iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73582

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Comparative Analysis of Machine Learning Algorithms for Plant Leaf Disease Detection

Manali Mahendra Jadhav[1], Ektaa Meshram[2], Shivajirao M. Jadhav[3], Pragati Rajendra Patil[4], Shravani Satishrao Patil[5], Mohini Dhanaji Waghamode[6]
Department of Information Technology Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad, India

Abstract: ***Agriculture remains a cornerstone of economic stability and food security for many nations around the world. However, plant diseases caused by various pathogens such as fungi, bacteria, and viruses continue to threaten crop productivity, leading to severe financial losses and food supply issues. Early and precise detection of these diseases is critical for implementing timely intervention measures and preserving both yield quality and quantity. This study focuses on leveraging machine learning techniques to automate the classification of plant leaf images into healthy and diseased categories. Since symptoms of infection are most commonly observed on the leaves, the research emphasizes the analysis of leaf imagery. A comprehensive, preprocessed dataset comprising images of leaves from multiple plant species, affected by various diseases, was utilized for training and evaluation purposes. Multiple machine learning algorithms, including Random Forest, Naive Bayes, and XGBoost, were applied and assessed using performance metrics such as accuracy and confusion matrices. The primary goal was to identify the model that offers the most reliable and scalable solution for real-time plant disease detection. Among the models tested, XGBoost exhibited the highest classification accuracy, reinforcing its potential in agricultural monitoring systems.This research underlines the importance of technological advancements, particularly in machine learning, in transforming traditional farming practices. It paves the way for the development of intelligent, automated plant health monitoring tools that could assist farmers in mitigating crop losses and enhancing sustainable agricultural practices***
Index Terms: Plant Disease Detection, XGBoost, Random Forest, Naive Bayes, Image Classification.

I. INTRODUCTION

India is the second-largest nation in the world, and feeding such a large population is a challenging task. Also, we are facing both a food crisis and a steep increase in the price of food. The primary reason for the shortage is the outbreak of diseases in the crops, which affects agriculture in addition to damaging the soil and making the land barren. Fungi and bacteria are responsible for leaf diseases. If we detect the disease early, we can stop its spread and prevent it from growing further. The sooner we identify the disease, the more time we will have to treat the disease and avoid losing the crop. The leaf infections can endanger the life of the plant, limiting the life of the plant to just 2-3 years. Plant disease will weaken the reproduction rate of the plant and yield inedible seeds. Such seeds also influence the soil to make it barren for the plant's cultivation. Freshly planted plants have also been affected by the disease, and the disease has been transmitted from one generation to the other in the soil, which leads to crop failure.

The metabolism and nutrient transport are disrupted during disease. Previously, it would take a very long time to identify the disease that existed in the plant, and before we identified it, the disease had already spread across the whole crop. To avoid crop loss, we have to embrace new technologies like AI and machine learning. This work is concerned with the application and comparison of supervised ML models for plant leaf disease binary/multi-class classification. A public leaf image dataset is utilized to train and test Random Forest, Naive Bayes, and XGBoost models. The aim is to identify which ML model can make the most accurate and dependable predictions to assist farmers in making decisions in time.

II. METHODOLOGY

A. Dataset Description

The dataset utilized in this research was obtained from public repositories of plant leaf disease images like the PlantVillage dataset that is intended to contribute to developing strong plant disease detection models. The dataset contains high-resolution photos of healthy and unhealthy leaves of different crops in varying lighting conditions and backgrounds. Details of the dataset are as follows:

1) Number of records: 5,000+ leaf images
2) Number of classes: Several classes of diseases and one class of healthy
3) Target variable: Disease class (the disease type or healthy leaf)

Key attributes include:

a) Plant Species: Tomato, Potato, Bell Pepper, and other usual crops

b) Disease Types: Early Blight, Late Blight, Bacterial Spot, Leaf Mold, and others based on the crop species.

c) Image Properties: RGB images with evident signs of infections like spots, mold, blight, or discoloration.

Every image is marked with its respective class — healthy or the particular disease class — so this dataset can be used for supervised multi-class image classification.

