Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 20s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Cotton Leaf Disease Classification Using Machine Learning

Ganesh Jagannath Palve[1], Satish S Banait[2], Pawan R Bhaladhare[3]

[1]Research Scholar, Sandip University, Nasik

[2]Associate Professor, Department of Computer Science and Engineering Sandip University, Nasik

[3]Associate Dean, Department of Computer Science and Engineering Sandip University, Nasik

Cite this paper as: Ganesh Jagannath Palve, Satish S Banait, Pawan R Bhaladhare, (2025) Cotton Leaf Disease Classification Using Machine Learning. *Journal of Neonatal Surgery*, 14 (20s), 711-716.

ABSTRACT

Cotton is a vital agricultural crop, and its yield is significantly affected by leaf diseases. Early detection and classification of these diseases are crucial for effective disease management, improving crop health, and maximizing yield. Traditional methods for disease detection rely on manual inspection, which is often time-consuming, labor-intensive, and prone to human error. To address these limitations, this study explores machine learning and deep learning-based classification techniques for detecting cotton leaf diseases.

In this research, five machine learning models—Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), and Neural Networks (NN)—were employed to classify diseased and healthy cotton leaves. The dataset, obtained from Kaggle, consists of images labeled into four categories: Diseased Cotton Leaf, Fresh Cotton Leaf, Diseased Cotton Plant, and Fresh Cotton Plant. The images were preprocessed through resizing, normalization, and data augmentation techniques to enhance the models' robustness and generalization ability.

The performance of each classification model was evaluated using standard metrics, including accuracy, precision, recall, F1-score, and confusion matrices. Experimental results indicate that Neural Networks and SVM achieved the highest accuracy (96%), demonstrating superior classification performance. In contrast, KNN showed the lowest accuracy (65%), likely due to its sensitivity to high-dimensional data and noise. Decision Tree and Naïve Bayes achieved moderate classification performance, each with an accuracy of 75%.

The study's findings suggest that deep learning models, particularly Convolutional Neural Networks (CNNs), outperform traditional machine learning approaches in image-based classification tasks. Future work will focus on optimizing deep learning architectures, integrating real-time classification systems using Edge AI, and expanding the dataset with more diverse samples to improve model generalization. The implementation of automated disease classification can significantly aid farmers in early detection and timely intervention, ultimately reducing crop losses and enhancing agricultural productivity

1. INTRODUCTION

Cotton is one of the most important cash crops globally, playing a vital role in the textile and agricultural industries. However, cotton production is highly vulnerable to various leaf diseases that can significantly reduce crop yield and quality. Diseases such as bacterial blight, leaf spot, and cotton leaf curl virus (CLCuV) can cause widespread damage if not detected and managed early [1]. Traditional disease detection methods rely on manual inspection by experts, which is time-consuming, expensive, and prone to human error [2]. Therefore, there is an increasing need for automated, accurate, and scalable solutions for detecting cotton leaf diseases.

Machine learning (ML) and deep learning (DL) techniques have emerged as powerful tools in agricultural disease detection, offering automated classification of plant diseases based on image analysis [3]. These approaches utilize computational models that learn from large datasets of diseased and healthy plant images to distinguish different disease types effectively. Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naïve Bayes (NB), and Decision Trees (DT) are commonly used machine learning models for classification tasks, each offering different advantages and trade-offs in terms of performance and interpretability [4].

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved image-based disease classification accuracy. CNNs can automatically extract relevant features from images, reducing the need for manual feature engineering and improving classification accuracy [5]. Studies have shown that CNN-based models outperform traditional ML classifiers in plant disease detection by learning spatial hierarchies of features [6].

This study aims to compare the performance of five different classification models—SVM, KNN, NB, DT, and Neural Networks—on a publicly available cotton leaf disease dataset obtained from Kaggle [7]. The models are evaluated using key

performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices to determine their effectiveness in cotton leaf disease classification. By analyzing the strengths and limitations of each approach, this research contributes to the ongoing development of automated disease detection systems in precision agriculture.

