


Machine Learning-Driven Identification of Cotton Leaf Diseases for Precision Agriculture

Tushar Mohite Patil¹, Sanjay Pandey², Ravindra Duche³

¹Research Scholar,^{2,3} Professor,  [0009-0007-6695-2644](https://orcid.org/0009-0007-6695-2644)
Department of Physics/Electronics, ISBM University, Nawapara, Chhattisgarh, India

Email of Corresponding Author: tvmohitepatil@vpmthane.org

Received on: 9 May, 2025

Revised on: 09 June, 2025

Published on: 10 June, 2025

Abstract – In this study, we propose a machine learning-based method for automatic detection of cotton leaf diseases based on Random Forest classifier. Other features extracted are color based (RGB) and texture based (GLCM) that helps a lot in increasing the classification accuracy. It yielded a 92.5% accuracy rate, which indicates that combining these features was the right way to go! Class imbalance and similar looking diseases were tackled using data augmentation. Further work comprises applying deep learning techniques and IoT real-time based monitoring for precision agriculture. Our findings underline the power of machine learning for better diagnosis of cotton diseases and for achieving a sustainable economy.

Keywords- Cotton leaf disease, Machine learning, Random Forest, Feature extraction, Plant Village

I. INTRODUCTION

Prediction and control of cotton leaf diseases are vital issues that have a considerable impact on agricultural economics and sustainability, especially in light of the increasing worldwide demand for cotton. So are cotton plants, they are vulnerable to many diseases, in particular the cotton leaf curl disease (CLCuD), a disease complex caused by begomoviruses that can cause significant yield losses if not accurately diagnosed and controlled (Mubin et al., 2010). Conventional approaches to detecting disease, which rely on direct assessment and subject matter expertise, can be labor- and time-intensive, prone to error, and susceptible to the manifestations of variability among field-grown plants, where symptoms can be confounded with those of healthy plants. Emergence of machine learning (ML)

technologies gives a ray of hope to tackle these agricultural problems. During the cotton germination period, fungal infection is a significant issue in agriculture, affecting the quality and yield of the crops. Deep learning has recently gained tremendous attention for accurate classification of the health status of cotton leaves using convolutional neural networks (CNNs). These networks have shown an accuracy of above 90% for identifying different diseases in Cotton (Kumar et al., 2020; Noon et al., 2021). Such a potential opens the door to deploying ML models in real-world scenarios whereby timely interventions can help escape considerable revenue losses (Sharif et al., 2018; Barbedo, 2018). There was the emergence of several algorithms in the literature that offer various options for researchers and practitioners. The classifiers like Support Vector Machine (SVM) have been proven to be used alongside advanced imaging techniques for discrimination between diseased and healthy plant tissues (Mehmood et al., 2023; Hyder & Talpur, 2024). On the other hand, new methods such as bilinear coordinate attention models and hyperspectral imaging have also been applied to improve disease recognition performance considering the challenges brought by complicated background images of leaves (Shao et al., 2022; Noon et al., 2021). You have up to date on evidence for these technologies until October 2023. Furthermore, such knowledge of pathology on the molecular level is very important in the next and most vital step of improving the disease prediction models being used. Big data machine learning techniques can be better explored with the plant disease prediction by using genetic marker information from the hosts. Familiarity with the activity of particular virulence determinants, like the function of the β C1 protein in disease dynamics, establishes the framework for development of targeted measures (Tahir et al. 2011; Saeed et al. 2015). By blending concepts from plant-microbe interactions with advancements in machine learning techniques, this case

study also reflects the importance of multidisciplinary collaborations in improving disease detection systems for cotton farming. So overall, ML techniques to enhance the detection of diseases in cotton leaves has a lot of potential to revolutionize agricultural management practices. With the technology advancing, it is become more and more likely that this could lead to more precise accuracy and quicker responses to plant health threats. Training will focus on data until 2024. This research will help to develop and optimise systems that support these innovations, resulting in effective and resilient strategies for cotton production that can cope with current disease pressures.