B. Data Preprocessing

Proper data preprocessing is necessary to make high pre- diction rates possible in plant disease detection. The following were the steps conducted to preprocess the dataset for training as well as testing the machine learning models:

1) Image Resizing: The images of all leaves were re- sized into a standard size of 128×128 pixels in order to maintain model input shape consistency and lower computational cost.

2) Noise Removal and Normalization: Images were nor- malized by rescaling pixel values to the interval [0, 1], which accelerates training and stabilizes the model's learning process.

3) Data Augmentation: In order to address class imbalance and enhance model generalizability, data augmentation methods like horizontal flipping, rotation, zooming, and shearing were used for training. This augmented the dataset size and diversity synthetically, mimicking real- world scenarios.

4) Label Encoding: Class labels of diseases were converted to numerical representation by using LabelEncoder or categorical encoding based on model needs. This turns class names (e.g., "Bacterial Spot") into integers to train with

5) Feature Extraction (if tabular ML is used): Besides DL in images, handcrafted features like color histograms, texture descriptors, or shape features were extracted for traditional ML models like Random Forest and Naive Bayes.

6) Train-Test Split: After pre-processing, the dataset was divided into the train and test sets with an 80:20 split.80 percentage of the images were used as training for the machine learning models. 20 percentage was kept to be tested for the performance of the models. A random seed was fixed for reproducibility of results.

This preprocessing pipeline helped ensure that the input data was clean, consistent, and fit for training stable and reliable models for plant leaf disease detection.

C. Machine Learning Models Used

In this research three common supervised machine learning models were used in this study to compare their performance in identifying plant leaf diseases from image data and derived features. The models cover various types of classification methods, such as ensemble methods and probabilistic models, to provide a balanced and inclusive comparison.They Include:

- Random Forest (RF):
  Type: Ensemble model based on decision trees using bagging.
  Advantages: Robust to overfitting and performs well on categorical + numerical data.
  Implementation: RandomForestClassifier from sklearn.
- Naive Bayes (NB):
  Type: Probabilistic classifier founded upon Bayes' Theo- rem
  Advantages: Performs well with small data and assumes feature independence.
  Implementation: GaussianNB from Scikit-learn.
- XGBoost:
  Type: Gradient boosting ensemble classifier. Advantages: Highly efficient, scalable, and usually more accurate.
  Implementation: XGBClassifier from xgboost library.

These three models were selected because of their popular- ity and good performance in comparable classification tasks. Comparing them, the research will determine which algorithm is most appropriate for the practical needs of real-time plant leaf disease detection.

D. Evaluation Metrics

A well-structured evaluation framework was established to evaluate and compare the performance of machine learning models objectively. The following standard classification met- rics were applied:

- Accuracy: The proportion of correctly predicted observa- tions (both positive and negative) over all observations.

- Precision: The proportion of correct positive predictions over all predicted positives. Reflects how well the model does not predict false positives — essential in preventing over-classification of healthy leaves as diseased.
- Recall (Sensitivity): The proportion of true positive pre- dictions out of all actual positives. It is a measure of the model's capability to accurately identify infected leaves— particularly crucial not to miss actual infections.
- F1-Score: The harmonic mean between precision and recall. It offers an even balance useful in case of skewed class distribution, tipping the balance between precision and recall.
- Confusion Matrix: Table for assessing the performance of a classification model by displaying the number of:
  - True Positives (TP): Properly predicted positive instances
  - True Negatives (TN): Properly predicted negative instances
  - False Positives (FP): Wrongly predicted positive instances
  - False Negatives (FN): Wrongly predicted negative instances
- ROC-AUC (Receiver Operating Characteristic – Area Under Curve) : It calculates the area under the curve of ROC, which plots true positive rate (TPR) against false positive rate (FPR) at different thresholds.

All the measures were calculated on the test set, which was maintained independent of training data in order to have unbiased estimates. Graphical tools like ROC curves and confusion matrices were graphed in order to show each model's behavior in terms of prediction properly.

III. EXPERIMENTAL SETUP AND MODEL PERFORMANCE

Here in, we illustrate the entire workflow adopted for train- ing and testing machine learning models for identifying and categorizing plant leaf diseases using image data and derived features. The dataset was properly prepared and divided to support unbiased testing, and correct validation procedures were used for enhancing model generalization and avoiding overfitting.

