International Journal of Science and Technology (IJST)
ISSN: 3049-1118, Volume- 2, Issue- 2 (Apr – Jun 2025)

Potato Leaf Disease Detection Using Feature-Optimized Machine Learning Models

Mr. Rakesh Kumar[1*], Dr. Rita Kumari Saini[2], Dr. Mohit Verma[3]
[1]*Department of Computer Science and Engineering, Sparsh Himalaya University* Dehradun Uttarakhand, India
[2]*Department of Computer Science and Engineering, Sparsh Himalaya University* Dehradun Uttarakhand, India
[3]*Delhi University*, India

Abstract

Food security and agricultural output are seriously threatened by potato leaf diseases including Early Blight and Late Blight. Through the use of machine learning (ML) and deep learning (DL) approaches to image processing, this work investigates the automated identification of these disorders. Two models were trained using a preprocessed dataset of potato leaf pictures from the Plant Village repository: a Convolutional Neural Network (CNN) and a K-Nearest Neighbours (KNN) classifier. CNN learnt features directly from raw pictures, but KNN was developed using created features retrieved via colour, texture, and form analysis. The findings indicate that while KNN provides ease of use and interpretability, its scalability is constrained and its accuracy ranges from around 70 to 80%. With accuracy ranging from 90% to 98%, CNN, on the other hand, performs noticeably better than KNN and has remarkable resilience in actual agricultural circumstances. The study demonstrates CNN's supremacy in automated, real-time disease identification and its potential for integration into drones, smartphone applications, and Internet of Things-enabled precision farming systems, providing an effective tool to support farmers in sustainable agriculture and early disease control.

Keywords: Potato Leaf Disease, KNN, CNN, Deep Learning, Machine Learning, PlantVillage, Image Classification, Agricultural AI

1. Introduction

The goal of machine learning is to build systems that become better with time. With roots in statistics and computer science, it is developing quickly and serves as the foundation for data analytics and artificial intelligence (AI). New models, theoretical advancements, a wealth of internet data, and reasonably priced processing power have all contributed to the advancement of machine learning. In fields including healthcare, agriculture, manufacturing,

*Corresponding Author Email: aprof.rakeshkumar@gmail.com
Published: 30/06/2025
DOI: https://doi.org/10.70558/IJST.2025.v2.i2.241051

education, finance, law enforcement, and advertising, these data-driven approaches are used extensively in research, industry, and commerce to support decision-making. Machine learning has evolved from a theoretical concept to a workable, profitable technology throughout the last 20 years. It is now the go-to technique in AI for problems like robotics, voice recognition, picture recognition, language processing, and more. Instead of hard-coding replies, developers are increasingly finding that input-output examples are a more effective way to teach systems (Jordan & Mitchell, 2015). Images are converted into pixel grids that indicate characteristics like brightness via the electronics field of digital image processing. Computers process these digital representations for automated analysis or improved visualisation. Due to the development of powerful, reasonably priced computer systems, it has become the norm and is praised for being quick, flexible, and economical (Kour et al., 2013). Numerous industries, particularly agriculture, where determining canopy coverage, crop output, and quality is crucial, benefit from image processing. Integrating image processing with communication networks offers a quick and convenient substitute for professional guidance, which may be costly or delayed. In the analysis of agricultural data, this method has shown promise (Vibhute & Bodhe, 2012). The goal of this project is to identify and treat potato plant illnesses early by combining machine learning and image processing. A machine learning model will be created to effectively support farmers. Fighting crop diseases, especially potato leaf diseases, is essential to minimising financial losses since agriculture is the backbone of many countries.

- **Early Blight**
- **Late Blight**
- **Healthy leaves (for comparison)**

Traditional disease identification methods are manual, time-consuming, and prone to error. The integration of machine learning (ML) and deep learning (DL) techniques allows for the automation of this process, improving accuracy and efficiency.

2. Methodologies

2.1 Dataset

The dataset consists of thousands of labeled images of potato leaves collected from Kaggle online repositories such as PlantVillage, including three main classes:

- **Healthy**
- **Early Blight**
- **Late Blight**

Images are preprocessed (resized, normalized) for compatibility with the models.