II. LITERATURE REVIEW

Machine learning is an increasing essential tool in agriculture in recent years, which can be applied to the monitoring and forecasting of plant diseases, for example cotton diseases. Well-established cotton leaf diseases such as cotton leaf curl disease and others are rising too fast to ignore, demanding competent techniques right in time to save high yield and sustainable agriculture with minimal cost. A large number of studies demonstrate remarkable progress in this application using various machine learning techniques. Kumar et al. describe a baseline study where they adopt deep learning architectures to classify cotton diseases by using a convolutional neural network (CNN) to automate the disease plant detection from leaf images. It also highlights how such neural architectures can correctly classify cotton leaf disease with an accuracy level of up to 100% (Kumar et al., 2020). Likewise, Liang's work combines metric learning and few shot-learning approaches to effectively classify cotton leaf spot diseases. They segmented spots of disease, then used classical CNNs to improve detection accuracy (Liang, 2021). Earlier efforts like those by Pujari et al. (2016) also showcased the strength of texture-based analysis in fungal disease detection through classical image processing techniques. This innovation also extends to Shao et al.'s method, which obtains amplified disease identification via a bilinear coordinate attention mechanism. This approach avoids losing valuable feature details due to complex backgrounds that are generally contained in images of infected cotton leaves, which also indicate that achieving precise feature extraction is important in machine learning systems (Shao et al., 2022). Furthermore, dataset characteristics such as size and color space significantly influence model performance in deep learning-based plant disease detection systems, as highlighted by Barbedo (2018). Additionally, Kumar et al. using multiple algorithms, including support vector machines (SVM) and random forests, in a comparative analysis, other studies have indicated the capabilities of several models for the identification of both organic and non-organic cotton diseases (Kumar et al., 2021). The comparison also helps improve understanding of the strengths and weaknesses

of various algorithms on cotton disease prediction. Random Forest classifiers, known for their robustness, have been effectively applied in plant disease recognition, including citrus crops, leveraging texture features (Sharif et al., 2018). Other papers have focused on integrating machine learning and computer vision, as well as other approaches. Jajja et al. (2020) used image processing coupled with SVM classifiers to provide better disease detection capabilities (Bhimte & Thool, 2018). compared several deep learning models to traditional methods and found that deep learning frameworks outperformed traditional machine learning algorithms in this context (Jajja et al., 2022). Additionally, Sharma et al. highlights a spectrum of machine learning applications in precision agriculture that are important to help monitor crop health and optimize agricultural practices (Sharma et al., 2021). Less about a specific technique than on a general level, the implications of machine learning are a real paradigm shift for agriculture as a whole. Machine learning techniques have been shown to enhance disease prediction accuracy as well as the global efficiency of agriculture itself, as reviewed in detail outlining the challenge ahead and potential for integration of such methods in agriculture (Benos et al., 2021; Araújo et al., 2023). Furthermore, the literature highlights the urgency for the availability of large, well-annotated datasets, which is essential for effectively training machine learning models. Zhang et al. and other scholars note that diverse datasets including healthy and diseased plant images are important to further improve model accuracy and generalizability under different conditions (Benos et al., 2021; Zhang et al., 2021). Conventional methods of data collection, acquisition, cleaning, formatting, and storage largely contribute to the generalization of machine learning applications in agriculture (Chen et al., 2021; Sharma et al., 2023).

III. METHODOLOGY

a. Dataset collection and preprocessing

The dataset used in this study was obtained from the PlantVillage repository and consisted of images classified into four categories: Bacterial Blight, Alternaria Leaf Spot, Cotton Leaf Curl Virus, and Healthy Leaves. Each image was pre-labeled to ensure accurate class representation. To enhance reliability, dataset annotations were cross-verified with botanical disease databases, and quality assessments were further refined through consultations with agronomy experts within our team. This validation process significantly improved the accuracy and credibility of the dataset. Before training the machine learning model, multiple preprocessing steps were applied to ensure uniformity and enhance model performance. All images were resized to 224×224 pixels for consistency and normalized by scaling pixel values to a range between 0

and 1 to facilitate faster model convergence. Gaussian and median filtering were used to reduce noise and enhance image clarity. Additionally, data augmentation techniques—including rotation ($\pm 20^\circ$), horizontal and vertical flipping, brightness adjustments, and zooming—were employed to increase dataset diversity and mitigate the risk of overfitting.

b. Feature Extraction Methods

Feature extraction is a crucial step in cotton leaf disease prediction, where relevant characteristics are derived from images to enhance the performance of machine learning models. The following techniques were used for feature extraction:

i. Color-Based Features

Color histograms were extracted in different color spaces such as RGB and HSV. The histogram represents the frequency distribution of pixel intensities across the image (Barbedo, 2018)[21]. For a given image $I(x, y)$ with pixel intensities in channel c , the normalized histogram $H_c(k)$ is computed as:

$$H_c(k) = \frac{N_k}{N} \tag{1}$$

ii. Gray-Level Co-occurrence Matrix (GLCM)

