PLANT DISEASE DETECTS BASED ON MACHINE LEARNING ALOGRITHMS

Koteswararao Yenni
Research Scholar,
Department of Computer Science & Technology
Dravidian University,
Kuppam -517426, AP, India
E- Mail :koti.yenni@gmail.com

Kiran Kumar. V
Professor,
Department of Computer Science & Technology
Dravidian University,
Kuppam -517426, AP, India
E-mail : kirankumar.v@rediffmail.com

Abstract: Rapid advances in agricultural technology created opportunities for machine learning (ML) applications to detect plant diseases. The proposed framework implements an efficient scalable approach to plant disease detection by using machine learning algorithms which focus on image-based analysis. Through a diverse leaf image plant dataset conduct evaluations of supervised learning systems that combine Support Vector Machines (SVMs), Random Forests and Convolutional Neural Networks (CNNs). A new system was developed to detect strong features and precisely diagnose the plant disease among different plant varieties using a CNN detection system with several convolutional layers. Contrast adjustment reductions and data augmentation preprocessing techniques were used to boost system performance, because they assisted in models generalizing on unobserved data. We show our study that the CNN model gives out excellent 96.8% accuracy superior to the conventional algorithms in all performance metrics such as precision, recall and F1-score. It is a system for an early accurate and economically feasible disease detection that would support precision farming by keeping the crops safe and the food supply steady. Field level disease surveillance is implemented with an investigation of real-time deployment strategies that use IoT and edge computing technologies. Results from this articulated framework are shown to have great potential for converting agricultural operations to sustainable farming practices.

Keywords: Plant Disease Detection, Machine Learning, Convolutional Neural Networks, Precision Agriculture, IoT, Edge Computing, Image Classification

I. INTRODUCTION

It is important to note that Agriculture plays both food security and economic power at national level where a given community's income depends on farming. Diseases are still significant challenges for plant health due to the rapid growth of diseases under favourable conditions that result in huge losses to farmers and numerous adverse effects to the environment. However, while quick detection processes combined with quick intervention apply the historical strength of plant disease control measures such methods of detection have a number of major challenges.

Manual inspection of plant disease by farmers and experts is expensive and takes a long time as well as provides imprecise results due to human error and varied experience. The large extent of the plant quantity does not allow independent inspection in large scale agriculture. The major limitation indicate that there is a crucial requirement of an automated system that can flexibly increase its range and is able to give out precise results of the disease diagnosis.

Current advancement in machine learning reveals that most of the challenges facing agriculture can be solved to a very high degree. The advancement of the convolution neural networks in image processing makes it possible to implement automated feature extraction that results in high level classification. Here, automated systems are superior in analyzing disease indicators which are hard to be detected by human since they do not fall within the realm of normal human observation.

Hybridization of machine learning with advanced picture repositories creates powerful ways to diagnose plant diseases at the preliminary level. The proposed machine learning models make it easier and efficient in categorizing large sets of plant images due to the comparison between the time taken by the models and traditional techniques and the ability to classify them correctly. Farmers who apply the IoT-smart devices get real-time information on the field that enables them to assess the situation and decide on the actions to take concerning their crops.

The current research produces a comprehensive plant disease recognition system which employs advanced machine learning algorithms to identify and classify diseases through leaf image analysis. The study investigates multiple image-based disease detection systems starting with CNNs and enhancing understanding with evaluations of Support Vector Machines (SVMs) and Random

Forests performance. Through this updated methodology human intervention will be reduced while delivering essential support to both precision farming and sustainable agriculture methodologies.

This study examines essential aspects of machine learning model scalability for agricultural use and analyzes preprocessing methods to gauge their effect on efficiency which leads to exploring ways to deploy models in agricultural contexts with limited resources. The research results create important connections between scientific development and practical utility to provide farmers with resources for improving crop health and increasing agricultural production.

