Plant Disease Classification Using Artificial Intelligence: A Comparative Study of Machine Learning Algorithms

Varun Raj[1], Jwala Jose[2], Gibi K S[3]
[1]*Student, Department of Computer Science, Don Bosco College, Sulthan Bathery*
[2,3]*Assistant Professor, Department of Computer Science, Don Bosco College, Sulthan Bathery*

***Abstract*—Plant diseases are a significant threat to global agriculture, impacting crop yields and food security. Early detection and accurate classification of these diseases are crucial for minimizing their adverse effects. Traditional methods of plant disease diagnosis, often reliant on expert knowledge and manual inspection, are time-consuming and impractical for large-scale applications [1]. This paper investigates the use of machine learning (ML) algorithms for automating the detection and classification of plant diseases. A comparative study was conducted using four popular ML models: Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Convolutional Neural Networks (CNN). The models were evaluated on the publicly available PlantVillage dataset, which contains labeled images of diseased and healthy plant leaves. Several performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, were used to assess the efficacy of each algorithm. The results indicate that CNN outperforms traditional machine learning algorithms in terms of classification accuracy and robustness, achieving a high accuracy of 96.4%.[2] Random Forest and Gradient Boosting also demonstrated strong performance, while SVM showed relatively lower accuracy. This study highlights the potential of deep learning techniques for plant disease classification and offers insights into the strengths and limitations of different machine learning models for agricultural applications.**

***Index Terms*—Plant disease classification, machine learning, deep learning, Convolutional Neural Networks (CNN), Random Forest, Support Vector Machines (SVM), Gradient Boosting, image processing, agricultural automation, plant health monitoring, AUC-ROC, feature extraction, dataset, AI in agriculture.**

I. INTRODUCTION

Agriculture remains a cornerstone of human civilization, feeding billions of people worldwide. However, plant diseases pose a significant threat to crop yields, leading to extensive economic losses each year. Early detection and classification of plant diseases are crucial for mitigating their effects. Traditionally, plant disease identification has been performed manually by expert botanists and agricultural specialists. While effective, this approach is time-consuming, expensive, and highly dependent on the availability of expert resources. With the advent of Artificial Intelligence (AI) and Machine Learning (ML), there has been a marked shift toward automating the process of plant disease detection, making it faster, more efficient, and accessible.[3]

In recent years, a variety of machine learning algorithms have been proposed for plant disease classification, offering promising solutions. These models can be trained to analyze plant images, detect symptoms of diseases, and classify them accurately. This paper provides a comparative study of several popular machine learning algorithms used for plant disease detection, including Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Convolutional Neural Networks (CNN). These algorithms differ in their underlying mechanics, performance, and suitability for specific datasets. This study aims to provide insights into the strengths and limitations of these approaches in the context of plant disease classification.[4]

II. LITERATURE REVIEW

In recent years, the application of AI for plant disease classification has gained significant attention. Various studies have explored different machine learning models for detecting plant diseases from images. The most common approach involves using datasets consisting of leaf images captured under

various conditions, and training models to recognize specific symptoms of plant diseases.[5]

For example, in 2018, **Nitin et al.** proposed a hybrid system combining SVM and k-means clustering to identify diseases in plants based on leaf images. Their approach achieved an accuracy of 90%, highlighting the potential of machine learning for plant disease classification. Similarly, **Patel et al. (2020)** used deep learning techniques, particularly CNNs, to classify plant diseases from a large dataset of plant images. Their study demonstrated that CNN-based methods significantly outperformed traditional machine learning approaches like SVM and RF in terms of classification accuracy and processing time.[6]

Other studies, such as **Li et al. (2022)**, have focused on combining traditional machine learning models with data augmentation techniques to improve classification performance. Data augmentation helps mitigate issues related to limited training datasets, a common challenge in plant disease classification. The incorporation of transfer learning in deep learning models has also been explored to enhance performance, especially in cases where the dataset size is relatively small.

This paper builds upon these foundational studies by comparing the performance of multiple machine learning models on a common dataset, focusing on their strengths and weaknesses, and offering insights into their practical applicability.

III. METHODOLOGY

3.1 Dataset Description

The dataset was preprocessed to standardize image size and resolution and was split into training and test sets using a 70:30 ratio. Data augmentation [7] techniques such as rotation, flipping, and zooming were applied to improve model robustness and prevent overfitting, particularly given the inherent class imbalance in the dataset.

3.2 Preprocessing and Feature Extraction

Image preprocessing plays a critical role in enhancing model performance. The raw leaf images undergo several preprocessing steps before being fed into the machine learning algorithms. These steps include:

- Resizing and Normalization: All images were resized to a uniform size of 224x224 pixels. Additionally, pixel values were normalized to a range of [0, 1] to ensure consistency in the input data.
- Color Space Conversion: The images were converted from RGB to grayscale and HSV color spaces for easier analysis of plant symptoms, as certain diseases [8] manifest more clearly in specific channels like the hue and saturation of the image.
- Feature Extraction: For traditional machine learning algorithms like Random Forest, SVM, and GBM, feature extraction was performed using techniques like Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Gabor filters. These methods aim to extract high-level features like edges, textures, and shapes from the leaf images.

