Content is available at: CRDEEP Journals
Journal homepage: http://www.crdeepjournal.org/category/journals/ijrem/

International Journal of Research in Engineering and Management (ISSN: 2456-1029)
A Peer Reviewed UGC Approved Quarterly Journal

SJIF: 4.45

Research Paper

Enhancing Potato Leaf Disease Classification through Optimized Machine Learning and Advanced Feature Selection Methods

Rakesh Kumar[1]; Dr. Rita Kumari Saini[1] and Dr. Nishant Saxena[21]
1-Department of Computer Science and Engineering, Sparsh Himalaya University Dehradun Uttarakhand, India
2-Co-Supervisor: Additional Director, Tula's Institute Dehradun Uttarakhand, India

ARTICLE DETAILS

Corresponding Author:
Dr Nishant Saxena

Key words:
Early blight · Feature selection · Late blight · Machine learning · Neural network Sustainable agriculture · Weather parameters

ABSTRACT

This study investigates the use of various machine learning (ML) models to forecast early and late blight in potatoes, diseases that significantly reduce crop yield and quality. Using a dataset of over 4,000 weather condition records, key factors such as temperature, humidity, wind speed, and atmospheric pressure were analyzed using techniques like K-means clustering, PCA, and copula analysis. ML models applied included logistic regression, gradient boosting, MLP, SVM, and KNN—with and without feature selection. Feature selection, notably using binary Greylag Goose Optimization (bGGO), significantly improved model accuracy. The MLP model with feature selection achieved the highest accuracy at 98.3%, highlighting the value of optimized features. This approach offers a reliable tool for early disease detection, supporting sustainable farming and food security through automated and effective disease control.

1. Introduction

Potatoes are a vital food crop and economic asset, especially in India, the world's second-largest producer with over 43 million tons produced in 2018. However, increasing global demand faces a major threat from leaf diseases like early and late blight, which cause significant yield losses. Early blight appears as black spots, while late blight leads to blistering and rotting of leaves. Traditionally, disease detection relied on manual inspection and basic weather models, which lack accuracy and scalability. This research leverages machine learning (ML), particularly CNNs, to identify and differentiate potato leaf diseases using meteorological data (temperature, humidity, rainfall, wind speed). Advanced ML models can uncover complex patterns and offer early warnings, enabling timely interventions. The goal is to build predictive systems that support sustainable farming by improving disease forecasting and reducing crop losses, thereby strengthening food security and supporting farmers' livelihoods. The core problem is the need for rapid and accurate forecasting of potato leaf diseases like early and late blight based on weather conditions. Existing systems often fall short in reliability and precision, revealing a clear research gap. This study addresses that gap by introducing an updated dataset and testing alternative machine learning models to improve forecasting accuracy, offering a more dependable tool for farmers and agricultural stakeholders. This research aims to tackle the challenges posed by potato leaf diseases, particularly early and late blight, through several key goals. It seeks to identify critical environmental factors influencing disease prevalence and develop an advanced climate-based prediction model using a refined dataset. The study also compares the performance of various models under different weather conditions to enhance forecasting accuracy. By addressing existing research gaps, the project proposes a more personalized and effective model. Ultimately, it aims to equip farmers with a reliable tool for disease management, contributing to reduced crop losses and improved food security.

2.Related Work

The integration of artificial intelligence (AI) and machine learning (ML) into agriculture is transforming traditional, experience-based practices that previously lacked algorithmic support. This section reviews the existing literature and

[1]Author can be contacted at: *Tula's Institute, Dehradun*
Received: *20-05-2025*; Sent for Review on: 27*-05- 2024*; Draft sent to Author for corrections: *10-06-2025*; Accepted on*: 12-06-2025*; Online Available from 25*-06-2025*
DOI: 10.13140/RG.2.2.20398.01607
IJREM: -8899/

examines key advancements in the field, with sub-sections dedicated to exploring how AI and ML technologies are being applied to agricultural disease detection and control.

Revolutionary Applications of Machine Learning and Artificial Intelligence

The use of AI and machine learning in agriculture marks a revolutionary shift from traditional, empirical methods of plant disease forecasting, which relied on historical outbreak and weather data but lacked adaptability. AI introduces advanced technologies that analyze complex datasets to predict when and where diseases may occur. By integrating real-time ground sensors, satellite imagery, and high-resolution drone data, machine learning algorithms can detect patterns and anomalies indicative of early disease development, offering more precise and timely disease management.

Climate Clues for Disease Prediction

This study focuses on developing AI models that incorporate key environmental factors—especially temperature, humidity, wind speed, and direction—to predict potato leaf diseases like early and late blight. These weather conditions strongly influence disease occurrence and fungal spore distribution. By optimizing AI models with thorough data collection, preprocessing, and feature selection from a dataset of 4,020 weather records, the research enhances prediction accuracy for real-field applications. This approach supports sustainable potato farming, boosts crop resilience, and contributes to food security amid changing climate conditions.

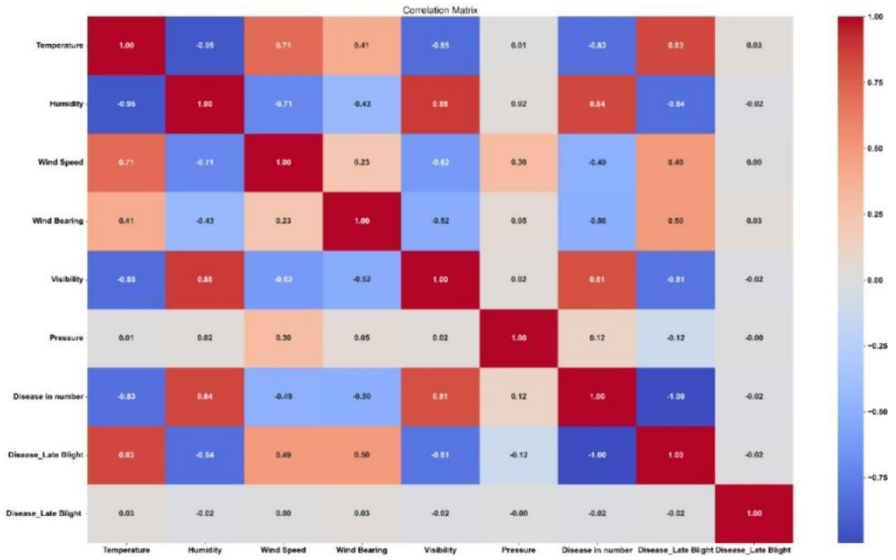


Fig. 1 Original dataset features' correlation matrix

The dataset includes columns for "Disease name" and "Due to a number of diseases," which categorize different potato leaf diseases, focusing specifically on early blight and late blight for machine learning model training. Early blight, caused by *Alternaria solani*, thrives in warm, humid conditions and appears as dark brown patches on leaves that impair photosynthesis. In contrast, late blight, caused by *Phytophthora infestans*, is more prevalent in cold, humid climates and can rapidly devastate entire fields under favorable conditions. Notably, temperature and humidity often exhibit a strong positive correlation, creating environments conducive to both early and late blight outbreaks.

Data preprocessing

Data preprocessing is crucial for cleaning, standardizing, and preparing datasets for analysis. In this study, data normalization and encoding techniques are applied, followed by Principal Component Analysis (PCA) to reduce dimensionality and simplify data processing. K-means clustering further groups data points by similarity, revealing natural patterns in weather variables. These clusters help link specific weather patterns to potato diseases, providing high-quality training data that improves machine learning model accuracy and reduces bias.

Copula Analysis

Copula analysis is a statistical technique used to uncover hidden relationships among variables by generating synthetic datasets that simulate future scenarios. These synthetic datasets, which include correlation matrices, help machine learning models recognize weather-disease patterns and improve prediction accuracy. By replicating the original data's statistical properties, copula synthesis supports robust model evaluation across diverse conditions, enhancing the understanding of weather-driven disease outbreaks.

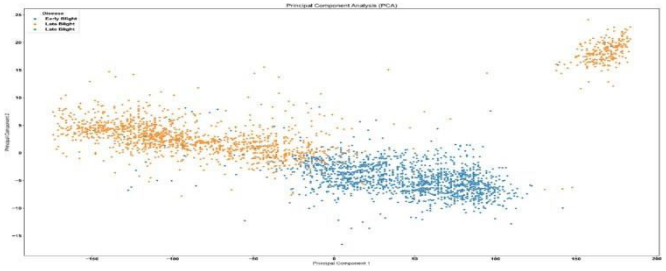
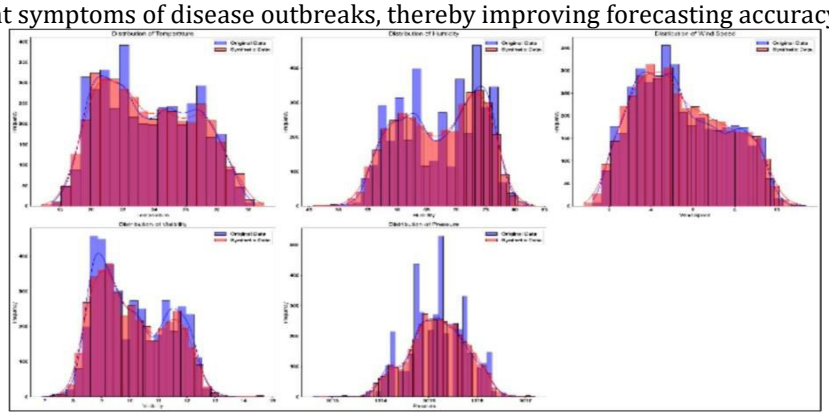


Fig. 2 Cluster analysis (*K*-means) of tested dataset

Fig. 3 Distribution of dataset features

Figure illustrates the variability of distinct weather attributes in relation to weather conditions, indicating that each attribute exhibits unique behaviors. Recognizing the extent and distribution of these features enhances models' capacity to detect subtle yet significant symptoms of disease outbreaks, thereby improving forecasting accuracy.



Feature Selection

Feature selection is vital for identifying key variables that predict diseases by converting features into binary form (0 or 1) to remove irrelevant or duplicate data. This streamlines models, improving accuracy and efficiency, with temperature and humidity often ranking highest. Advanced binary optimization algorithms like binary Greylag Goose Optimization (bGGO) and binary Waterwheel Plant Algorithm (bWWPA), along with others such as bGWO, bPSO, and bGA, are used to isolate the most relevant features. Table 1 summarizes feature selection evaluation metrics, including best/worst fitness, average error, and standard deviation.

Machine Learning Models

This research evaluates various machine learning models for potato disease prediction, each with unique strengths:

1. Logistic Regression – effective for binary classification by analyzing feature correlations.
2. Neural Network (MLP) – captures complex, non-linear data patterns.
3. Random Forest – ensemble of decision trees improving accuracy for diverse data.
4. Support Vector Machine (SVM) – uses hyperplanes for clear classification.
5. K-Nearest Neighbors (KNN) – classifies based on proximity to known data points.
6. Naive Bayes – probabilistic model predicting categorical outcomes.
7. Decision Tree – hierarchical decision rules based on weather variables.
8. Gradient Boosting – sequentially reduces prediction errors for higher accuracy.
9. SVM (RBF Kernel) – SVM variant using kernels for non-linear classification.

Models are assessed using accuracy, sensitivity, and specificity to find the best fit for forecasting. Table 1 summarizes the evaluation metrics ensuring reliable performance and clear differentiation of disease presence.

3.Experimental Results

The experimental outcome presents a comprehensive evaluation of machine learning models' ability to predict potato leaf diseases using meteorological data.

Table 1: Classification model evaluation criteria

Evaluation Criteria	**Value**
Accuracy	No. of correct predictions / Total predictions
Sensitivity (Recall)	True positives / True positives + false negatives
Specificity	True negatives / True negatives + false positives

Positive predictive value	True positives / (True positives + false positives)
Negative predictive value	True negatives / (True negatives + false negatives)
F1 score	2 × (precision × sensitivity) / (precision + sensitivity)

Machine Learning Model Results with Feature Selection

KNN demonstrates that feature selection effectively removes noise, focusing models on the most important predictors, improving sensitivity and specificity. Figure 10 shows that after feature selection, all models, especially MLP, perform better, with KNN and random forest also strong. This highlights the crucial role of selecting key features in boosting accuracy. Figure 11's pair plot compares performance before and after feature selection, showing models like MLP and random forest consistently improve across metrics, confirming that optimal feature selection enhances prediction accuracy.

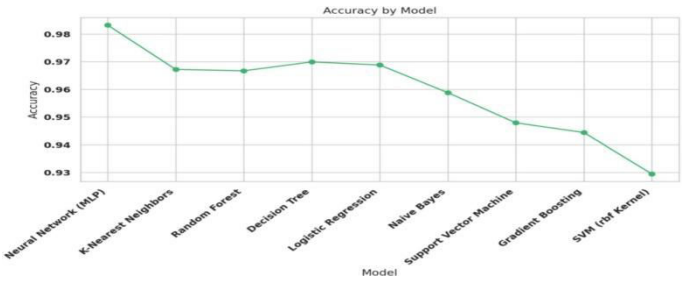


Fig 4: Accuracy by model for machine learning models with feature selection

4. Conclusion

This project explored using weather factors and various machine learning models to predict potato leaf diseases. Before feature selection, logistic regression (LR) and neural networks (MLP) achieved over 94% accuracy. Feature selection techniques, especially bGGO and bWWPA, improved all models by identifying key features and reducing errors. MLP performed best, reaching 98.3% accuracy and efficient data reduction. The study highlights the importance of optimizing ML models for effective disease prediction, helping farmers make accurate, timely decisions to minimize crop losses. Future work should expand datasets to include more crops and diseases, apply advanced feature selection, and develop user-friendly management tools to support sustainable agriculture. Overall, this research demonstrates the strong potential of AI to enhance agricultural disease control.

References

Aditya Shastry K, Sanjay HA (2021) A modified genetic algorithm and weighted principal compo- nent analysis based feature selection and extraction strategy in agriculture. Knowl-Based Syst 232:107460. https://doi.org/10.1016/j.knosys.2021.107460

Albulescu CT, Tiwari AK, Ji Q (2020) Copula-based local dependence among energy, agriculture and metal commodities markets. Energy 202:117762. https://doi.org/10.1016/j.energy.2020.117762

Alhussan AA, Abdelhamid AA, El-Kenawy El-S M, Ibrahim A, Eid MM, Khafaga DS, Em AA (2023) A binary waterwheel plant optimization algorithm for feature selection. IEEE Access 11:94227– 94251. https://doi.org/10.1109/ACCESS.2023.3312022

Ali MZ, Abdullah A, Zaki AM, Rizk FH, Eid MM, El-Kenway EM (2024) Advances and challenges in feature selection methods: a comprehensive review. J Artif Intell Metaheuristics 7(1):67–77. https:// doi.org/10.54216/JAIM.070105

Arshaghi A, Ashourian M, Ghabeli L (2023) Potato diseases detection and classification using deep learn- ing methods. Multimed Tools Appl 82(4):5725–5742. https://doi.org/10.1007/s11042-022-13390-1

Ayoub Shaikh T, Rasool T, Rasheed Lone F (2022) Towards leveraging the role of machine learn- ing and artificial intelligence in precision agriculture and smart farming. Comput Electron Agric 198:107119. https://doi.org/10.1016/j.compag.2022.107119

Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D (2021) Machine learning in agri- culture: a comprehensive updated review. Sensors 21(11):3758. https://doi.org/10.3390/s2111 3758

Bhat SA, Huang N-F (2021) Big data and AI revolution in precision agriculture: survey and chal- lenges. IEEE Access 9:110209–110222. https://doi.org/10.1109/ACCESS.2021.3102227

Cravero A, Pardo S, Sepúlveda S, Muñoz L (2022) Challenges to use machine learning in agricul- tural big data: a systematic literature review. Agronomy 12(3):748. https://doi.org/10.3390/agron omy12030748

Das S, Das J, Umamahesh NV (2022) Copula-based drought risk analysis on rainfed agriculture under stationary and non-stationary settings. Hydrol Sci J 67(11):1683–1701. https://doi.org/10.1080/ 02626667.2022.2079416

Dhal P, Azad C (2022) A comprehensive survey on feature selection in the various fields of machine learning. Appl Intell 52(4):4543–4581. https://doi.org/10.1007/s10489-021-02550-9

Dolničar P (2021) Importance of potato as a crop and practical approaches to potato breeding. In: Dobnik D, Gruden K, Ramšak Ž, Coll A (eds) Solanum tuberosum: methods and protocols. Springer, US, New York, NY, pp 3–20

El-kenawy El-SM, Khodadadi N, Mirjalili S, Abdelhamid AA, Eid MM, Ibrahim A (2024) Grey- lag Goose Optimization: nature-inspired optimization algorithm. Expert Syst Appl 238(Part E):122147. https://doi.org/10.1016/j.eswa.2023.122147

Fenu G, Malloci FM (2020) Artificial intelligence technique in crop disease forecasting: a case study on potato late blight prediction. In: Czarnowski I, Howlett RJ, Jain LC (eds) Intelligent decision technologies. Springer, Singapore, pp 79–89

Gao J, Westergaard JC, Sundmark EHR, Bagge M, Liljeroth E, Alexandersson E (2021) Automatic late blight lesion recognition and severity quantification based on field imagery of diverse potato genotypes by deep learning. Knowl-Based Syst 214:106723. https://doi.org/10.1016/j.knosys. 2020.106723

Garske B, Bau A, Ekardt F (2021) Digitalization and AI in European agriculture: a strategy for achieving climate and biodiversity targets? Sustainability 13(9):4652. https://doi.org/10.3390/ su13094652

Gold KM, Townsend PA, Chlus A, Herrmann I, Couture JJ, Larson ER, Gevens AJ (2020) Hyperspec- tral measurements enable pre-symptomatic detection and differentiation of contrasting physiological effects of late blight and early blight in potato. Remote Sensing 12(2):286. https://doi.org/10.3390/ rs12020286

Gold KM, Townsend PA, Herrmann I, Gevens AJ (2020) Investigating potato late blight physiological differences across potato cultivars with spectroscopy and machine learning. Plant Sci 295:110316. https://doi.org/10.1016/j.plantsci.2019.110316

Gupta N, Khosravy M, Patel N, Dey N, Gupta S, Darbari H, Crespo RG (2020) Economic data ana- lytic AI technique on IoT edge devices for health monitoring of agriculture machines. Appl Intell 50(11):3990–4016. https://doi.org/10.1007/s10489-020-01744-x

Hamrani A, Akbarzadeh A, Madramootoo CA (2020) Machine learning for predicting greenhouse gas emissions from agricultural soils. Sci Total Environ 741:140338. https://doi.org/10.1016/j.scitotenv. 2020.140338

Javidan SM, Banakar A, Vakilian KA, Ampatzidis Y (2023) Diagnosis of grape leaf diseases using auto- matic K-means clustering and machine learning. Smart Agric Technol 3:100081. https://doi.org/10. 1016/j.atech.2022.100081

Kang F, Li J, Wang C, Wang F (2023) A lightweight neural network-based method for identifying early- blight and late-blight leaves of potato. Appl Sci 13(3):1487. https://doi.org/10.3390/app13031487