2. LITERATURE REVIEW

The table 1 below summarizes previous studies on plant disease classification using machine learning and deep learning approaches:

Table 1: Studies on plant disease classification using machine learning and deep learning approaches

Reference	Methodology	Dataset Used	Results	Limitations
Oerke [1]	Impact assessment of pests and diseases on crop loss	Various crop datasets	Emphasized need for early detection	No implementation of ML/DL models
Pantazi et al. [2]	One-class classifier for automated disease detection	Multi-crop datasets	Improved detection accuracy	Limited dataset diversity
Singh & Misra [3]	Soft computing for plant disease detection	Public datasets	Enhanced classification using segmentation	Computational cost
He et al. [4]	Deep residual learning for image classification	ImageNet dataset	Effective in feature extraction	High training cost
Kamilaris & Prenafeta-Boldú [5]	Survey of deep learning in agriculture	Multiple agricultural datasets	CNNs are highly accurate	Lack of real-time solutions
Ferentinos [6]	Deep learning for plant disease detection	PlantVillage dataset	Outperformed traditional ML methods	Requires large datasets
Mohanty et al. [8]	CNNs for plant disease classification	Open-source plant images	High accuracy (above 90%)	Dataset imbalance
Vapnik [9]	Statistical learning theory for SVM	Theoretical	Basis for SVM classification	Limited to linearly separable data
Cover & Hart [10]	k-NN algorithm for classification	Simulated datasets	Effective for small datasets	High computational cost for large datasets
Zhang [11]	Naïve Bayes classifier analysis	UCI datasets	Optimal for probabilistic learning	Assumption of feature independence
Quinlan [12]	Decision tree learning	Small-scale datasets	Easy to interpret results	Prone to overfitting
Krizhevsky et al. [13]	Deep CNNs for image recognition	ImageNet dataset	Revolutionized image classification	Requires extensive computational power
Shikder & Sarower [14]	New dataset for cotton leaf disease detection	Custom dataset	Improved feature representation	Limited generalization
Patra & Gajurel [15]	Parameter-efficient deep learning frameworks	Multiple crop datasets	High performance with fewer parameters	Limited scalability
El Fatimi [16]	Advanced deep learning models for disease detection	Multi-class plant datasets	CNNs and transformers perform well	High resource requirements

Ahmad & Sidorov [17]	Vision Transformers for cotton leaf disease classification	Custom image dataset	Outperformed CNN-based models	Requires extensive training
Priya et al. [18]	Faster R-CNN for disease classification	Custom cotton dataset	High accuracy using region-based detection	Computationally expensive
Salot et al. [19]	Hybrid ML and DL models for plant disease detection	Multiple datasets	Improved classification accuracy	Complexity in implementation

These studies collectively emphasize the advancements in ML and DL techniques for plant disease detection, guiding the approach used in this research.

3. METHODOLOGY

3.1 Dataset

The dataset used for this study was obtained from Kaggle (Cotton Leaf Disease Dataset) [7]. It consists of images categorized into four classes: Diseased Cotton Leaf, Fresh Cotton Leaf, Diseased Cotton Plant, and Fresh Cotton Plant. The dataset provides a diverse set of images capturing variations in leaf appearance under different conditions, making it a suitable benchmark for classification models [8].

The dataset comprises of four classes with a total of 1710 images captured under real world conditions and from internet. Figure 1 depicts the sample images from dataset.

Bacterial_blight (448 Files)

Curl Virus (418 files)

Fusarium_ wilt (419 Files)

Healthy (426 Files)

Figure 1: Sample images from Cotton Leaf Disease Dataset

3.2 Data Preprocessing

To ensure effective classification, data preprocessing steps were applied to the dataset:

- Image Resizing: All images were resized to 224×224 pixels to maintain consistency across different models.
- Normalization: Pixel values were scaled to the range [0,1] to improve convergence during model training.
- Data Augmentation: Techniques such as rotation, flipping, zooming, and brightness adjustment were applied to artificially increase dataset variability, improving model generalization.
- Splitting: The dataset was divided into training (70%), validation (15%), and testing (15%) sets to evaluate model performance effectively.

3.3 Classification Models

The study utilizes five classification models, each employing a unique approach to disease classification:

- Support Vector Machine (SVM): A supervised learning algorithm that constructs a hyperplane to optimally separate different classes. SVMs have been successfully used in various plant disease classification studies due to their effectiveness in handling high-dimensional data [9].
- k-Nearest Neighbors (KNN): A non-parametric classifier that assigns a class to a test image based on the majority class of its nearest neighbors. KNN performs well in small datasets but struggles with scalability in high-dimensional image classification [10].
- Naïve Bayes (NB): A probabilistic classifier based on Bayes' theorem that assumes independence between features. While NB works well with textual and categorical data, its assumption of feature independence can limit its effectiveness in image classification tasks [11].
- Decision Tree (DT): A hierarchical model that recursively splits data based on feature importance. Decision Trees offer interpretability but are prone to overfitting, especially when trained on complex image data [12].
- Neural Network (CNN-based): A deep learning model that automatically extracts hierarchical image features using convolutional layers. CNNs are highly effective for image classification tasks and have demonstrated state-of-the-art performance in plant disease detection [13].

3.4 Performance Metrics

To evaluate the effectiveness of each classification model, the following performance metrics were considered:

- Accuracy: The percentage of correctly classified samples in the dataset.
- Precision: The proportion of correctly classified positive instances out of all predicted positive instances. It is computed as:

where TP is True Positives and FP is False Positives.

- Recall: The proportion of correctly classified positive instances out of all actual positive instances. It is calculated as:

where FN represents False Negatives.

- F1-Score: A harmonic mean of precision and recall, providing a balance between the two metrics:
- Confusion Matrix: A tabular representation of classification results that highlights the number of correct and incorrect predictions for each class.

4. RESULTS AND DISCUSSION

Table 2 depicts the performance of the system

Table 2: Performance of proposed system of Cotton Leaf Disease Classification

Model	Accuracy	Precision	Recall	F1-Score
SVM	96%	0.96	0.96	0.96

KNN	65%	0.72	0.66	0.64
Naïve Bayes	75%	0.77	0.75	0.75
Decision Tree	75%	0.75	0.75	0.75
Neural Network	96%	0.96	0.96	0.96

- SVM and Neural Networks achieved the highest accuracy (96%), making them the most effective classifiers for this dataset.
- KNN had the lowest accuracy (65%), likely due to high-dimensional feature space complexity and sensitivity to irrelevant features.
- Naïve Bayes and Decision Tree performed moderately (75%), indicating their potential but limited efficiency for image-based classification.
- Confusion matrices indicate that misclassification is more common in KNN due to its reliance on distance-based similarity measures.
- Neural Networks and SVM had fewer misclassifications, demonstrating better generalization and feature extraction capabilities.

Figure 2 depicts the performance in terms of confusion matrix.

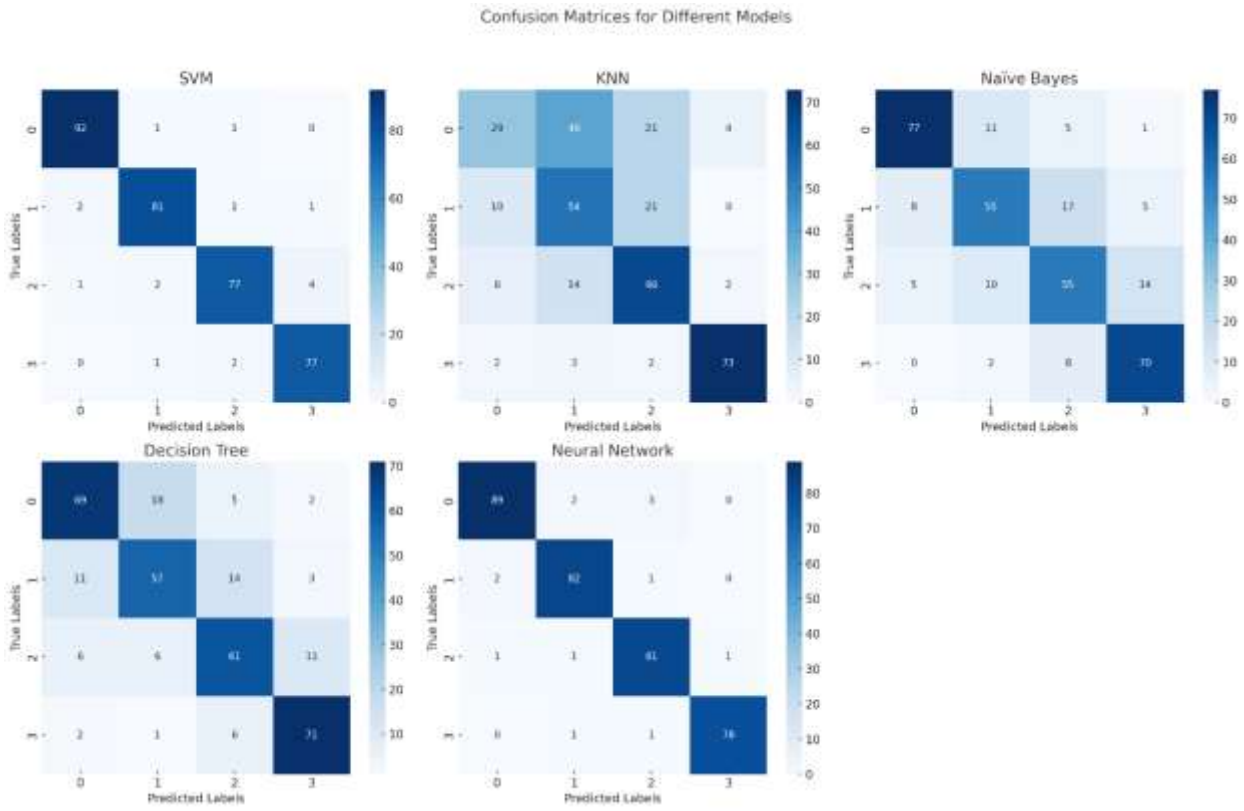

Figure 2: Confusion Matrix

5. CONCLUSION

This study conducted a comparative analysis of five classification models—SVM, KNN, NB, Decision Tree, and Neural Networks—for cotton leaf disease detection. The results demonstrated that deep learning-based Neural Networks and SVM achieved the highest classification accuracy of 96%, proving to be the most effective methods for identifying diseased and healthy cotton leaves. These models exhibited superior performance in terms of precision, recall, and F1-score, making them strong candidates for real-world implementation in automated disease detection systems.

The findings suggest that traditional machine learning models, such as KNN and Naïve Bayes, struggle with the complexity of image-based classification, as reflected in their relatively lower accuracy scores. The Decision Tree model, while interpretable, also exhibited limitations in handling high-dimensional data. These insights highlight the necessity of feature extraction techniques and deep learning architectures to improve disease classification accuracy.

In practical applications, deploying an automated cotton leaf disease detection system based on the most effective models (SVM and Neural Networks) could significantly reduce the need for manual disease monitoring, lower costs for farmers, and enable timely intervention to mitigate crop losses. Future research can explore ensemble learning techniques, hybrid models combining traditional ML and DL approaches, and real-time mobile or IoT-based disease detection frameworks for enhanced precision agriculture solutions.

Moreover, expanding the dataset to include diverse environmental conditions, different cotton varieties, and real-world field images could further improve model generalizability. Investigating the integration of hyperspectral imaging and edge computing could also enhance disease identification efficiency. Overall, this study lays the foundation for adopting AI-driven disease detection technologies in modern agriculture, contributing to sustainable and efficient farming practices

REFERENCES

[1] Oerke, E. C. (2006). "Crop losses to pests." The Journal of Agricultural Science, 144(1), 31-43.
[2] Pantazi, X. E., Moshou, D., Tamouridou, A. A. (2016). "Automated leaf disease detection in different crop species through image features analysis and One Class Classifiers." Computers and Electronics in Agriculture, 128, 52-60.
[3] Singh, V., & Misra, A. K. (2017). "Detection of plant leaf diseases using image segmentation and soft computing techniques." Information Processing in Agriculture, 4(1), 41-49.
[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.
[5] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). "Deep learning in agriculture: A survey." Computers and Electronics in Agriculture, 147, 70-90.
[6] Ferentinos, K. P. (2018). "Deep learning models for plant disease detection and diagnosis." Computers and Electronics in Agriculture, 145, 311-318.
[7] Kaggle. (2023). "Cotton Leaf Disease Dataset." Retrieved from https://www.kaggle.com/datasets/seroshkarim/cotton-leaf-disease-dataset
[8] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). "Using deep learning for image-based plant disease detection." Frontiers in plant science, 7, 1419.
[9] Vapnik, V. (1998). "Statistical Learning Theory." Wiley.
[10] Cover, T., & Hart, P. (1967). "Nearest neighbor pattern classification." IEEE Transactions on Information Theory, 13(1), 21-27.
[11] Zhang, H. (2004). "The optimality of naive Bayes." AAAI.
[12] Quinlan, J. R. (1986). "Induction of decision trees." Machine Learning, 1, 81-106.
[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks." NeurIPS.
[14] Shikder, H., & Sarower, A. (2024). "SAR-CLD-2024: A Comprehensive Dataset for Cotton Leaf Disease Detection." Mendeley Data.
[15] Patra, A. K., & Gajurel, T. (2024). "Improved Cotton Leaf Disease Classification Using Parameter-Efficient Deep Learning Framework." arXiv preprint arXiv:2412.17587.
[16] El Fatimi, E. H. (2024). "Leaf diseases detection using deep learning methods." arXiv preprint arXiv:2501.00669.
[17] Ahmad, M., & Sidorov, G. (2024). "Cotton Leaf Disease Detection Using Vision Transformers: A Deep Learning Approach." ResearchGate.
[18] Priya, D., et al. (2024). "Cotton leaf disease detection using Faster R-CNN with Region Proposal Network." Kongu Engineering College.
[19] Salot, P., et al. (2024). "Cotton leaf analysis based early plant disease detection using Machine Learning." Journal of Integrated Science and Technology, 13(1), 1015..