3.3 Model Implementation

3.3.1 Random Forest (RF): Random Forest (RF) is an ensemble learning method that combines multiple decision trees to improve classification performance. Each tree in the forest is trained on a random subset of features, which ensures diversity among the trees and prevents overfitting. The final output is determined by the majority vote of all the individual trees. Random Forest has been widely used in many image classification tasks due to its simplicity, high accuracy, and ability to handle large datasets with minimal tuning.

3.3.2 Support Vector Machines (SVM): Support Vector Machines (SVM) are supervised learning algorithms that find the optimal hyperplane to separate different classes in the feature space. For multi-class classification, the "one-vs-one" or "one-vs-all" strategy is typically used. SVM is known for its ability to handle high-dimensional data and works well in cases where the class boundaries are complex. In this study, we used the RBF (Radial Basis Function) kernel to map the input data into a higher-dimensional space for better separation.

3.3.3 Gradient Boosting Machines (GBM)

Gradient Boosting is an ensemble technique where models are trained sequentially, with each new model trying to correct the errors made by the previous one. GBM is known for its strong predictive performance and ability to handle various types of data, including images. In this study, we utilized the XGBoost implementation, which is highly optimized for speed and accuracy.

3.3.4 Convolutional Neural Networks (CNN): Convolutional Neural Networks (CNN) are a class of deep learning algorithms that have revolutionized image classification tasks. CNNs automatically learn spatial hierarchies of features by applying convolutional filters to the input images. The layers in CNNs gradually build complex features, making them particularly suitable for image-based tasks. In this study, we used a simple architecture with three convolutional layers followed by two fully connected layers for classification.

IV. ALGORITHM COMPARISON

4.1 Evaluation Metrics

To assess the performance of each model, we used several evaluation metrics, including:

- Accuracy: The proportion of correct predictions made by the model.
- Precision: The ratio of correctly predicted positive observations to the total predicted positives.
- Recall: The ratio of correctly predicted positive observations to all observations in the actual class.
- F1-Score: The harmonic means of precision and recall, providing a balanced measure of model performance.
- AUC-ROC: The Area Under the Receiver Operating Characteristic Curve, which measures the model's ability to distinguish between classes.

4.2 Results

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Random Forest (RF)	92.3	89.6	90.5	90.0	0.94
Support Vector Machines (SVM)	89.4	88.0	89.0	88.5	0.91
Gradient Boosting (GBM)	91.2	89.2	90.0	89.6	0.93
Convolutional Neural Network (CNN)	96.4	94.5	95.1	94.8	0.97

The performance metrics of the four machine learning algorithms, namely Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Convolutional Neural Networks (CNN), reveal significant differences in their ability to classify plant diseases effectively. CNN demonstrated the highest overall performance, achieving an accuracy of 96.4%, a precision of 94.5%, recall of 95.1%, and an F1-score of 94.8%. This indicates that CNN not only excelled in correctly identifying plant diseases but also in balancing precision and recall, making it the most robust model for this task. Furthermore, CNN's AUC-ROC value of 0.97 signifies its excellent capability to distinguish between healthy and diseased plants across a wide range of thresholds, further supporting its effectiveness in plant disease classification.

Random Forest (RF),[9] while not outperforming CNN in every metric, still achieved strong results with an accuracy of 92.3%, precision of 89.6%, recall of 90.5%, and an F1-score of 90.0%. RF performed admirably, with slightly lower precision and recall compared to CNN, but it still demonstrated a strong ability to identify both healthy and diseased plants. The AUC-ROC of 0.94 further reinforces its reliability in distinguishing between the classes, although it was outperformed by CNN in this regard.

Gradient Boosting Machines (GBM)[10] followed closely behind RF, with an accuracy of 91.2%, precision of 89.2%, recall of 90.0%, and an F1-score of 89.6%. GBM's performance was slightly weaker than RF's, particularly in terms of precision, and it demonstrated a slightly lower AUC-ROC value of 0.93. However, it still proved to be a competitive algorithm for plant disease classification, especially considering its ability to correct errors sequentially and improve its performance over iterations.

Lastly, Support Vector Machines (SVM)[11] showed the lowest performance across all metrics, with an accuracy of 89.4%, precision of 88.0%, recall of 89.0%, and an F1-score of 88.5%. While SVM demonstrated decent performance, it struggled in comparison to the other models, particularly in precision and recall, indicating that it was more prone to misclassifications. The AUC-ROC value of 0.91

further highlights its relative difficulty in distinguishing between classes compared to the other models.