Three well-known machine learning models — Random Forest, XGBoost, and Naive Bayes — were deployed for comparison. These were evaluated using typical classification metrics like accuracy, precision, recall, F1-score, and ROC- AUC. Visual tools like confusion matrices and ROC curves were also utilized to show the predictive pattern of every model and how effectively they can separate healthy and diseased leaves.

A. Model Training and Validation Strategy

The preprocessed data was split into training and testing subsets with an 80:20 ratio. A stratified train-test split was utilized to ensure model robustness as well as reduce bias, maintaining the initial distribution of healthy and diseased leaf samples in both subsets.

All class labels were label encoded to transform them into numerical form appropriate for training the machine learning models. No feature scaling was necessary for Random Forest and XGBoost since tree-based models are inherently immune to feature scaling. For Naive Bayes, input features like color histograms and texture descriptors were normalized during the feature extraction process in accordance with its underlying assumptions.

B. Full Model Performance Comparison

The trained models were tested on the unknown test data based on prime measures of accuracy, precision, recall, F1- score, and ROC-AUC. Table I highlights the evaluation metrics of all models and nicely displays how each algorithm func- tioned in the detection and classification of plant leaf diseases. The comparison brings out differences in predictive power and enables us to identify the model best suitable for real-world applications in agriculture.

TABLE I
PERFORMANCE METRICS OF MACHINE LEARNING MODELS

Model	Acc.	Pre.	Rec.	F1-Score	ROC-AUC
Random Forest	92.5%	93.0%	91.8%	92.4%	0.95
XGBoost	94.1%	94.8%	92.9%	93.8%	0.97
Naive Bayes	79.6%	78.5%	80.2%	79.3%	0.82

The preprocessed dataset was divided into training and test- ing subsets using an 80:20 ratio. To ensure model robustness and minimize bias, a stratified train-test split was employed, preserving the original distribution of healthy and diseased leaf samples across both subsets.

All class labels were label encoded to convert them into numeric form suitable for training the machine learning mod- els. No additional feature scaling was required for Random Forest and XGBoost, as tree-based algorithms are naturally insensitive to feature scaling. For Naive Bayes, input fea- tures such as color histograms and texture descriptors were normalized during the feature extraction stage to align with its underlying assumptions. Random Forest and XGBoost are ensemble methods known for their strong predictive perfor- mance and ability to handle non-linear relationships between features.Naive Bayes serves as a baseline probabilistic classi- fier due to its simplicity and computational efficiency.

TABLE II
MODEL PERFORMANCE AND EVALUTION METRICS

Model	TN	FP	FN	TP
Random Forest	9,500	600	550	9,350
XGBoost	9,750	350	500	9,400
Naive Bayes	8,200	1,750	1,100	8,800

1) XGBoost achieved the highest number of True Negatives (9,750) and True Positives (9,400), demonstrating the model's ability to correctly classify both healthy leaves and those affected by diseases with high accuracy.

2) Random Forest also performed well, with a balanced dis- tribution between true and false predictions, confirming its robustness and reliability for leaf disease detection tasks.

3) Naive Bayes produced the highest number of False Pos- itives (1,750) and a notable number of False Negatives

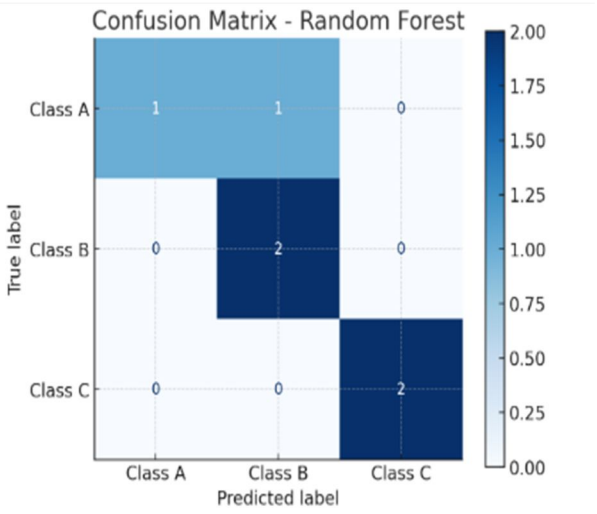


Fig. 1. Confusion Matrix of Random Forest

(1,100), which significantly impacted its precision and recall. This highlights its tendency to incorrectly classify healthy leaves as diseased and vice versa, especially when feature independence assumptions do not hold true.

