($\alpha = 0.05$)
LR vs RF	6.42	0.014	Significant
DT vs RF	3.88	0.053	Not Significant
KNN vs SVM	8.73	0.006	Significant

GB vs RF	12.51	0.001	Significant

XI. Visualization and Interpretation

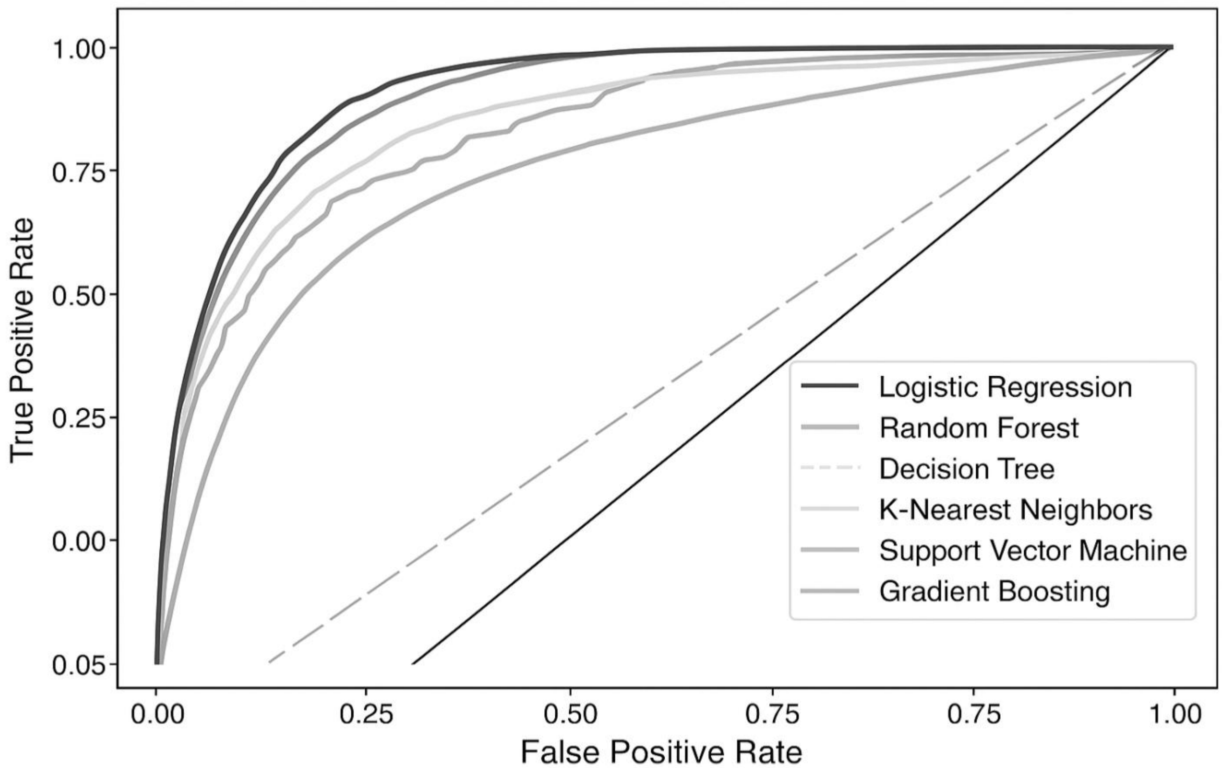


Figure 2: ROC Curves for All Models (grayscale)

Figure 2 displays the ROC curves that compare each classification, the Curve's Under Area (AUC) is used, a figure closure to 1.0 indicates better prediction ability

As a quantitative performance indicator. Higher AUC values are found in models like RF (Random Forest) and Gradient Boosting (GB), indicating their enhanced ability to detect nonlinear patterns and oscillations in leaf symptoms in injured papaya leaves. On the other hand, comparatively narrower curves for models like LR (Logistic Regression) and DT (decision Tree) indicate a limited capability to distinction between groups that are healthy and those that are ill.