In conclusion, while [12] CNN clearly outperforms the traditional machine learning algorithms in plant disease classification, Random Forest and Gradient Boosting are still highly competitive models for applications with more limited computational resources. SVM, while useful in certain contexts, appears to be less effective for this specific task. The results suggest that deep learning models, particularly CNN, are the most suitable choice for accurately identifying plant diseases when sufficient computational resources and data are available.

V. DISCUSSION

The experimental results show that deep learning models, especially CNN, provide superior performance in plant disease classification tasks. CNNs automatically learn the relevant features from images, making them highly effective for image-based tasks. On the other hand, traditional machine learning models like Random Forest, SVM, and Gradient Boosting require manual feature extraction, which may limit their performance when compared to CNNs. However, these models are computationally less expensive and may perform adequately in scenarios where computational resources are limited.

While CNNs demonstrated the highest accuracy, they also require substantial computational resources and a large amount of labeled data to train effectively. In contrast, traditional machine learning models are less resource-intensive and can perform well with smaller datasets. Therefore, the choice of algorithm depends on the specific application, dataset size, and available computational resources.

VI. CONCLUSION

In this study, we have compared several machine learning algorithms for plant disease classification using a publicly available leaf image datase t[13]. The results suggest that deep learning models, particularly Convolutional Neural Networks, offer the best performance in terms of accuracy and robustness. However, traditional machine learning algorithms like Random Forest and Gradient Boosting provide strong alternatives for scenarios with limited data or computational resources.

Future work could focus on optimizing CNN architectures, exploring transfer learning techniques, and incorporating additional data sources, such as environmental conditions, to further improve classification accuracy. Additionally, deploying these models in real-world agricultural settings would be a valuable step toward their practical application in the early detection of plant diseases.

REFERENCE

[1] S. He and K. M. Creasey Krainer, “Pandemics of People and Plants: Which Is the Greater Threat to Food Security?,” *Molecular Plant*, vol. 13, no. 7, pp. 933–934, Jul. 2020, doi: https://doi.org/10.1016/j.molp.2020.06.007.

[2] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, “A Systematic Literature Review on Machine Learning Applications for Sustainable Agriculture Supply Chain Performance,” *Computers & Operations Research*, vol. 119, no. 1, p. 104926, Feb. 2020, doi: https://doi.org/10.1016/j.cor.2020.104926.

[3] J. D. Floros *et al.*, “Feeding the World Today and Tomorrow: The Importance of Food Science and Technology,” *Comprehensive Reviews in Food Science and Food Safety*, vol. 9, no. 5, pp. 572–599, Aug. 2010, doi: https://doi.org/10.1111/j.1541-4337.2010.00127.x.

[4] A. Jafar, N. Bibi, Rizwan Ali Naqvi, Abolghasem Sadeghi-Niaraki, and D. Jeong, “Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations,” *Frontiers in Plant Science*, vol. 15, Mar. 2024, doi: https://doi.org/10.3389/fpls.2024.1356260.

[5] P. P. Shinde and S. Shah, “A Review of Machine Learning and Deep Learning Applications,” *IEEE Xplore*, Aug. 01, 2018. https://ieeexplore.ieee.org/document/8697857.(accessed Dec. 20, 2020).

[6] A. Chug, A. Bhatia, A. P. Singh, and D. Singh, “A novel framework for image-based plant disease detection using hybrid deep learning approach,” *Soft Computing*, Jun. 2022, doi: https://doi.org/10.1007/s00500-022-07177-7.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, Jul. 2019, doi: https://doi.org/10.1186/s40537-019-0197-0.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, doi: https://doi.org/10.1109/iccv.2017.244.

[9] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geology Reviews*, vol. 71, no. 71, pp. 804–818, Dec. 2015, doi: https://doi.org/10.1016/j.oregeorev.2015.01.001.

[10] Z. Liu, Y. Li, F. Ren, W. L. Goh, and H. Yu, "SqueezedText: A Real-Time Scene Text Recognition by Binary Convolutional Encoder-Decoder Network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: https://doi.org/10.1609/aaai.v32i1.12252.

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, doi: https://doi.org/10.1109/cvprw.2010.5543262.

[12] P. K. Malaker *et al.*, "First Report of Wheat Blast Caused by Magnaporthe oryzae Pathotype triticum in Bangladesh," *Plant Disease*, vol. 100, no. 11, pp. 2330–2330, Nov. 2016, doi: https://doi.org/10.1094/pdis-05-16-0666-pdn.

[13] A. Pandey and K. Jain, "A robust deep attention dense convolutional neural network for plant leaf disease identification and classification from smart phone captured real world images," *Ecological Informatics*, vol. 70, p. 101725, Sep. 2022, doi: https://doi.org/10.1016/j.ecoinf.2022.101725.