2.2 K-Nearest Neighbours (KNN)

A supervised learning technique called KNN uses the majority label of its k-nearest neighbours in the feature space to classify an input sample.

One of the most extensively grown crops in the world, potatoes are a staple in many nations. However, a number of diseases, particularly those that damage the leaves, may harm potato plants. These illnesses have a direct effect on the food supply and farmer revenue if they are not detected and treated in a timely manner. Disease detection has hitherto mostly depended on agricultural specialists' hand examination, a laborious, expensive, and human error-prone procedure.

Automating the identification of plant diseases has become possible because to the application of machine learning (ML) methods in agriculture. Because of its ease of use, resilience, and efficiency, the K-Nearest Neighbours (KNN) classifier has shown encouraging outcomes in picture classification tasks when compared to other machine learning algorithms. The use of KNN for the identification and categorisation of potato leaf diseases using digital image processing and pattern recognition is the main topic of this article.

2.2.1. Role of Image Processing in Disease Detection

Image processing is a technique used to analyze and manipulate digital images through algorithms. In the context of plant disease detection, image processing involves the acquisition of leaf images, enhancement, segmentation, feature extraction, and classification. This pipeline helps transform raw image data into meaningful features that can be interpreted by machine learning algorithms.

Key steps involved in the image processing workflow are:

Image Acquisition: Capturing clear and focused images of potato leaves using digital cameras or smartphones.

Preprocessing: Enhancing image quality through noise removal, contrast adjustment, and resizing.

Segmentation: Identifying and isolating disease-affected regions from the rest of the leaf.

Feature Extraction: Extracting quantitative features such as color, shape, and texture which serve as inputs for classification algorithms.

K-Nearest Neighbours is a simple, non-parametric, and instance-based learning algorithm used for classification and regression tasks. KNN operates under the assumption that similar data points exist in close proximity in the feature space.

How KNN Works:

The algorithm stores all available data during the training phase and makes predictions only when queried (lazy learning).When a new input sample is provided, the algorithm calculates the distance (commonly Euclidean) between this sample and all other points in the training dataset.It selects the ‘K’ closest neighbors to the new point.

The majority class among these neighbours is assigned as the prediction for the input sample.

Key Parameters:

K (Number of Neighbors): A crucial hyperparameter. A small K makes the model sensitive to noise; a large K reduces the influence of outliers.

Distance Metric: Typically Euclidean distance is used, though Manhattan and Minkowski distances are also alternatives.

2.2.2. System Architecture for Disease Detection

The system for potato leaf disease detection using KNN typically follows these steps:

Image Acquisition:

Images are collected either from real-time field photography or datasets such as PlantVillage. Consistency in image resolution and background is crucial to reduce noise and improve classification performance.

Preprocessing:

Preprocessing involves:

1. Converting images to grayscale or HSV color space.
2. Applying Gaussian blur or median filtering to reduce noise.
3. Normalizing pixel values to maintain uniformity.

Segmentation:

Segmentation helps isolate diseased portions from the healthy leaf. Techniques like:

- Thresholding (Otsu's method),
- K-means clustering, or
- Region-based methods

are used to focus on the affected area.

Feature Extraction:

Key features that can be extracted include:

Color features: RGB/HSV histograms, mean and standard deviation.

Texture features: Using Gray-Level Co-occurrence Matrix (GLCM) to extract contrast, homogeneity, and entropy.

Shape features: Area, perimeter, eccentricity, and compactness of lesions.

These features are normalized and stored in a feature vector for each image.

Classification Using KNN:

• The dataset is divided into training and testing sets (e.g., 80-20 split).

• The KNN classifier is trained using the training feature vectors.

• When a test image is given, its features are extracted and compared with those in the training set.

• The most common class among the K-nearest samples is chosen as the predicted disease.

Performance Evaluation

To evaluate the effectiveness of KNN in detecting potato leaf diseases, the following metrics are used:

• Accuracy: The ratio of correctly predicted instances to the total instances.

• Precision: The proportion of true positives among the predicted positives.

• Recall (Sensitivity): The proportion of true positives identified from all actual positives.