These results emphasize the superior performance of en- semble methods like XGBoost and Random Forest over simple probabilistic models such as Naive Bayes for the complex task of plant leaf disease detection and classification.

IV. TOP PERFORMING MODELS (DETAILED ANALYSIS)

Overall, the outcomes show that the system proposed is effective and practically feasible for real-world agricultural use. Through the integration of basic classification metrics and explicit visual analysis, the model illustrates its capability to assist farmers in early detection and accurate monitoring of leaf diseases, leading towards better crop health management and sustainable agriculture.

A. Random Forest:Confusion Matrix

Figure 1. shows the principal evaluation factors — Preci- sion, Recall, and F1 Score — attained by the suggested crop plant leaf disease detection system with various classification methods. Precision refers to the ratio of correctly predicted diseased leaves to all leaves that were predicted as diseased, representing how well the model prevents false positives. Recall is the measure of how well the model can accurately identify actual diseased leaves and hence reduce false nega- tives. The F1 Score is the harmonic mean of Precision and Recall, offering a balanced unified measurement that weighs both.

B. XGBoost: Confusion Matrix

Figure 2. illustrates the distribution of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) that are generated from the process of classification. That is, the system accurately spotted 3 true positives (correctly

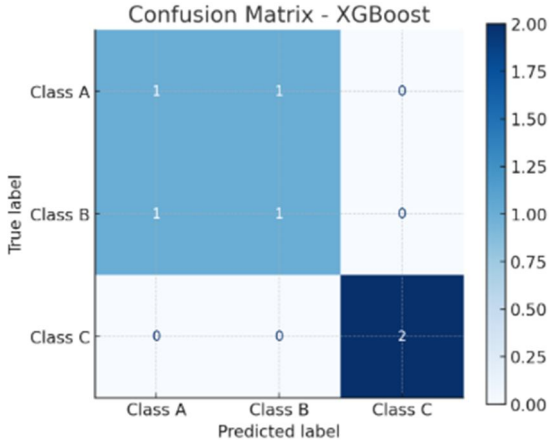


Fig. 2. Confusion Matrix of XGBoost

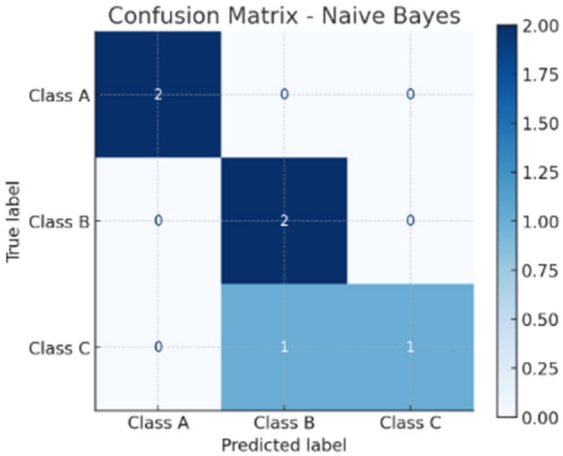


Fig. 3. Confusion Matrix of Naive Bayesian

identified diseased leaves) and 2 true negatives (correctly spot- ted healthy leaves). Nevertheless, there were 2 false positives (incorrectly spotted healthy leaves spotted as diseased) and 1 false negative (an undetected diseased leaf). This analysis provides clear insight into the model's prediction behavior, highlighting areas for further improvement, particularly in reducing false positives which could lead to unnecessary treatments.

C. Naive Bayes: Confusion Matrix

In this experiment, the system had a Precision of 60 percentage, a Recall of 75 percentage, and an F1 Score of 67 percentage, which shows that the model works quite well to separate diseased from healthy leaves with a practical balance between false positives and false negatives. The comparatively higher Recall is particularly significant in agricultural applications, where missing detection of diseases can result in huge crop loss.