GLCM represents the spatial relationship between pixel intensities at a given offset (d, θ) . The matrix element $P(i, j)$ is defined and from GLCM, statistical texture descriptors such as **contrast, correlation, energy, and homogeneity** were computed[22]:

Contrast: $C = \sum_{i,j} P(i, j)(i - j)^2$ (2)

Correlation: $R = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)P(i, j)}{\sigma_i \sigma_j}$ (3)

Energy: $E = \sum_{i,j} P(i, j)^2$ (4)

Homogeneity: $H = \sum_{i,j} \frac{P(i, j)}{1 + |i - j|}$ (5)

c. Classification

Random Forest consists of an ensemble of N decision trees, where each tree is trained on a random subset of the dataset. The RF classifier is built using Bootstrap Aggregation (Bagging) [23]:

- Random Sampling: Each tree is trained on a random subset of the dataset.
- Feature Selection: At each node split, a random subset of features is considered.

- Majority Voting: The final prediction is obtained by aggregating outputs from all trees.

Mathematically, the RF classifier can be represented as:

$$f(X) = \frac{1}{N} \sum_{t=1}^N h_t(X) \tag{6}$$

Each tree follows a decision rule based on Gini impurity

$$G = 1 - \sum_{i=1}^c p_i^2 \tag{7}$$

IV. Results and Discussion

The Random Forest classifier's performance was assessed with respect to the accuracy, precision, recall, and F1 metrics. The model trained on the cotton leaf disease dataset and was tested using an unseen validation set. Performance metrics are presented in Table 1. Classification accuracy was improved considerably with the subsequent addition of Color-Based Features and GLCM (texture) The fact that removing each of the features set both resulted in a 2-5% decrease in accuracy further reveals their relevance. The model had trouble with diseases that looked similar (e.g. Alternaria Leaf Spot vs. Bacterial Blight). The results were affected by class imbalance (because there were significantly less healthy leaf images); using data augmentation techniques can improve performance.

Table 1: Performance Matric

Metric	Value (%)
Accuracy	92.5
Precision	91.8
Recall	93.2
F1-Score	92.5

V. Conclusion and future scope

A Random Forest-based model for cotton leaf disease prediction was developed through color-based features (RGB/HSV histograms) and GLCM (Gray Level Co-occurrence Matrix) texture features, achieving an accuracy of 92.5%. GLCM features provided the extra discrimination power by capturing the texture patterns in the image, while color features captured the pigmentation specific to diseases. The full integration of both features improved accuracy, demonstrating they were complementary. Future work involves integrating deep learnings, hybrid models, data augmentation, and real-time applications to land monitoring in precision agriculture. The incorporation of explainable AI (XAI) and IoT based disease monitoring will be of even more practical usability for crop protection.