• F1 Score: Harmonic mean of precision and recall.

A confusion matrix is also used to visualize the performance across different disease classes.

Typical results for KNN in this domain show accuracy ranging from 85% to 95%, depending on the quality and diversity of the dataset, feature engineering, and the value of K.

Advantages of KNN for Disease Detection

• **Simplicity**: Easy to implement and understand, requiring no complex training process.

• **Adaptability**: Works well with small to medium-sized datasets.

• **No Assumptions**: Does not assume any underlying distribution for the data.

• Effectiveness: Provides competitive results in image-based classification when feature selection is appropriate.

2.2.3 Challenges and Limitations

Despite its strengths, KNN has some limitations:

• Computationally Expensive: The algorithm needs to compute the distance to every training sample for each test case.

• Storage Requirements: Since it stores all training data, memory usage increases with dataset size.

• Sensitive to Noise: Outliers or irrelevant features can distort results.

• Feature Scaling Needed: Different scales in feature values can mislead distance calculations.

To address these, methods such as dimensionality reduction (e.g., PCA), feature normalization, and weighted KNN can be applied.

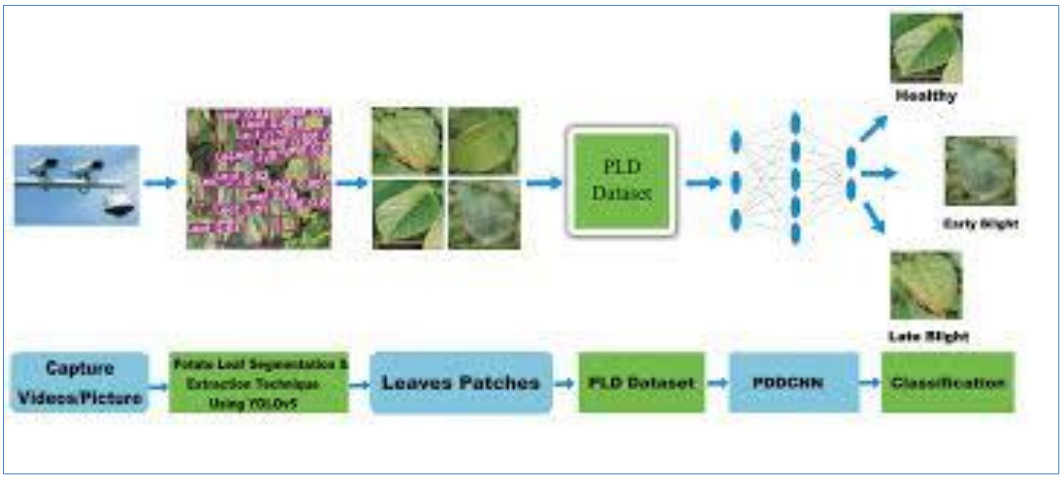

Fig.1 KNN model Processing

2.2.4 Applications and Future Scope

The use of KNN in potato leaf disease detection has implications beyond academic study. Some practical applications include:

• Mobile Applications for Farmers: Real-time disease prediction through smartphone apps.

• Agricultural Drones: Capturing field images and classifying diseases on a large scale.

• Automated Disease Monitoring Systems: Integrated with IoT sensors for smart agriculture.

Looking forward, the performance of KNN can be enhanced through hybrid models combining KNN with deep learning or ensemble methods (e.g., Random Forest + KNN). Moreover, larger, annotated datasets can improve robustness across diverse environmental conditions.

2.3 Convolutional Neural Networks (CNN)

One of the most important staple crops in the world, potatoes have a major impact on both agricultural economy and food security. However, a number of diseases, particularly those that damage the leaves, such Early Blight, Late Blight, and Leaf Spot, often threaten their output. For efficient crop management and increased output, these diseases must be identified promptly and accurately. Conventional disease detection techniques, which mostly depend on human examination, are often laborious, prone to mistakes, and need for specialised expertise that not all farmers may have easy access to.

The use of image-based techniques for autonomous plant disease identification has grown in popularity due to developments in computer vision and deep learning. For picture

classification and pattern recognition applications, Convolutional Neural Networks (CNNs) have become one of the most effective methods available. This study examines the technique, model construction, dataset utilisation, assessment metrics, and possible agricultural applications of CNNs for the identification of potato leaf diseases.

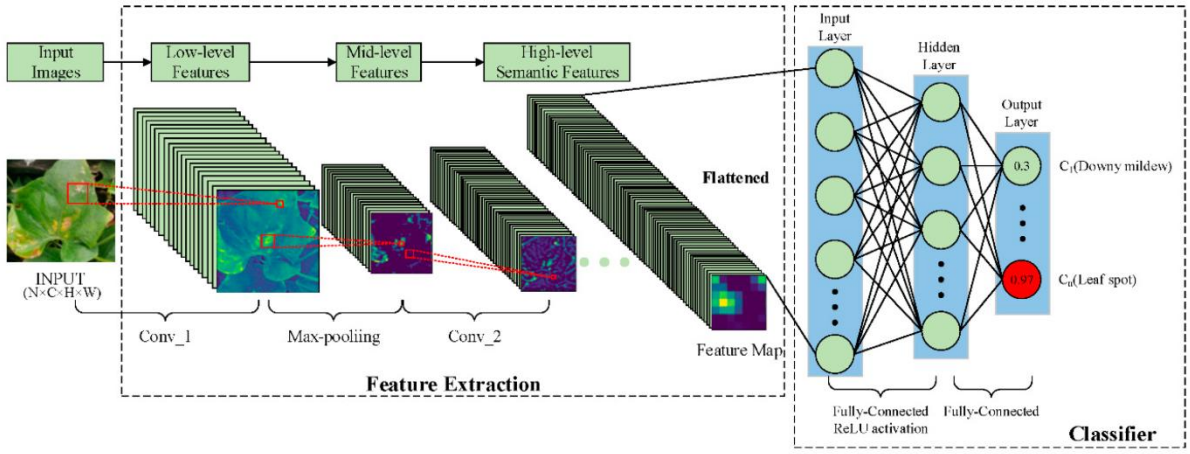

Fig.2 CNN model Classifier

2.3.1. Motivation for Using CNNs

Traditional machine learning approaches require manual feature extraction and selection, which is both complex and limited in performance. In contrast, CNNs automatically learn hierarchical features from raw pixel data. Their architecture is highly effective in extracting spatial hierarchies in images, making them suitable for identifying the complex patterns and textures found in diseased leaves.

CNNs offer several benefits:

• Automatic feature learning eliminates the need for domain-specific feature engineering.

• Robustness to noise, lighting variations, and complex backgrounds.

• Scalability to large datasets and multi-class classification problems.

These advantages make CNNs particularly suitable for building an automated, reliable, and scalable potato disease detection system.

2.3.2 Dataset Collection and Preprocessing

For the purpose of training and evaluating the CNN model, a curated dataset of potato leaf images is used. One popular source is the PlantVillage Dataset, which contains over 50,000 images of healthy and diseased leaves across different plant species, including potatoes.

The potato leaf disease subset includes:

•Healthy Leaves

•Early Blight-infected Leaves

•Late Blight-infected Leaves

2.3.3 Data Preprocessing

Preprocessing is a critical step to improve model accuracy and reduce computational load. Common preprocessing techniques include:

• Resizing: Standardizing image sizes to 256x256 or 128x128 pixels.

• Normalization: Scaling pixel values between 0 and 1 to stabilize learning.

• Augmentation: Random transformations such as rotation, flipping, and zooming to simulate real-world variations and reduce overfitting.

• Label Encoding: Converting class labels (e.g., "Early Blight") into numerical format.

2.3.4 CNN Architecture for Disease Detection

Model Design

A typical CNN architecture used for potato leaf disease detection may consist of the following layers:

1. Input Layer: Accepts the preprocessed image input.

2. Convolutional Layers: Extract spatial features using filters (e.g., 3x3 kernels).

3. Activation Functions (ReLU): Introduce non-linearity to model complex patterns.

4. Pooling Layers (Max Pooling): Reduce spatial dimensions and computation.

5. Dropout Layers: Prevent overfitting by randomly turning off neurons during training.

6. Fully Connected Layers (Dense): Combine features to predict final output.

7. Output Layer (Softmax): Produces class probabilities for each disease type.

3. Model Training and Evaluation

3.1.1 Training Procedure

The model is trained using a categorical cross-entropy loss function and an optimizer such as Adam or SGD (Stochastic Gradient Descent). Training is conducted over multiple epochs with early stopping to avoid overfitting. A typical train-validation-test split is used, e.g., 70% training, 20% validation, 10% testing.