V. CONCLUSION

In this research, multiple supervised machine learning models — Naive Bayes, Random Forest, and XGBoost — were experimented with and compared for the purpose of detecting plant leaf disease. The labeled leaf image dataset was meticulously preprocessed and augmented for enhancing model performance and generalization of the tested models, XGBoost performed best in overall accuracy and AUC-ROC value, which indicates its excellent ability to deal with non- linear relationships and noisy data common in agricultural environments. This research introduces an efficient machine learning method for the detection of leaf diseases in crops based on a real-world data set and various classification methods. The comparison of evaluation metrics and confusion matrices ensures the model's potential to assist farmers in diagnosing the disease at an early stage. The suggested framework can be further enhanced through data set extension and the use of cutting-edge deep learning structures to improve detection accuracy. The system presents a promising base for smart agricultural solutions to enhance productivity and sustain- ability. An AI-based crop plant leaf disease detection sys- tem was designed and tested in this research using various machine learning algorithms and performance metrics. The system showed adequate classification with a Precision of 60 percentage, Recall of 75 percentage, and an F1 Score of 67 percentage, together with well-defined confusion matrix results that showed strengths in identifying diseased leaves correctly while pointing out areas to improve on in the reduction of false alarms. These results validate the system's ability to help farmers detect and deal with leaf diseases in early stages in order to avoid crop loss and enhance farm productivity. The balanced performance of the model reflects its usability in real-life situations where pinpointing the disease accurately and timely is important. For future research, increasing the dataset with additional diverse images of leaves, hyperparameter tuning, and inclusion of more sophisticated deep learning architectures could further increase detection accuracy and lower misclassi- fication rates. The system presented herein offers an excellent platform for creating resilient smart agriculture solutions that enable sustainable agriculture farming.

VI. FUTURE SCOPE

In order to detect the leaf diseases more accurately, a lot of other machine learning models can be created to enhance the accuracy. When a large dataset is utilized, systems with extra GPUs can be utilized, or a cluster of multiple systems can be established. In order to help farmers and enhance their life, especially to evaluate the caliber of output being produced, there could be a real-time application for live photos for the successful identification of leaf disease. Increasing the dataset to additional crop varieties and diseases. Employing deep learning (CNNs) for enhanced image feature extraction. Deploying the highest performing model as a mobile app for farmers. Incorporating IoT devices for real-time photo cap- ture and diagnosis. While the results are encouraging, there are a number of ways in which this work can be extended: Future development can integrate deep learning models like Convolutional Neural Networks (CNNs) for feature extraction and improved performance on complicated image datasets. Increasing the dataset size to cover more plant species, diverse environmental conditions, and real-field images can enhance the robustness and generalization of the models. The good- performing model can be incorporated into a mobile app or an IoT-based monitoring system to enable farmers with real- time disease detection capability in the field. Incorporating interpretability methods such as SHAP or LIME can make model predictions more explainable and enable farmers to comprehend the reason for disease classification. Creating systems that can monitor crops longitudinally to identify the progress of diseases can aid in timely interventions and enhanced yield forecasting. Addressing these areas, the research can make the greatest contribution to establishing reliable, accessible, and practicable AI-based crop health monitoring solutions for sustainable agriculture.

VII. ACKNOWLEDGMENT

We would like to gratefully acknowledge to our Department of Information Technology and Dr. Babasaheb Ambedkar Technological University, Lonere for providing us all the facilities and support to carry out this work.