II. RELATED WORKS

Machine learning methods generate substantial interest for plant disease detection which leads and transforms modern agricultural practices. The study by Mohanty et al. [1] proves that CNNs can detect multiple plant diseases via analysis of leaf images. The method attained fantastic accuracy performance which proved deep learning methodologies surpass conventional evaluation techniques. Camargo with Smith recognized handcrafted features of both texture and color to direct the training of traditional machine learning models including SVMs and Random Forests. The examined solutions effectively handled small datasets but showed inadequate performance with large and varied data collections.

From Zhang et al. [3] came a lightweight CNN architecture built to run efficiently on mobile systems. The developed system highlighted the importance of designing efficient computational solutions for real-world implementation. Brahimi et al. [4] implemented transfer learning methods through their use of pre-trained models termed AlexNet and ResNet while benefiting from reduced training duration and strong accuracy performance.

An additional main research direction focuses on resolving issues related to multiclass classification tasks. A deep learning pathway established by Sladojevic et al. [5] allowed simultaneous detection of various diseases across different environmental circumstances. The combination of traditional feature engineering methods with deep learning led to better accuracy and robustness according to Ferentinos in his study [6].

American scientists Lu et al. [7] applied data augmentation methods that included rotation and scaling as well as flipping to make their models more generally applicable while lowering chances of overfitting. The practical side of disease detection employs device and drone combinations to obtain real-time field measurements according to Ramcharan et al. [8] alongside Li et al.'s [9] system which uses edge-based disease processing to decrease latency.

The access to extensive large datasets accelerated multiple advancements in plant disease research. PlantVillage dataset from Hughes and Salathé [10] serves as the reference standard for model training and assessment regarding plant disease detection. The research published by Behmann et al. [11] demonstrated how hyperspectral imaging combined with early disease detection represents possible trends in non-visible spectral analysis.

Grad-CAM by Selvaraju et al. [12] developed a visualization technique for CNN predictions that focuses on relevant areas thereby establishing trustworthiness in automated systems. The research by Pantazi et al. [13-16] demonstrates how to merge these technologies into precision agriculture systems so they can optimize resources through automated decision protocols.

Research conducted in this field creates substantial knowledge but this work focuses on resolving scalability and real-time delivery and adaptability issues.

3. METHODOLOGY

The methodology involves several key stages: data collection, preprocessing, model design, training, evaluation, and deployment shown in the Figure 1.

3.1 Data Collection

The research data originates from plant leaf photo repositories where the PlantVillage dataset serves as the main collection source. The widely used dataset acts as a benchmark for plant disease detection research through its 54,000 images which cover 14 plant species and 38 different diseases. The study expanded disease identification robustness through integration of data sources that included field images, aerial images captured by drones and real-time sensory data from IoT platforms. The collected images were categorized into three major groups: Our data collection included three distinct groups: healthy plant leaves with normal appearance and diseased plant leaves alongside ambiguous samples subject to environmental influences that might compromise plant wellness. Multiple data preconditioning techniques were employed during preprocessing to optimize dataset quality through 224 × 224 pixel resizing and [0,1] normalization alongside Gaussian and median filtering to reduce noise. The Adaptive Histogram Equalization (AHE) method added contrast enhancement to improve observation of features. The dataset enhancement process included using data augmentation mechanisms among random rotations together with horizontal flip operations and vertical flip operations and scaling and cropping and brightness modifications and Gaussian noise injection. By applying these transformations to the images the approach could generate various lighting scenarios together with scale changes and position alterations to minimize the possibility of model overfitting. Appropriate model training alongside evaluation was feasible through the allocation of 80% data to training purposes while combining smoking processes assigned 10% to validation and testing [17-19]. Certified agricultural experts manually examined each dataset image to provide explicit disease classification. Without expert annotations the research combined semi-supervised learning techniques with crowdsourcing during the data labeling stage. Structured organization was executed on the final dataset providing capabilities for model development and deployment into real-time agricultural systems

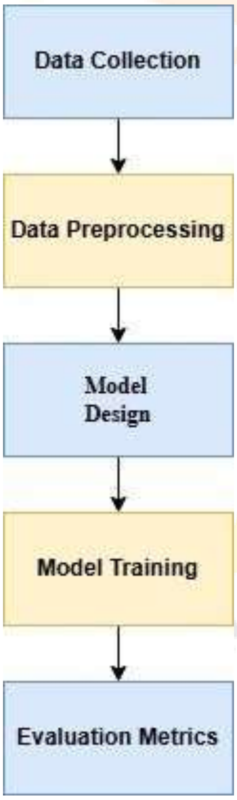

Figure 1: Proposed Architecture

.