REFERENCES

- [1] Hyder, U. and Talpur, M. (2024). Detection of cotton leaf disease with machine learning model. *Turkish Journal of Engineering*, 8(2), 380-393. <https://doi.org/10.31127/tuje.1406755>.
- [2] Kumar, K., Chandra, G., & Sukheja, D. (2020). Cotton disease detection using deep learning. *International Journal of Innovative Technology and Exploring Engineering*, 9(4), 152-156. <https://doi.org/10.35940/ijitee.d1391.029420>.
- [3] Mehmood, S., Memon, F., Nighat, A., Memon, F., & Saba, E. (2023). Comparative analysis of feature extraction methods for cotton leaf diseases detection. *Vfast Transactions on Software Engineering*, 11(3), 81-90. <https://doi.org/10.21015/vtse.v11i3.1626>.
- [4] Mubin, M., Amin, I., Amrao, L., Briddon, R., & Mansoor, S. (2010). The hypersensitive response induced by the v2 protein of a monopartite begomovirus is countered by the c2 protein. *Molecular Plant Pathology*, 11(2), 245-254. <https://doi.org/10.1111/j.1364-3703.2009.00601.x>
- [5] Noon, S., Amjad, M., Qureshi, M., & Mannan, A. (2021). Computationally light deep learning framework to recognize cotton leaf diseases. *Journal of Intelligent & Fuzzy Systems*, 40(6), 12383-12398. <https://doi.org/10.3233/jifs-210516>.
- [6] Saeed, M., Briddon, R., Dalakouras, A., Krczal, G., & Wassenegger, M. (2015). Functional analysis of cotton leaf curl kokhran virus/cotton leaf curl multan betasatellite rna silencing suppressors. *Biology*, 4(4), 697-714. <https://doi.org/10.3390/biology4040697>.
- [7] Shao, M., He, P., Zhang, Y., Zhou, S., Zhang, N., & Zhang, J. (2022). Identification method of cotton leaf diseases based on bilinear coordinate attention enhancement module. *Agronomy*, 13(1), 88. <https://doi.org/10.3390/agronomy13010088>.
- [8] Tahir, M., Amin, I., Briddon, R., & Mansoor, S. (2011). The merging of two dynasties—identification of an african cotton leaf curl disease-associated begomovirus with cotton in pakistan. *Plos One*, 6(5), e20366. <https://doi.org/10.1371/journal.pone.0020366>.
- [9] Araújo, S., Peres, R., Ramalho, J., Lidon, F., & Barata, J. (2023). Machine learning applications in agriculture: current trends, challenges, and future perspectives. *Agronomy*, 13(12), 2976. <https://doi.org/10.3390/agronomy13122976>.
- [10] Benos, L., Tagarakis, A., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: a comprehensive updated review. *Sensors*, 21(11), 3758. <https://doi.org/10.3390/s21113758>.
- [11] Bhimte, N. and Thool, V. (2018). Diseases detection of cotton leaf spot using image processing and svm classifier., 340-344. <https://doi.org/10.1109/iccons.2018.8662906>.
- [12] Chen, Z., Goh, H., Sin, K., Lim, K., Chung, N., & Liew, X. (2021). Automated agriculture commodity price prediction system with machine learning techniques. *Advances in Science Technology and Engineering Systems Journal*, 6(4), 376-384. <https://doi.org/10.25046/aj060442>.
- [13] Jajja, A., Abbas, A., Khattak, H., Niedbala, G., Khalid, A., Rauf, H., & Kujawa, S. (2022). Compact convolutional transformer (cct)-based approach for whitefly attack detection in cotton crops. *Agriculture*, 12(10), 1529. <https://doi.org/10.3390/agriculture12101529>
- [14] Kumar, K., Chandra, G., & Sukheja, D. (2020). Cotton disease detection using deep learning. *International Journal of Innovative Technology and Exploring Engineering*, 9(4), 152-156. <https://doi.org/10.35940/ijitee.d1391.029420>.
- [15] Kumar, S., Jain, A., Shukla, A., Singh, S., Raja, R., Rani, S., & Masud, M. (2021). A comparative analysis of machine learning algorithms for detection of organic and nonorganic cotton diseases. *Mathematical Problems in Engineering*, 2021, 1-18. <https://doi.org/10.1155/2021/1790171>.
- [16] Liang, X. (2021). Few-shot cotton leaf spots disease classification based on metric learning. *Plant Methods*, 17(1). <https://doi.org/10.1186/s13007-021-00813-7>.
- [17] Shao, M., He, P., Zhang, Y., Zhou, S., Zhang, N., & Zhang, J. (2022). Identification method of cotton leaf diseases based on bilinear coordinate attention enhancement module. *Agronomy*, 13(1), 88. <https://doi.org/10.3390/agronomy13010088>.
- [18] Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine learning applications for precision agriculture: a comprehensive review. *IEEE Access*, 9, 4843-4873. <https://doi.org/10.1109/access.2020.3048415>.
- [19] Sharma, P., Dadheech, P., Aneja, N., & Aneja, S. (2023). Predicting agriculture yields based on machine learning using regression and deep learning. *IEEE Access*, 11, 111255-111264. <https://doi.org/10.1109/access.2023.3321861>.
- [20] Zhang, J., Liu, J., Chen, Y., Feng, X., & Sun, Z. (2021). Knowledge mapping of machine learning approaches applied in agricultural management—a scientometric review with citespace. *Sustainability*, 13(14), 7662. <https://doi.org/10.3390/su13147662>
- [21] Barbedo, J. G. A. (2018). "Impact of Dataset Size and Color Space on Deep Learning-Based Plant Disease Detection." *Biosystems Engineering*, 172, 31-43. [DOI: 10.1016/j.biosystemseng.2018.05.012].
- [22] Pujari, J. D., Yakkundimath, R., & Byadgi, A. S. (2016). "Image Processing-Based Detection of Fungal Diseases in Plants." *Procedia Computer Science*, 85, 329-336. [DOI: 10.1016/j.procs.2016.05.251]
- [23] Sharif, M., Khan, M. A., Rashid, M., Iqbal, Z., & Azam, F. (2018). "Detection and Classification of Citrus Diseases Using Texture Features and Random Forest Classifier." *Computers and Electronics in Agriculture*, 153, 12-22. [DOI: 10.1016/j.compag.2018.07.03]