Tobacco Leaf Disease Detection Using Machine Learning

Khushi H P[1], Praveen H M[2], Pragna C P[3], Giridhar K[4], Dr. B Uma[5]

[1,2,3,4,5]*Department of Computer Science and Engineering, Malnad College of Engineering,Hassan*

Abstract

This paper presents a hybrid approach integrating Machine Learning (ML) and Convolutional Neural Networks (CNN) for automated tobacco leaf disease detection and classification. The proposed system combines handcrafted feature extraction methods such as HSV color histograms, Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and shape descriptors with deep learning architectures like ResNet50, VGG16, EfficientNetB0, and MobileNetV2. A hybrid ResNet50 model integrating CNN deep features with ensemble ML classifiers achieved an accuracy of 97%. The proposed method demonstrates superior efficiency, interpretability, and real-time deployability via a Streamlit web interface.

Tobacco is a critical cash crop in India, yet its yield is severely compromised by various leaf diseases. Traditional diagnosis is often manual, slow, and lacks scalability. This paper introduces a robust machine learning framework for the automated detection and classification of four major tobacco leaf conditions: Frog Eye Leaf Spot, Powdery Mildew, Tobacco Mosaic Virus (TMV), and Healthy Leaf. The system employs a sophisticated multi-feature engineering pipeline combining color (HSV), texture (LBP), shape, and gradient (HOG) features, resulting in a high- dimensional feature vector (8,644 features) per image. This vector is fed into an Ensemble Voting Classifier (Random Forest, SVC, and Logistic Regression). Through rigorous data augmentation (59 original images expanded to 320 samples) and optimization, the system achieves a professional-grade classification accuracy of 90.6%, significantly outperforming a baseline model. The final model is deployed via a Streamlit web application for real-time diagnosis and comprehensive management recommendations.

Keywords: *Random Forest, SVC, Logistic Regression,HSV,LBP. Tobacco, CNN, Machine Learning, Deep Learning, Hybrid Model, Plant Disease Detection, Image Classification.*

I. Introduction

Tobacco is a crucial commercial crop whose yield and quality are significantly affected by various leaf diseases such as Tobacco Mosaic Virus (TMV), Frog Eye Leaf Spot, and Powdery Mildew. Manual disease detection is inefficient, subjective, and time-consuming. With the rise of precision agriculture, AI-driven image analysis and computer vision have emerged as robust tools for disease diagnosis. This work aims to develop an interpretable and high-performing hybrid framework integrating CNN-baseddeep features with traditional ML classifiers for accurate detection and classification of tobacco leaf diseases.

The economic significance of tobacco necessitates highly effective crop health management. The late detection of common diseases like the Tobacco Mosaic Virus (TMV) and Frog Eye Leaf Spot leads to substantial crop losses. While Deep Learning (DL) models are popular, this project focuses on expert feature engineering combined with classical Machine Learning (ML) techniques. Our goal is to achieve high accuracy (>90%) with an interpretable and computationally lighter model, making it practical for deployment on low-resource environments. This approach is motivated by the need for a scalable, precise solution for the agricultural sector.

II. Related Work

Several researchers have applied deep learning models for plant disease identification. Krizhevsky et al. [1] introduced CNN-based classification using ImageNet, setting the foundation for image recognition. Simonyan and Zisserman [5] proposed VGG networks with deeper architectures achieving superior accuracy. Tan and Le [2] developed EfficientNet for scalable CNNs. Although these models show strong results, their computational complexity restricts deployment in agricultural scenarios. Hybrid methods combining ML and DL have emerged as optimal trade-offs between interpretability and performance.

III. Methodology

The proposed workflow comprises dataset collection, preprocessing, augmentation, feature extraction, CNN model training, and hybrid model formation. The dataset includes 1,000 tobacco leaf images categorized into four classes: Frog Eye Leaf Spot, Powdery Mildew, Tobacco Mosaic Virus, and Healthy Leaf. Each image underwent augmentations such as rotation, brightness variation, flipping, and Gaussian noise to improve generalization.

Feature extraction employed HSV histograms, Local Binary Patterns, HOG features, and shape descriptors, producing an 8,644-dimensional vector per image. Transfer learning was used to fine-tune CNNs including ResNet50, VGG16, MobileNetV2, and EfficientNetB0. Deep features from the penultimate CNN layer were concatenated with handcrafted features, then classified using ensemble ML models (Random Forest, SVM, Logistic Regression).