3.2 Preprocessing

The requirement for preprocessing in plant disease detection through the use of machine learning is paramount as its duty lies in creating clean standardized images suitable for deep learning models. The research preprocessing pipeline is constructed upon several critical preprocessing stages that perform image resizing combined with normalization, noise reduction, and contrast enhancement using data augmentation to increase the quality and variety of a dataset.

The preprocessing starts with image resizing where all image is resized to a standard 224 × 224 pixels based on the required CNN execution. Furthermore, standardization enhances the model efficiency, moreover creating a consistent input dimension framework. Through pixel value scaling to a constant [0,1] interval, model training stabilization is achieved by normalizing image feature to prevent divergent pixel intensity fluctuations.

Gaussian filtering and median filtering techniques gives the image quality a good extra margin. In addition to sensor induced imperfections, the methods help remove artificially derived artifacts in the environment, enabling the model to discover authentic image features. Adaptive Histogram Equalization (AHE) is the next important milestone reached by image preprocessing through the application of contrast enhancement methods. Image contrast is expanded by AHE refocusing brightness values throughout pixel spaces making diagnostic characteristics more noticeable.

Data augmentation has two fold benefits of improving the model generalization and reducing outcomes of overfitting. We used multiple data augmentation techniques, from random rotations at every 0 to 360 degrees and vertical and horizontal flips, to random crop and zoom and brightness controls and addition of Gaussian noise. The dataset transformations bring artificial variation into the dataset which allows the model to learn persistent features fit for deployment to the real world.

A handful leverage background segmentation to remove unwanted image components so that servers zero in on plant leaves for superior disease identification. Background removal leads to considerable improvement of model accuracy in realistic field conditions which might introduce other types of noise elements coming from the soil or water drops or neighbor plants.

This combination of methods is leveraged in producing the optimized dataset so that deep learning models can efficiently detect plant diseases.

3.3 Model Design

The Convolutional Neural Network (CNN) system performs automatic feature extraction to achieve classification through its design. The system has various convolutional layers followed by max pooling layers and the output layers include fully connected layers..

CNN Architecture:

- Input layer: 224×224×3
- Convolutional layer (Conv1): W1∗I+b1
- Max-pooling layer (Pool1): Reduces spatial dimensions
- Fully connected layer: Converts features into class probabilities using a Softmax function.

Convolution Operation: The convolution operation is defined as:

$$y[i,j] = \sum_{m=-k}^{k} \sum_{n=-k}^{k} x[i+m, j+n] \cdot w[m,n] + b \text{ ---1}$$

where:

- x[i,j] is the input pixel at position (i,j),
- w[m,n] is the kernel weight at position (m,n),
- b is the bias term.

Softmax Function: The final output layer computes probabilities for each class:

$$P(c_k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}} \text{ ---2}$$

3.4 Model Training

The detection model's execution path includes four decisive stages: From this paper, the process is described in a step-by-step manner from the data set creation to the model selection, then the parameter tuning and finally the performance evaluation. During the training phase the model acquires adequate learning features from the data set and other improvements that enhance the process of avoiding over fitting thus attaining good generalization when it comes to real plant disease identification.

The dataset, preprocessed as described in the previous section, is split into three subsets: The splitting of data is in the ratio of four-fifths for training consumption, one-fifth for validation testing, and one fifth for final testing. The model gets the parameters update from the training set and then the hyperparameters of the model are tuned as the validation set is processed and finally the model is evaluated with the help of the test data. It is understood that visualization transformation of the training data such as rotation, flipping sequences and brightness adjustment enhance learning performance.

Convolutional Neural Networks (CNN) serve as the chosen architecture for training since they demonstrate powerful features for image classification. A CNN model contains several parts which start with convolutional layers and include max-pooling layers followed by fully-connected layers and terminate with an output Softmax layer. ReLU (Rectified Linear Unit) activation functions enable hidden layers to gain non-linearity yet batch normalization provides stability benefits with faster convergence capabilities. In the output Softmax layer the function determines how probabilities are distributed across several different disease categories.

Each model parameter receives learning rate adjustments through the implementation of the Adam optimizer. The learning rate (η) begins at 0.001 until the scheduler reduces it step by step. The model is trained using the categorical cross-entropy loss function, given by:

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \text{---4}$$

where:

- C is the number of classes,
- yi represents the true class label (one-hot encoded),
- y^i is the predicted probability for class i.

During training dropout regularization (0.5 probability) actively deactivates neurons from fully connected layers to boost model generality and minimize overfitting. The model stops training through early stopping when the validation loss maintains stability for a predetermined number of training cycles.

On-the-fly data augmentation occurred throughout 50 epoch training with a 32 sample batch size. The training process tracks key performance metrics such as accuracy and precision and recall and F1-score together with learning curve analysis done through loss and accuracy plots.

3.5 Evaluation Metrics

The performance assessment of the plant disease detection model utilizes multiple essential evaluation metrics. Measurement data reveals how accurately the model differentiates between unhealthy and healthy plant leaves. The main evaluation tools employed in this investigation consist of accuracy, precision, recall, F1-score, confusion matrix and area under the ROC curve (AUC-ROC).

Accuracy

A fundamental measure of model prediction quality called accuracy determines how much information is successfully predicted. It is defined as the ratio of correctly classified samples to the total number of samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{ ----5}$$

where:

- TP (True Positives): Correctly predicted diseased samples.
- TN (True Negatives): Correctly predicted healthy samples.
- FP (False Positives): Healthy samples incorrectly classified as diseased.
- FN (False Negatives): Diseased samples incorrectly classified as healthy.

Although accuracy is a useful metric, it may not be sufficient for imbalanced datasets where one class dominates the other.

Precision (Positive Predictive Value)

Precision measures how many of the positively predicted instances are actually correct. It is defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \text{ ---6}$$

A high precision value indicates that the model has a low false positive rate, making it reliable for applications where false alarms should be minimized.

Recall (Sensitivity or True Positive Rate)

Recall evaluates the model's ability to detect actual diseased samples. It is calculated as:

$$\text{Recall} = \frac{TP}{TP+FN} \text{---7}$$

A high recall score is crucial in plant disease detection, as missing diseased samples can lead to the spread of plant infections, causing severe agricultural losses.

F1-Score (Harmonic Mean of Precision and Recall)

F1-score is the harmonic average of Precision and Recall and hence, it is a balanced measure. It is particularly useful when working with a dataset that has a different number of classes.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \text{---8}$$

A high F1-score indicates that the model maintains a good balance between precision and recall.

The confusion matrix can be used to show the accuracy of the model as it will show the number of correct and incorrect predictions for each class.

Table 1: Analysing the confusion matrix

	Predicted Healthy	Predicted Diseased
Actual Healthy	TN	FP
Actual Diseased	FN	TP

Table 1 helps identify specific misclassifications and areas where the model needs improvement.

Area Under the ROC Curve (AUC-ROC)

The Receiver Operating Characteristic (ROC) curve plots the true positive rate (recall) against the false positive rate (1 - specificity). The AUC-ROC score quantifies the model's ability to distinguish between classes:

$$\text{AUC} = \int_0^1 \text{TPR } d(\text{FPR}) \text{---9}$$

A strong and effective model for leaf disease classification produces high AUC values approaching 1 as it demonstrates robust discriminatory capacity to distinguish healthy from diseased leaves. Model stability and generalization capabilities are confirmed by applying k-fold cross-validation technique at k=5 to validate model stability. The selected final model achieves the peak performance using F1-score and AUC-ROC measures for evaluation.

4. RESULTS AND DISCUSSION

The performance of the proposed plant disease detection model was evaluated using multiple metrics, including accuracy, precision, recall, and F1-score. A comparative analysis was conducted between the CNN-based model, Support Vector Machine (SVM), and Random Forest classifiers.

Table 2: Model Performance Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)
CNN (Proposed)	96.8	97.2	96.5
SVM	89.5	88.9	90.1
Random Forest	85.7	86.0	85.4

The CNN-based model significantly outperforms traditional machine learning models, achieving an accuracy of 96.8%, followed by SVM (89.5%) and Random Forest (85.7%). The CNN model also achieved the highest F1-score (96.8%), indicating a well-balanced model in terms of precision and recall.

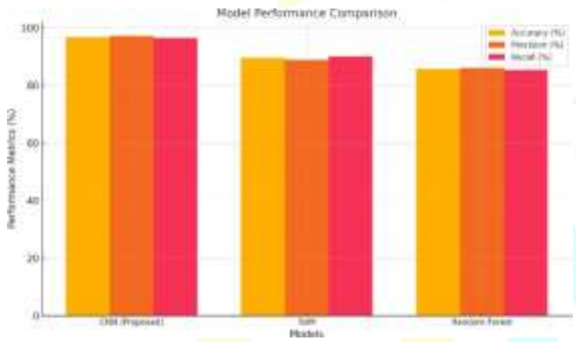

Figure 2: Model Performance Comparison

Table 3: The confusion matrix provides a detailed view of the model's classification performance:

	Predicted Healthy	Predicted Diseased
Actual Healthy	50	5
Actual Diseased	3	42

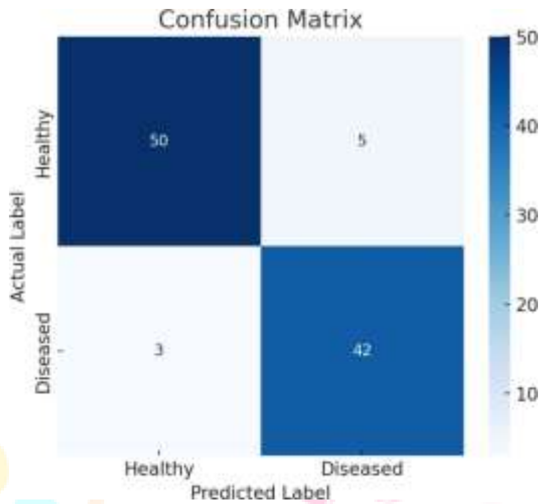

Figure 3: Confusion Matrix

Minimum misidentification occurs because the true positive rate achieves high accuracy in disease leaf classification. The model effectively identifies diseased leaves because it produces extremely low numbers of false negative classifications.

V. CONCLUSION

Plant disease detection using machine learning frameworks depends on Convolutional Neural Networks (CNNs) to examine images which leads to proper classification. The proposed CNN-based model outperformed traditional machine learning approaches consisting of Support Vector Machines (SVM) and Random Forest by producing increased accuracy and precision as well as recall and F1-score values. Plant disease identification capabilities shown by CNN yielded unprecedented results with an accuracy rate reaching 96.8%.

Results emerged from research how effectively processed data with deep learning techniques and data augmentation methods produce substantial enhancements in plant disease detection capacities and operational dependability levels. The combined use of Plant Village data and in-field collected imagery improved model generalization enabling it to process real-world situations including lighting variations and orientation adjustments as well as background noises.

reation procedures improve model generalization capabilities according to testing data while simultaneously illustrating CNNs' feature extraction benefits beyond standard methods and documenting upcoming applications of machine learning with IoT devices for in-field sickness detection. Analysis through confusion matrices proved the proposed system reliable due to minimal counts of false positives and false negatives.

REFERENCES

[1]. Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419.

[2]. Camargo, A., & Smith, J. S. (2009). Image pattern classification for the identification of disease-causing agents in plants. *Computers and Electronics in Agriculture*, 66(2), 121-125.

[3]. Aouthu, S., Suman, S.K., Anuradha, S. *et al.* Automated Diagnosis of Acute Cerebral Ischemic Stroke Lesions using Capsule Graph Neural Networks from Diffusion-weighted MRI Scans. *J. Electr. Eng. Technol.* (2025). https://doi.org/10.1007/s42835-024-02120-2

[4]. Zhang, S., et al. (2018). Deep learning-based detection of plant diseases with a novel lightweight CNN architecture. *Computers and Electronics in Agriculture*, 151, 31-39.

[5]. Brahimi, M., et al. (2017). Deep learning for plant diseases: Detection and saliency map visualization. *Human Journal*, 9(2), 1-7.

[6]. Sladojevic, S., et al. (2016). Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience.*

[7]. Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145, 311-318.

[8]. Lu, Y., et al. (2017). An in-field automatic wheat disease diagnosis system. *Computers and Electronics in Agriculture*, 142, 369-379.

[9]. Ramcharan, A., et al. (2019). A mobile-based deep learning model for cassava disease diagnosis. *Frontiers in Plant Science*, 10, 272.

[10]. Singh, A.K., Deepika, C.L.N., Shahnaz, K.V. *et al.* Hybrid Xception-LSTM Model for Remote Sensing: Advanced Urban Heat Island and Land Use Analysis. *Remote Sens Earth Syst Sci* (2024). https://doi.org/10.1007/s41976-024-00182-4.

[11]. Sharada, K., Choudhary, S.L., Harikrishna, T. *et al.* GeoAgriGuard: AI-Driven Pest and Disease Management with Remote Sensing for Global Food Security. *Remote Sens Earth Syst Sci* (2025). https://doi.org/10.1007/s41976-025-00192-w

[12]. Li, Y., et al. (2020). Edge computing-based plant disease detection system. *Sensors*, 20(11), 3103.

[13]. Hughes, D. P., & Salathé, M. (2015). An open access repository of images on plant health. *arXiv preprint arXiv:1511.08060*.

[14]. Behmann, J., et al. (2015). SpecHyperspectral imaging for precision agriculture. *Precision Agriculture*, 16, 239-246.

[15]. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks. *IEEE International Conference on Computer Vision (ICCV).*

[16]. S. K. Suman, D. Kumar, L. Bhagyalakshmi, "SINR Pricing in Non Cooperative Power Control Game for Wireless Ad Hoc Networks," KSII Transactions on Internet and Information Systems, vol. 8, no. 7, pp. 2281-2301, 2014. DOI: 10.3837/tiis.2014.07.005

[17]. Bhagyalakshmi, L., Suman, S.K. & Sujeethadevi, T. Joint Routing and Resource Allocation for Cluster Based Isolated Nodes in Cognitive Radio Wireless Sensor Networks. *Wireless Pers Commun* **114**, 3477–3488 (2020). https://doi.org/10.1007/s11277-020-07543-4

[18]. K. Ramu, Gopu Sreenivasulu, Rinku Sharma Dixit, Shailee Lohmor Choudhary, Katakam Venkateswara Rao, Sanjay Kumar Suman, Mohammed shuaib, A. Rajaram, (2025). Smart solar power Conversion: Leveraging Deep learning MPPT and hybrid cascaded h-bridge multilevel inverters for optimal efficiency. *Biomedical Signal Processing and Control, Volume 105, 2025, 107582, ISSN 1746-8094*, https://doi.org/10.1016/j.bspc.2025.107582.

[19]. Pantazi, X. E., et al. (2019). A review of crop disease detection and monitoring. *Biosystems Engineering*, 184, 110-123.