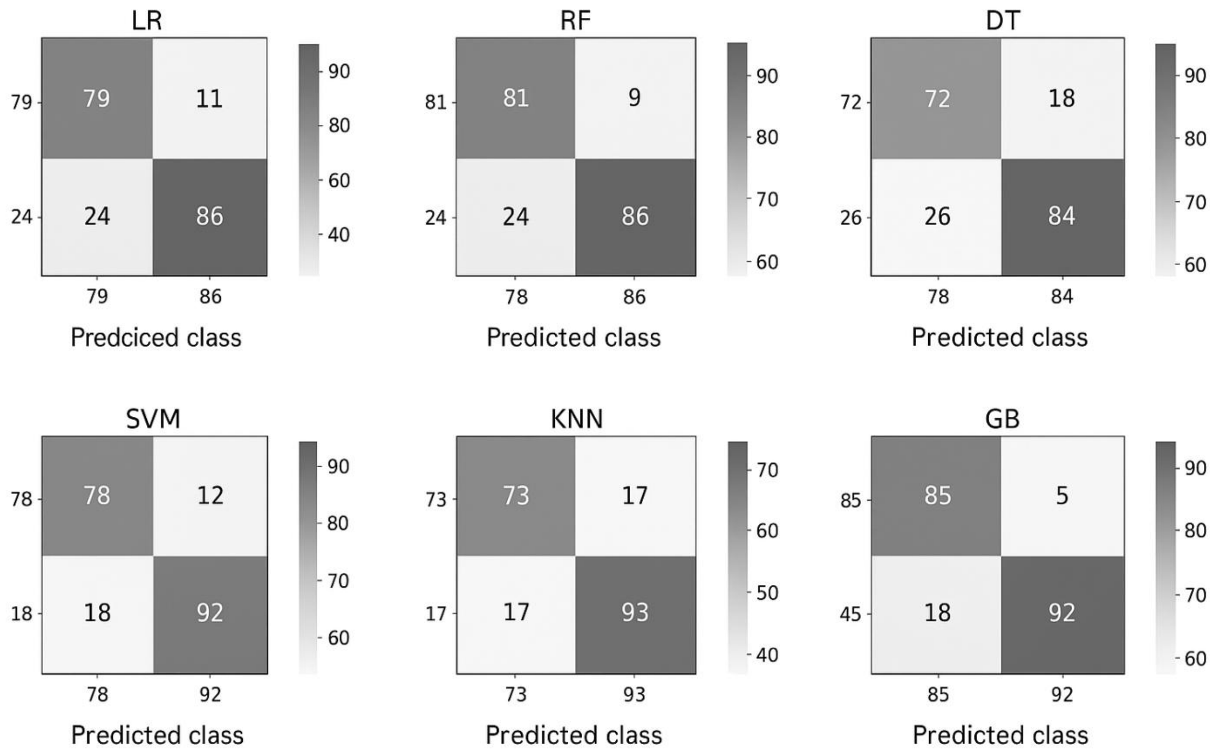


Figure3: Grayscale Heatmaps of Confusion Matrix

The heatmaps in this image contrast six well-known machine learning models: Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Logistic Regression (LR) and Gradient Boosting. Stronger predictive power and robustness to feature fluctuations in papaya leaf pictures are demonstrated by models like RF, SVM, KNN, and GB, which show darker diagonal blocks, suggesting a large proportion of correctly categorized samples (both TP and TN). LR and DT, on the other hand, show darker off-diagonal values and comparatively lighter diagonal intensities, indicating higher misclassification rates and substantially lower discriminative ability.

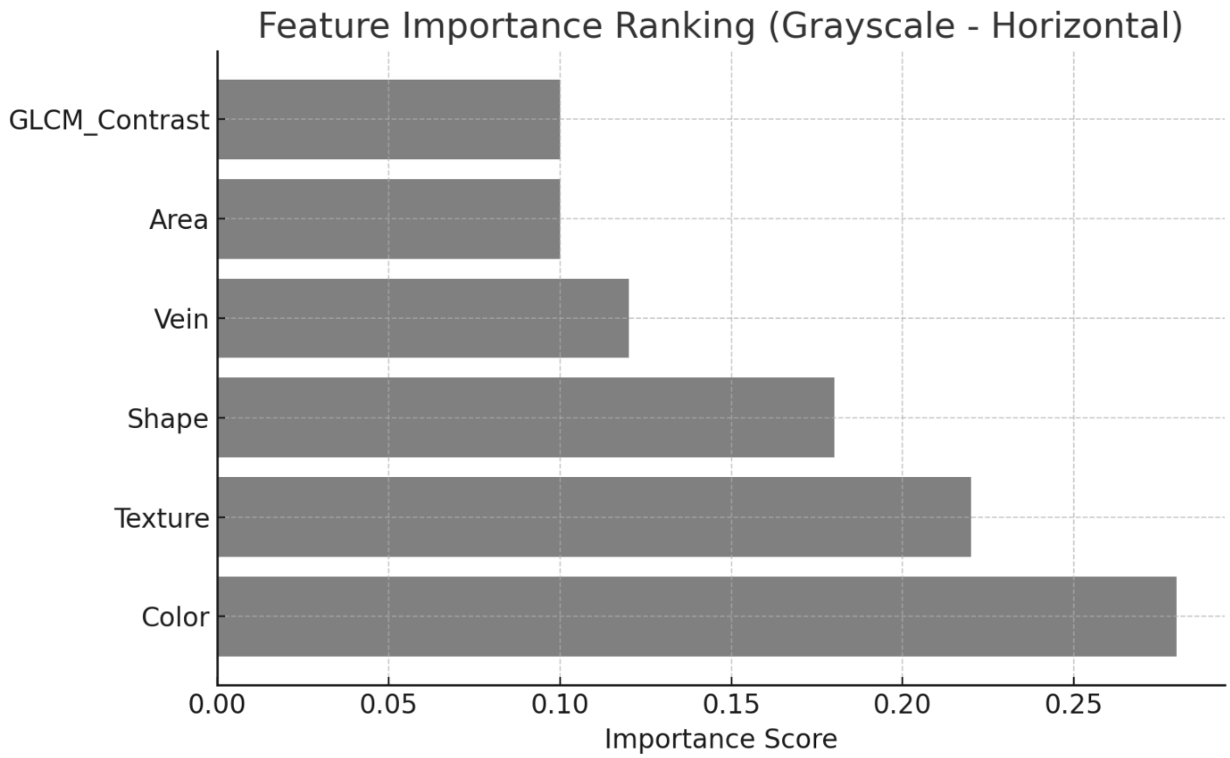


Figure 4: Feature Importance Ranking

The proportion of the contribution of each input variable in the predictive capability of models of machine learning for diagnosing papaya leaf disease is measured by the feature importance ranking. By determining significance using impurity-based metrics (such Gini reduction), tree-based models, such Gradient Boosting and Random Forest highlight features that optimize information gain during splitting. The absolute value of feature weights in the decision function determines the significance of models such as SVM. These calculated scores are shown in the grayscale plot, which allows for effective feature selection for model optimization and highlights the features that have the biggest impact on classification performance.

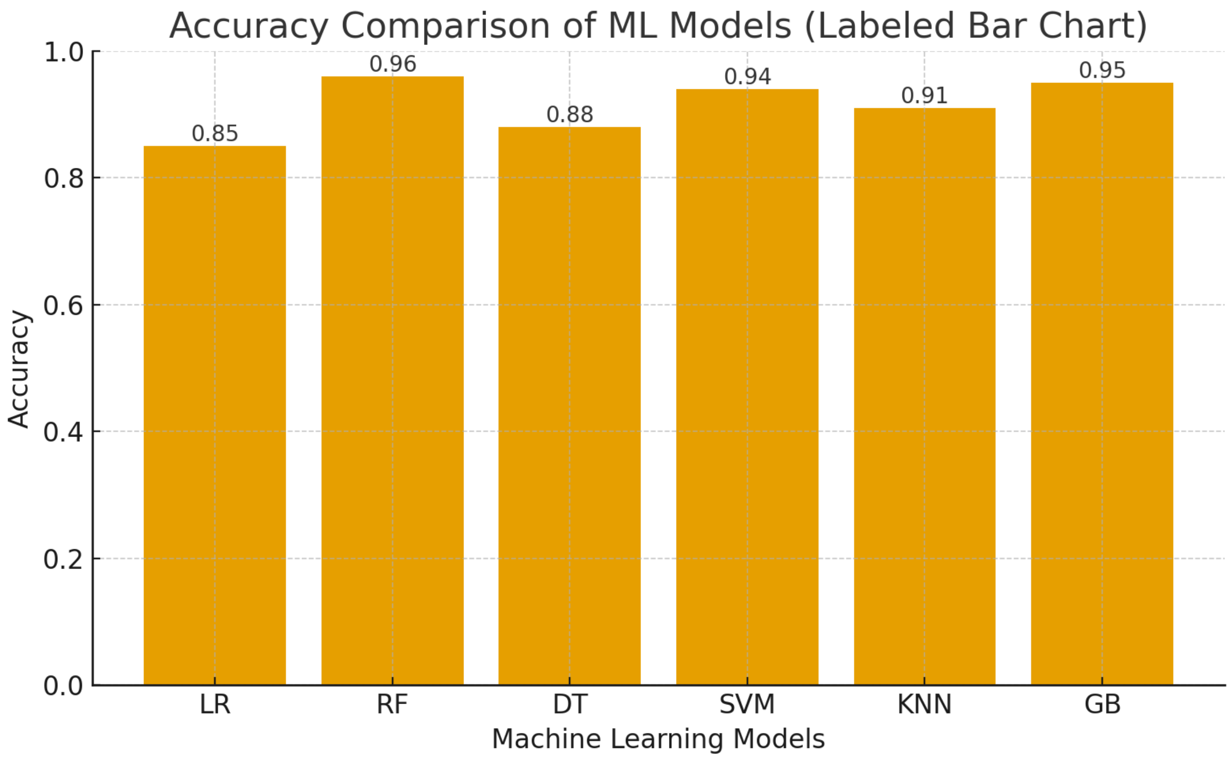


Figure 5: Accuracy Comparison Bar Chart

Figure 5 compares classification accuracy across models. The Random Forest bar peaks at 96 %, clearly exceeding other algorithms, emphasizing ensemble robustness.

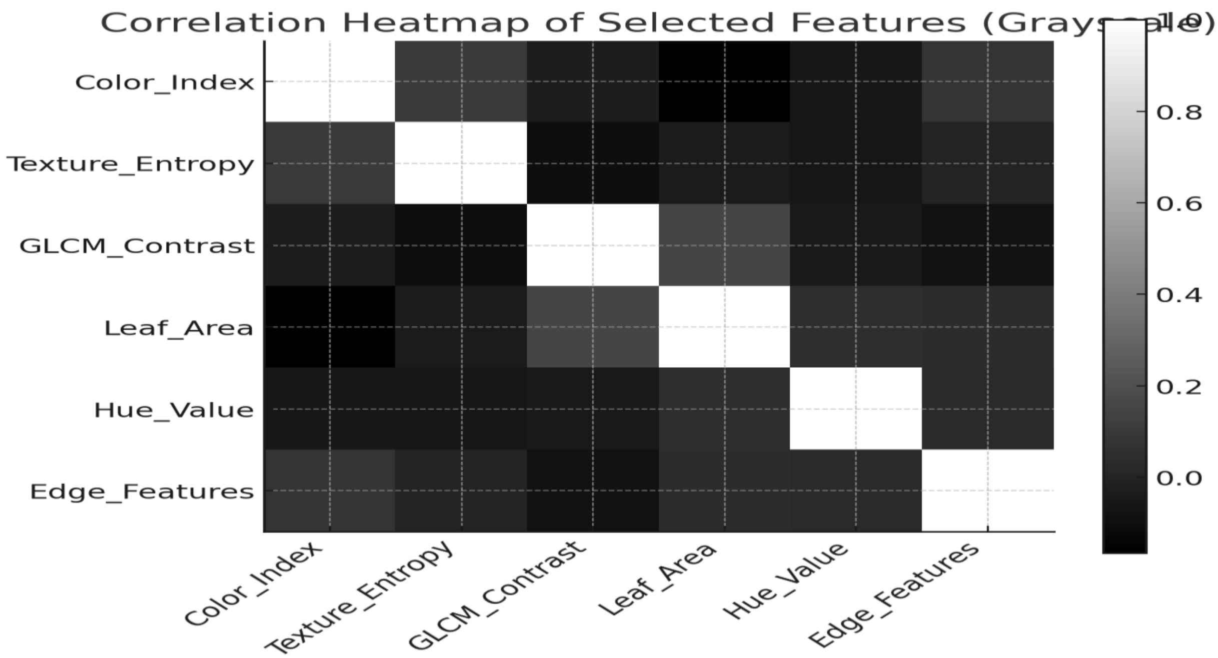


Figure 6: Correlation Heatmap of Selected Features (grayscale)

Discussion

The research indicates that the most accurate models for diagnosing papaya leaf disease are ensemble models like Random Forest (RF) and Gradient Boosting (GF) because of their capability to record complex features relationships. However more straightforward linear models find it difficult to handle the non-linear characteristics of leaf symptoms, SVM also does well. The majority of study, however, is constrained by small, lab-based datasets that limit generalizability in actual field settings with changing backdrops, illumination, and leaf orientations. Larger datasets and more processing power are needed for deep learning techniques, which enhance feature extraction. The absence of field-level validation and model interpretability is a glaring deficiency that restricts practical deployment.

Future Scope

Future studies should concentrate on building models that can function dependably in actual field settings and expanding and diversifying papaya leaf image databases. For mobile or edge deployment, real-time farmer support requires lightweight and comprehensible AI models. Smarter precision agriculture solutions can be made possible by integrating deep learning, hyperspectral imaging, and IoT-based monitoring to further improve automated disease identification.

Conclusion

Current research on machine learning for papaya leaf disease diagnosis shows significant progress in automated and accurate crop health assessment. Inspection by manually is very time consuming and stays vulnerable to human mistakes, although the fact that recent deep learning techniques with machine leaning (ML), mainly CNN (Convolutional Neural Network) have established exceptional precision in identifying papaya diseases including anthracnose, ringspot and mosaic. The grouping of extraction of features, ideal models for classification also picture preprocessing takings increased detection efficacy and robustness in a variety of field conditions. However, there are still problems with traditional background and illuminating variations, the lack of publicly available papaya-specific data, and the need for real-time mobile solutions. Large, annotated papaya leaf datasets, hybrid deep learning models, and lightweight architectures for field application must be the key attention of upcoming study. Ultimately, machine learning exposed itself to be a consistent and scalable approach designed for enhancing precision agriculture and sustainable papaya production.

References

[1] Raiyan Gani, Maherun Nessa Isty, Mohammad Rifat Ahmmad Rashid, Jubaer Ahmed, Tasmia Islam, Mahamudul Hasan, Raihan UI Islam, Shamin H. Ripon Ahmed Easif Reza, "Attention-guided convolutional neural network with explainable AI for papaya leaf disease detection," Scientific Reports, 15, Article number: 41489 (2025)

[2] Sumaya Mustofa, Md Taimur Ahad, Yousuf Rayhan Emon, Arpita Sarker, "BDPapayaLeaf A dataset of papaya leaf for disease detection and classification," https://doi.org/10.1016/j.dib.2024.110910, Volume 57, December 2024, 110910

[3] Putra Sumari, Mustafa M.Abd Zaid, Ahmed Kateb Jumaah Nussairi, Shaffiah Binti Mohd Ling Gaomian, Devinna Win Anak Boniface Emag , Chew Yee Ling, "Papaya fruit and leaf ringspot virus detection using CNN and transfer learning," International Journal of Environmental Sciences, 2025.

[4] R Sujatha, Sushil Krishan, Sushil Krishnan, Jyoti Moy Chatterjee, "Advancing plant leaf disease detection integrating machine learning and deep learning," Scientific Reports 15, Article number: 11552 (2025)

[5]Sumaya Mustofa, Md Taimur Ahad , Yousuf Rayhan Emon , Arpita Sarker , "BDPapayaLeaf : A dataset for papaya disease detection , classification, and analysis" , 2024 Sep 10;57:110910. doi: 10.1016/j.dib.2024.110910

[6] Snehal J. Banarase, Suresh D. Shirbahadurkar, "Enhancing Papaya Leaf Disease Detection with CNN and Transfer Learning Fusion for Precise Disease Diagnosis", J. Electrical Systems 20-2s (2024): 1015-1024

[7] M. V. N. Godapitiya; T. K. N. Mandira,"A Deep Learning Approach for Papaya Disease, Pest and Maturity Identification via Mobile Imaging**",** DOI**:** 10.1109/ICAC69156.2025.11361477

[8] Syed Asif Ahmad Qadri, Nen-Fu Huang; Taiba Majid Wani, Showkat Ahmad Bhat, “Advances and Challenges in Computer Vision for Image-Based Plant Disease Detection: A Comprehensive Survey of Machine and Deep Learning Approaches”, IEEE Transactions on Automation Science and Engineering (Volume: 22)

[9] P. Sharma et al., “Machine learning techniques for papaya leaf disease identification,” 2024.

[10] Chetan Harsha Lekkala; Sitadevi Bharatula; C. Sonika; K Kannaiah “ConvNeXt-Small: An Automated Deep Learning Method for Detecting Papaya Leaf Disease Using the BDPapayaLeaf Dataset”, 2025 7th International Conference on Intelligent Sustainable Systems (ICISS) , DOI: 10.1109/ICISS63372.2025.11076525

[11] Taifa Ayoub Mir; Deepak Banerjee; Mukesh Kumar; Ruchira Rawat; Yerrolla Chanti, “Hybridized Model for Improved Papaya Leaf Disease Classification: CNN and Random Forest Integration”, 2024 5th International Conference for Emerging Technology (INCET) DOI: 10.1109/INCET61516.2024.10593105

[12] M. V. N. Godapitiya; T. K. N. Mandira, “ A Deep Learning Approach for Papaya Disease, Pest and Maturity Identification via Mobile Imaging”, 2025 7th International Conference on Advancements in Computing (ICAC) DOI: 10.1109/ICAC69156.2025.11361477

[13] Mohammad Rifat Ahmmad Rashid, Md. AL Ehtesum Korim, Mahamudul Hasan, Md Sawkat Ali, Mohammad Manzurul Islam, Taskeed Jabid, Raihan Ul Islam, Maheen Islam, “ An Ensemble Learning Framework with Explainable AI for interpretable leaf disease detection”, Volume 26, July 2025, 100386 https://doi.org/10.1016/j.array.2025.100386

[14] A Sainath Chaithanya1, M Rachana, “Identification of Diseased Papaya Leaf through Transfer Learning”, Sainath Chaithanya & Rachana / Indian Journal of Science and Technology 2023;16(48):4676–4687

[15] Guoqiang Li, Qing Zhao, Yachao Wang, Peiyan Yuan, “PMVT: a lightweight vision transformer for plant disease identification on mobile devices,” volume 14 - 2023 | https://doi.org/10.3389/fpls.2023.1256773

[16] Ishak Pacal, Ismail Kunduracioglu, Mehmet Hakki Alma, Muhammet Deveci, Seifedine Kadry, Jan Nedoma, Vlastimil Slany & Radek Martinek , “A systematic review of deep learning techniques for plant diseases”, Volume 57, article number 304, (2024)

[17] K. Azarudeen; Prem Sangeeth V; R. Santhana Krishnan; Tamizhselvi Annamalai; J. Relin Francis Raj; M. Prem Nazeer, “Robust Deep Learning Framework for Image-based Classification of Papaya Leaf Diseases using EfficientNet-B0” , 2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN)

[18] A. Verma et al., “Automated classification of papaya leaf diseases using CNN,” 2024.

[19] Md. Faysal Ahamed, Abdus Salam, Md. Nahiduzzaman, M. Abdullah-Al-Wadud & S. M. Riazul Islam, "Streamlining plant disease diagnosis with convolutional neural networks and edge devices" Volume 36, pages 18445–18477, (2024)

[20] M. V. N. Godapitiya, T. K. N. Mandira, "A Deep Learning Approach for Papaya Disease, Pest and Maturity Identification via Mobile Imaging", 2025 7th International Conference on Advancements in Computing (ICAC) DOI:10.1109/ICAC69156.2025.11361477

[21] K. Sakthivel; S. Arularasi; S. Gopinath; M. Vinoth; G. Kowsalya; S. Lalitha ,"Deep Learning-Based Approach for Accurate Plant Disease Identification Using Image Analysis", 2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA) 2025 3rd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)

[22] Ravi Ranjan Kumar; Anuj Kumar Jain; Vikrant Sharma; Purushottam Das; Pooja Sahni "Enhancing Precision in Papaya Crop Health Management: A CNN-Based Approach for Early Disease Detection and Classification", DOI: 10.1109/ICITEICS61368.2024.10625272

[23] Putra Sumari, Mustafa M. Abd Zaid, Ahmed Kateb Jumaah Nussairi, Shaffiah binti Mohd, Ling Gaomian, Devinna Win Anak Boniface Emang, Chew Yee Ling, "Papaya Fruit and Leaf Ringspot Virus Detection Using Convolutional Neural Network (Cnn) And Transfer Learning", Vol. 11 No. 10s, 2025 https://theaspd.com/index.php

[24] Muskan Sharma, Sandhya Umrao, Aditee Mattoo, "Smart agriculture for disease detection in crops or leaves through AI and deep learning ," AIP Conf. Proc. 3297, 090025 (2025) https://doi.org/10.1063/5.0285946

[25] Abhishek Upadhyay, Narendra Singh Chandel, Krishna Pratap Singh, Subir Kumar Chakraborty, Balaji M. Nandede, Mohit Kumar, A. Subeesh, Konga Upendar, Ali Salem & Ahmed Elbeltagi , "Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture," volume 58, article number 92, (2025).

[26] Amin S. Ibrahim, Saeed Mohsen, I. M. Selim, Roobaea Alroobaea, Majed Alsafyani, Abdullah M. Baqasah ,Mohamed Eassa , "AI-IoT based smart agriculture pivot for plant diseases detection and treatment", Scientific Reports volume 15, Article number: 16576 (2025)

[27] Manju Bagga, Sonali Goyal, "Image-based detection and classification of plant diseases using deep learning: State-of-the-art review," 13 February 2024 https://doi.org/10.1002/uar2.20053

[28] Apri Junaidi; Siti Zaiton Mohd Hashim; Mohd Shahizan Bin Othman; Mohd Murtadha Bin Mohamad; Hitham Alhussian; Said Jadid Abdulkadir,"Deep Learning and Edge Computing in Agriculture: A Comprehensive Review of Recent Trends and Innovations", Volume: 13 DOI**:** 10.1109/ACCESS.2025.3595621.

[29] Anuja Bhargava; Aasheesh Shukla; Om Prakash Goswami; Mohammed H. Alsharif; Peerapong Uthansakul; Monthippa Uthansakul, "Plant Leaf Disease Detection, Classification, and Diagnosis Using Computer Vision and Artificial Intelligence: A Review", Volume: 12

[30]Munir Majdalawieh , Carla Martins , Mohammed Radi , Maher Alaraj , Shafaq Khan, "Preision agriculture in the age of AI: A systematic review of machine learning methods for crop disease detection", Volume 12, December 2025, 101491