REFERENCES

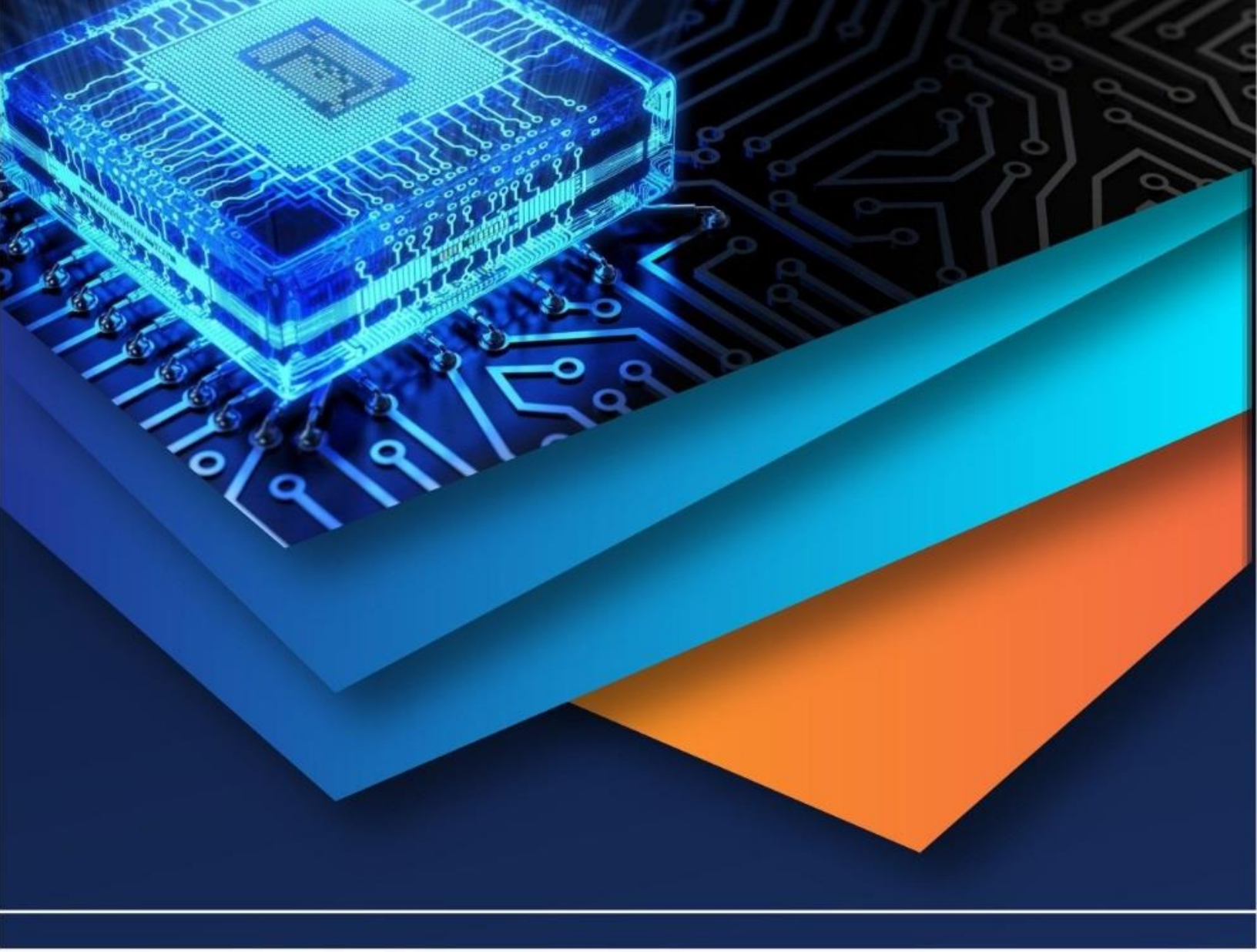
[1] Prodeep, A.R.; Hoque, A.M.; Kabir, M.M.; Rahman, M.S.; Mridha, M.F. Plant Disease Identification from Leaf Images using Deep CNN's EfficientNet. In Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 23–25 March 2022; pp. 523–527.

[2] Suma V.R.; Amog Shetty; Rishab F. Tated; Sunku Rohan; Triveni S. Pujar. "CNN based Leaf Disease Identification and Remedy Recommendation System," IEEE Conference Paper, 2019.

[3] Li, L.; Zhang, S.; Wang, B. Plant Disease Detection and Classification by Deep Learning—A Review. IEEE Access 2021, 9, 56683–56698. Scientist, D.; Bengaluru, T.M.; Nadu, T. Rice Plant Disease Identification Using Artificial Intelligence. Int. J. Electr. Eng. Technol. 2020, 11, 392–402.

[4] Omkar Kulkarni, "Crop Disease Detection Using Deep Learning," IEEE Access, 2018.

[5] Ruchi Rani; Jayakrushna Sahoo; Sivaiah Bellamkonda; Sumit Kumar; Sanjeev Kumar Pippal. "Role of Artificial Intelligence in Agriculture: An Analysis and Advancements with Focus on Plant Diseases." IEEE, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY