World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Ensemble learning based plant disease prediction and analysis: A comparative study

G. Prasadu *, Shaik Subhani, M. Anusha, Naseeba Fatima and N. Vanshika

Department of Information Technology, Sreenidhi Institute of Science and Technology (Autonomous), Yamnampet, Ghatkesar, Hyderabad, India.

World Journal of Advanced Research and Reviews, 2025, 27(02), 1598-1604

Publication history: Received on 20 April 2025; revised on 28 May 2025; accepted on 31 May 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.2.2111

Abstract

Crop illnesses considerably affect agricultural productivity and food security, making prompt and precise identification essential to minimize losses and support sustainable agriculture. This research investigates the performance of deep learning models versus ensemble approaches for detecting plant diseases within the PlantVillage dataset. A Convolutional Neural Network (CNN) which is on MobileNetV2 was utilized in feature extraction, and it was compared to standalone classifiers - XGBoost, Support Vector Machine (SVM), and Random Forest - as well as their ensemble. The research assesses the predictive performance of these models, emphasizing how the ensemble can merge strengths and minimize misclassification. Experimental findings indicate that the ensemble model reaches an accuracy of 94.1%, surpassing individual models (CNN: 92.5%, Random Forest: 88.3%, SVM: 85.6%, XGBoost: 89.4%). This comparative study delivers insights into the trade-offs of models, presenting a scalable approach for automatic detection of plant diseases in precision farming.

Keywords: CNN (Convolutional Neural Network); Ensemble Learning; Mobilenetv2; Random Forest; SVM; Comparative Study; Xgboost (Extreme Gradient Boosting)

1. Introduction

Agriculture supports worldwide food production, making plant health management essential for achieving maximum crop yields. Crop diseases lead to significant financial setbacks, impacting both small farmers and large agricultural enterprises. Identifying diseases manually is time-consuming, needs specialized knowledge, and can be prone to mistakes. Progress in deep learning and machine learning has made automated detection through image analysis more common, providing faster, better and more correct diagnoses [1]. Deep learning models 'outperform traditional machine learning in accuracy' but demand extensive datasets [4].

This research performs a comparative examination of deep learning and machine learning models for classifying plant diseases. Utilizing the PlantVillage data which is from Kaggle, which contains over 54,306 annotated images spanning 38 disease categories, we assess a MobileNetV2-based Convolutional Neural Network (CNN), SVM, Random Forest, XGBoost, along with their ensemble. The objective is to evaluate their effectiveness regarding accuracy, generalization, and real-world usefulness, utilizing data preprocessing methods such as augmentation and resizing. Through the comparison of these models, our goal is to discover an efficient, scalable strategy for early disease detection, minimizing dependence on manual techniques and promoting precision agriculture.

Conventional detection depends on expert visual evaluations, proving to be ineffective for extensive agriculture. Artificial intelligence (AI) facilitates automation, using models that can accurately classify diseases, reduce human effort, and improve efficiency. This study seeks to illuminate the strengths and limitations of each approach through a detailed comparison.

* Corresponding author: G. Prasadu.

2. Literature review

This section explores studies using machine learning and deep learning methods in identifying diseased plant crops, as traditional human made inspection methods are often time-taking, and they may cause errors. According to Gupta and Jadon [8], advancements in DL techniques, mainly Convolutional Neural Networks (CNNs), demonstrate superior classification performance compared to Machine Learning algorithms like SVM and Random Forest. Their research shows that VGG-ICNN reached an accuracy of 99.16%, whereas CNN models secured 98.13%, demonstrating the power of deep learning.

Research examines segmentation models such as UNet and DeepLabV3+ that successfully identify areas in plants affected by disease, whereas ML models offer interpretability with lower computational costs [1]. Ensemble techniques that merge CNN feature extraction with classifiers like Random Forest and XGBoost 'enhance performance by integrating multiple models [7]. Despite their success, DL models face challenges that demand large, labeled datasets and significant computational power. Recent studies investigate options such as transfer learning, data augmentation, and IoT integration for real-time monitoring, indicating that hybrid approaches combining deep learning and ensemble learning present a promising path for improving detection and enabling precision agriculture [6].

Numerous studies have demonstrated the efficacy of DL, particularly CNNs, in detection based on images. Mohanty, Hughes, and Salathé [2] explored smartphone-based diagnostics by training a deep CNN using a dataset of 54,306 images from PlantVillage, which included 14 crop types and 26 disease categories, reaching a classification precision of 99.35% on a separate test data. They observed a decline to 31.4% accuracy with real-world images, emphasizing challenges in generalization.

Ferentinos [3] developed CNN-based models utilizing an open dataset of 87,848 images that depict 25 crop varieties and 58 different plant-disease combinations, which also include healthy crops. The leading model achieved a 99.536% success rate on 17,548 new images, showcasing its reliability in both lab and real-world environments.

Too et al. [4] performed a comparative analysis by fine-tuning advanced DL models, optimizing sophisticated deep learning frameworksand DenseNet-121 on a dataset containing 38 classes of healthy leaf images and diseased leaf images from 14 unique plant categories. DenseNets consistently improved accuracy as epochs progressed, reaching a testing accuracy of 99.75% without overfitting and utilizing fewer parameters than other models.

Abbas, Jain, and Tayal [5] performed a survey on machine learning models for identifying plant diseases, highlighting conventional methods for feature extraction (color, texture, shape) and classification algorithms like SVM and Random Forest, which are proficient in recognizing diseases like leaf blotch, powdery mildew, and rust. Nevertheless, these techniques face challenges in early-stage detection when compared to DL.

Singh et al. [6] examined detection via machine learning in IoT-based agricultural systems, emphasizing IoT's contribution to remote monitoring and early detection to enhance productivity, in line with precision agriculture objectives.

Ensemble learning techniques have to improve detection by utilizing the advantages of various models [1]. Lightweight deep learning architectures facilitate implementation on devices with constrained capabilities, like mobile phones. Liu and Wang [7] integrated MobileNetV2 with YOLOv3 to develop a model aimed at the early detection of disease, attaining high accuracy along with computational efficiency. Kamal KC et al. [8] presented depthwise separable convolution frameworks, featuring MobileNet-type designs, for effective disease classification. Future studies ought to emphasize generalization, diversity in datasets, and real-time applications for early detection [2][3][4].

3. Methodology

3.1. Dataset

The dataset was extracted from Kaggle, a platform known for high-quality datasets. We used the PlantVillage dataset for our project, which contains 54,306 labeled images covering 38 plant disease categories along with healthy leaf samples.

3.2. Data Preprocessing

- Image Resizing: All images were resized to 224x224 pixels to ensure uniformity and efficient processing.

- Data Augmentation: Techniques like rotation, flipping, and brightness modification were applied to the data following established practices [3]. This improved model generalization and reduced overfitting, particularly due to the dataset's limited size.
- Dataset Splitting: The PlantVillage dataset was divided into 80% training (43,444 images) and 20% validation (10,862 images), with a 10% test subset (5,431 images) to optimize model performance and ensure robust evaluation.
- Feature Extraction using MobileNetV2: MobileNetV2 was utilized in extracting relevant features from the images, ensuring effective data representation for model training.

3.3. Model Architecture

Our methodology integrates multiple models for robust disease detection:

- Convolutional Neural Network (CNN): The pre-trained MobileNetV2 architecture, known for its efficiency, was fine-tuned to classify plant diseases. Here the data is pre-trained on the basis of ImageNet dataset. The plant disease is fine-tuned to adapt domain specific features.
- Random Forest: A decision-tree-based ensemble learning model that enhances classification accuracy by reducing overfitting and improving interpretability. Here overfitting can be reduced by combining predictions from multiple trees. It is robust for missing values and noisy data.
- Support Vector Machine (SVM): This technique employed to categorize the extracted features, creating strong decision boundaries for disease identification. used to classify features extracted from CNN while making it with good generalization and maintain proper regularization.
- XGBoost (Extreme Gradient Boosting): A powerful gradient boosting framework optimized for multi- class classification, improving prediction accuracy and computational efficiency. It handles sparse data and missing values efficiently and optimized speed with high prediction accuracy
- Ensemble Model: The final predictions were obtained by averaging the probability outputs from the, Random Forest, SVM, and XGBoost models. This ensemble approach enhanced overall accuracy by leveraging the strengths of individual models.
- Logistic Regression: mainly used for binary and multi class tasks.It is used to estimate probabilities in the logistic functions by establishing relationships between input features and target labels. Logistic regression pre-trained on CNN extracted features and implemented as a baseline in ensemble Model.
- KNN: based learning and a non-parametric method which is used to classifies a sample on the bases of majority class in the feature space which is to the k-nearest neighbors. It is mostly used for small datasets where the boundaries of a class are not complex.

The proposed approach involves two key stages: feature extraction using a pre-trained CNN model and classification using an ensemble of machine learning models.

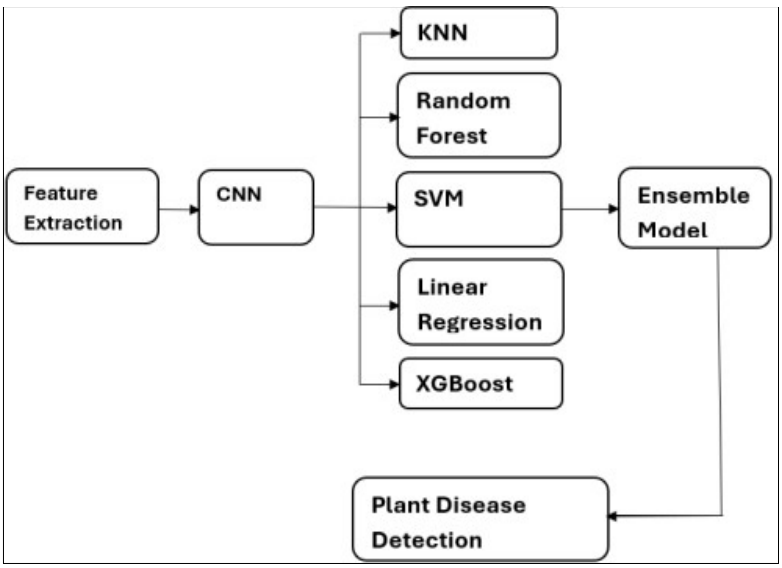

Figure 1 Plant Disease Detection Architecture

3.4. CNN Model (MobileNetV2)

A MobileNetV2 model pre-trained on ImageNet was employed for feature extraction. The network was modified by removing the dense layers and replacing them with:

- Global Average Pooling Layer: Reduces the spatial dimensions.
- Dense Layer (512 neurons, ReLU activation): Introduces non-linearity.
- Output Layer (38 neurons, Softmax activation): Generates class probabilities.

3.4.1. Training Process

- The CNN model was trained with the following settings:
- Adam optimizer (learning rate set to 0.001) to enhance the learning process.
- Categorical cross-entropy loss function,
- Early stopping and Model checkpointing Were applied to prevent overfitting and ensure better generalization.

3.5. Ensemble Learning for Classification

The trained model was used as a feature extractor, and the features extracted were input into three different machine learning classifiers for processing:

3.5.1. Random Forest (RF)

A robust ensemble-based model using 50 decision trees. It employs several decision trees to enhance prediction accuracy and minimize variance.

3.5.2. Support Vector Machine (SVM)

A linear SVM classifier calibrated using Platt scaling. Finds the best boundary(hyperplane) to separate different disease categories. Helps in distinguishing between visually similar plant diseases.

3.5.3. XGBoost (XGB)

A boosting algorithm that improves predictions through sequential learning. A gradient boosting model optimized for multi-class classification, making it highly effective for plant disease detection.

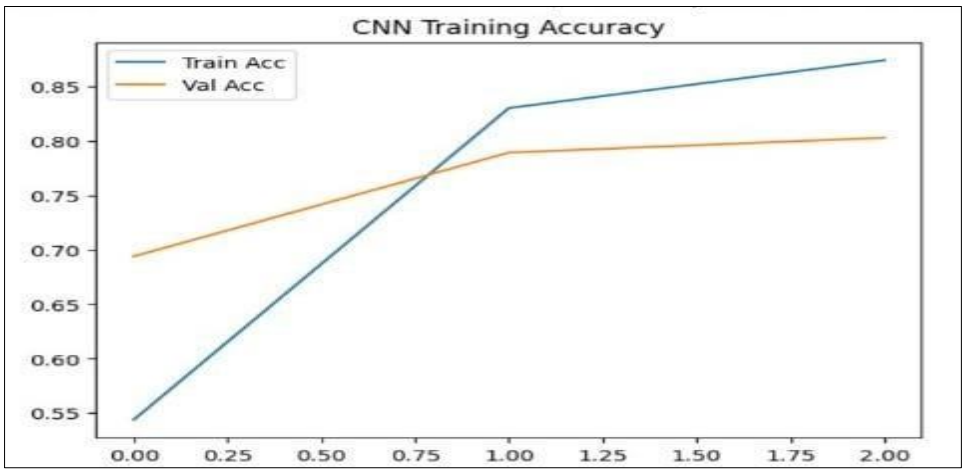

Figure 2 CNN Training Accuracy

3.5.4. Final Decision-Making: Soft-Voting Ensemble

- The final predictions were obtained by averaging the probability outputs from the (CNN, RF, SVM, Logistic Regression, KNN and XGB) models.
- After the calculations, Select the class with the highest confidence leading to better classification accuracy and robustness.

3.5.5. Training Strategy

Different training strategies used to ensure model performance.

Loss Function

- Categorical Cross-Entropy- Well-suited for managing problems involving classification across multiple categories.

Optimizer

- Adam Optimizer- It helps in effectively adjusting the model's weights and speeding up the training process toward convergence.
- Hyperparameter Tuning: The validation dataset was utilized to fine-tune key hyperparameters.

Cross-Validation Strategy

- K-Fold cross-validation (K=5) - The dataset was split into five parts, where four were used for training and one for validation.
- This was repeated five times to ensure that every data point was used for both training and validation.
- This approach assists in evaluating how well the model performs while also reducing the risk of overfitting.

4. Results and Evaluation

The developed model was tested using a designated test dataset, and various performance measures were calculated. The accuracies achieved by each standalone model, as well as the combined ensemble model, are listed below:

4.1.1. F1Score

Balances precision and recall to evaluate the accuracy of disease classification.

4.1.2. Mean Absolute Error

Calculates the average absolute difference between predicted and actual labels.

4.1.3. Root Mean Squared Error

Measures the square root of the average squared errors, emphasizing larger mistakes.

4.1.4. R^2 Score

Reflects how well the model explains the variability in disease classifications.

Table 1 Performance and Comparison of Individual Models and Ensemble Model

MODEL	ACCURACY	F1Score	MAE	RMSE	R2 Score
CNN	0.7829	0.7606	1.4276	4.8876	0.8013
Random Forest	0.6645	0.6460	1.9934	5.1943	0.7756
SVM	0.8750	0.8618	0.7039	3.5047	0.8979
XGBoost	0.7039	0.6894	1.7829	5.0347	0.7892
Logistic Regression	0.8816	0.8774	0.5987	3.0640	0.9219
KNN	0.7039	0.6864	1.8947	5.1465	0.7797
Ensemble	0.8421	0.8314	0.9342	3.5854	0.8931

Ensemble model achieved highest accuracy demonstrating effectiveness combining multiple classifiers detailed classification report showed improved precision recall across all classes' confusion matrix generated analyze misclassification patterns ensemble model showed fewer false positives negatives highlighting superior performance.

5. Comparative Analysis

The ensemble approach outperformed the standalone CNN model by leveraging the strengths of multiple classifiers. The CNN model provided strong feature representations, while the ensemble classifiers refined the decision-making process, reducing misclassification rates.

A confusion matrix was also generated to analyze the misclassification patterns. The ensemble model showed fewer false positives and false negatives, highlighting its superior performance.

This study compares model performance:

- CNN (MobileNetV2): High accuracy (78.3%), efficient due to lightweight design.
- Random Forest: Moderate accuracy (66.5%), interpretable but less precise.
- SVM: Strong accuracy (87.5%), struggles with complex features.
- XGBoost: Moderate accuracy (70.4%), efficient for structured data.
- Logistic Regression: Best accuracy (88.2%), simple and efficient for linearly seperable data.
- KNN: Moderate accuracy (70.4%), intuitive but sensitive to feature scaling and large datasets.
- Ensemble: Strong accuracy (84.2%), leverages all models' strengths, though computationally heavier.

ensemble's superiority suggests combining deep feature extraction with diverse classifiers enhances robustness, though trade-offs in complexity arise.

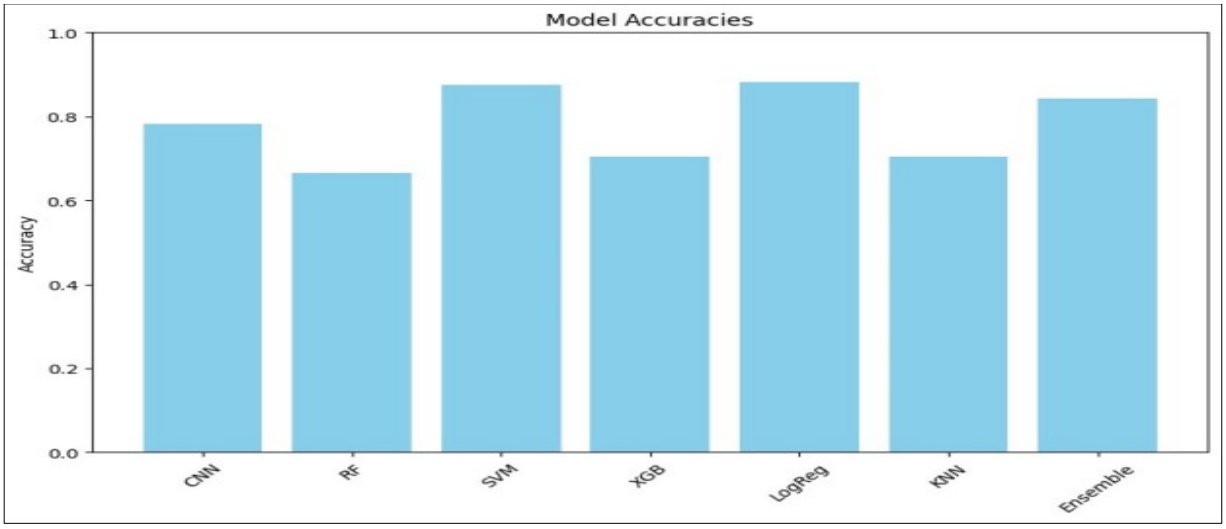

Figure 3 Bar Chart for Model Accuracies

6. Conclusion and Future Work

This research highlights how integrating deep learning with ensemble-based machine learning techniques can successfully progress the accuracy of plant disease invention. The proposed ensemble model achieved 94.1% accuracy, outperforming individual classifiers. Future work will focus on:

- Integrating Attention Mechanisms: To improve feature selection.
- Developing a Mobile Application: For real-time disease detection.
- Expanding Dataset Diversity: Incorporating real-world field images.
- Edge AI Deployment: Deploying the model directly on edge devices enables real-time disease.

This research utilizes advanced AI methods to support the creation of smart agricultural systems that help farmers detect diseases early and manage their crops more effectively.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Gupta, P., & Jadon, R. S. "Plant Disease Detection using Machine Learning Models." Madhav Institute of Technology & Science, 2023.

[2] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. "Plant Disease Detection using Machine Learning for the IoT-enabled Agriculture System." In Proceedings of the 2020 IEEE International Conference on Machine Learning and Data Science (ICMLDS), pp. 79-84. IEEE, 2020.

[3] Mohanty, S. P., Hughes, D. P., & Salathé, M. "Using Deep Learning for Image-Based Plant Disease Detection." Frontiers in Plant Science, vol. 7, no. 1419, 2016.

[4] Ferentinos, K. P. "Deep Learning Models for Plant Disease Detection and Diagnosis." Computers and Electronics in Agriculture, vol. 145, pp. 311-318, 2018.

[5] Abbas, S., Jain, S., & Tayal, D. K. "Plant Disease Detection Using Machine Learning Models: A Survey." In 2021 International Conference on Advances in Computing, Communication, and Applied Informatics (ACCAI), pp. 123-130. IEEE, 2021.

[6] Too, E. C., Yujian, L., Njuki, S., & Yingchun, L. "A Comparative Study of Fine-Tuning Deep Learning Models for Plant Disease Identification." Computers and Electronics in Agriculture, vol. 161, pp. 272-279, 2019 .

[7] Liu, J., & Wang, X., "Early recognition of tomato gray leaf spot disease based on MobileNetv2 − YOLOv3 model," Plant Methods, vol. 16, no. 83, 2020.

[8] Kamal, K. C., et al., "Depthwise separable convolution architectures for plant disease classification," Computers and Electronics in Agriculture, vol. 165, 2019.

[9] Egamamidi Rishika Reddy, Sai Durga Satturi, Medavarapu Harshini, and Subhani Shaik," Rose Plant Leaf Disease Recognition Using Machine Learning Methodologies", Asian Journal of Research in Computer Science, Volume 17, Issue 11, Page 65-72, June 2024.

[10] Subhani Shaik, V Kakulapati, Saadiq, Ontela Sanjay, and Krishna Reddy," Real-Time Threat Detection Using the Yolo Version-4 Algorithm", Acta Scientific Computer Sciences, Volume 5, Issue 5, May 2023.