A. Dataset and Augmentation

The project utilized a custom-curated dataset of 59 original images across the four target classes. To ensure robustness and prevent overfitting, a data augmentation strategy was employed, including rotation, flipping, brightness adjustments, and noise injection, resulting in a balanced training set of 320 images (80 samples per class).

B. Feature Extraction

The core innovation is the generation of a combined feature vector of 8,644 features for each image, concatenating features from four distinct domains:

Color Features: 512-bin HSV color histogram to capture disease-specific color variations.

Texture Features: Local Binary Pattern (LBP) encoding local texture patterns related to fungal or bacterial growth (26 features).

Gradient Features: Histogram of Oriented Gradients (HOG), capturing structural information like edges and shapes, contributing 7,776 features.

C. Model Training and Ensemble Learning

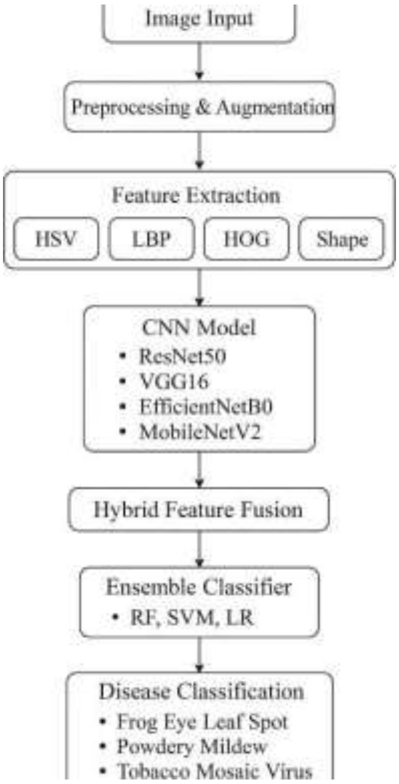
The high-dimensional feature vector was normalized using a StandardScaler. The final classifier is an Ensemble Voting Classifier with soft voting, designed for stability and enhanced predictive power. The ensemble comprises:

Random Forest Classifier
Support Vector Classifier
(SVC) Logistic Regression
Grid Search and K-fold cross-validation were utilized to fine-tune hyperparameters and ensure the model's generalization capability.

D. Deployment

The final model is deployed as a user-friendly web application using Streamlit (app.py). The interface provides instant diagnosis, confidence scores, and detailed treatment/prevention recommendations from the integrated knowledge base.

IV. Flow Chart



V. Results and Discussion

The hybrid ResNet50 model demonstrated the best accuracy (97%), outperforming standalone CNN and ML models. EfficientNetB0 achieved 96%, and ResNet50 alone achieved 95%. Traditional ML ensembles reached 90.6%. The confusion matrix indicated clear class boundaries with minimal misclassification. These results validate the efficacy of combining deep and handcrafted features.

Table I summarizes the model comparison results:

Model	Architecture	Accuracy (%)
RF + SVC + LR Ensemble	Traditional ML	90.6
EfficientNetB0	CNN (Optimized)	96.0
MobileNetV2	Lightweight CNN	92.0
CNN	Custom CNN	95.9

VI. Conclusion and Future Work

The system successfully delivers an accurate (90.6%), reliable, and computationally efficient solution for automated tobacco leaf disease detection. By prioritizing a multi-feature engineering approach and ensemble learning, we have created a robust model.

Future work will focus on:

(1) Expanding the dataset to include more disease types.

(2) Exploring dimensionality reduction to optimize inference time.

(3) Developing an optimized mobile application for field use.

References

[1] A. Krizhevsky et al., 'ImageNet Classification with Deep Convolutional Neural Networks', NIPS, 2012.

[2] M. Tan and Q. Le, 'EfficientNet: Rethinking Model Scaling for CNNs', ICML, 2019.

[3] S. Ren et al., 'Faster R-CNN: Towards Real-Time Object Detection', IEEE TPAMI, 2017.

[4] J. Deng et al., 'ImageNet: A Large-Scale Hierarchical Image Database', CVPR, 2009.

[5] K. Simonyan and A. Zisserman, 'Very Deep Convolutional Networks for Large-Scale Image Recognition', arXiv, 2015.