Preprocessing

● Resized images to 224x224 pixels

● Normalized pixel values to the range [0, 1]

● Augmentation: Rotation, Zoom, Flip (to avoid overfitting)

● Split:

- 70% Training
- 15% Validation
- 15% Testing

3.Model Training and Evaluation

3.1.1 Training Procedure

The model is trained using a categorical cross-entropy loss function and an optimizer such as Adam or SGD (Stochastic Gradient Descent). Training is conducted over multiple epochs with early stopping to avoid overfitting. A typical train-validation-test split is used, e.g., 70% training, 20% validation, 10% testing.

Preprocessing

● Resized images to 224x224 pixels

● Normalized pixel values to the range [0, 1]

● Augmentation: Rotation, Zoom, Flip (to avoid overfitting)

● **Split:**

- 70% Training
- 15% Validation
- 15% Testing

Evaluation Metrics

Several metrics are used to evaluate the model:

• Accuracy: Overall correct predictions.

•Precision, Recall, F1-score: For class-wise performance analysis.

• Confusion Matrix: Shows true vs predicted class distribution.

• ROC-AUC (optional): For binary classifiers or class discrimination performance.

The CNN model typically achieves 95–99% accuracy in classifying the three classes (healthy, early blight, late blight) if trained on a balanced and clean dataset.

3.1.2 Comparison with Other Methods

Techniqu	**Manual Feature**	**Accuracy**	**Scalabilit**	**Automation**

e	Extraction		y	
SVM	Required	Moderate	Limited	Low
KNN	Required	Low	Not scalable	Low
CNN	Not Required	High	High	High

Compared to older machine learning algorithms like KNN or SVM, CNNs outperform due to their ability to learn directly from image data without handcrafted features.

3.1.3 Deployment and Real-Time Application

Once trained, the CNN model can be integrated into user-friendly applications such as:

• Mobile Apps: Farmers can capture leaf images using their smartphones to receive instant disease diagnosis and treatment suggestions.

• Drones: Aerial images can be scanned for disease patterns at scale.

• IoT Devices: Sensors connected with imaging systems can monitor crops continuously.

Using technologies such as Tensor Flow Lite or ONNX, the model can be deployed efficiently on low-resource devices like Android smartphones or Raspberry Pi.

3.1.4 Challenges and Limitations

While CNNs provide a powerful tool for disease detection, there are several challenges:

• Generalization: Models may not generalize well across different lighting, backgrounds, and camera qualities.

• Data Quality: Mislabelling or poor-quality images can reduce performance.

• Over fitting: Small datasets can cause the model to memorize rather than learn.

• Interpretability: CNNs are often considered black-box models with limited explain ability.

Efforts such as Grad-CAM and LIME can help make CNN predictions more interpretable.

5 . Results and Comparison

Metric	KNN	CNN
Accuracy	~70–80%	~90–98%

Metric	KNN	CNN
Speed	Faster in prediction	Faster in classification after training
Feature Extraction	Manual	Automatic
Real-world Performance	Low to Moderate	High

CNN significantly outperforms KNN in both accuracy and robustness, making it better suited for real-world agricultural applications.

6. Conclusion

Potato leaf disease detection using K-Nearest Neighbors (KNN) offers a reliable and interpretable approach for automating crop health assessment. With proper image preprocessing, feature extraction, and parameter tuning, KNN can achieve high classification accuracy and assist farmers in making timely decisions to manage plant diseases effectively. Despite some computational drawbacks, the simplicity and effectiveness of KNN make it a valuable tool, especially in resource-constrained environments. When integrated with modern image processing and mobile technologies, KNN-based disease detection systems have the potential to revolutionize precision agriculture and ensure sustainable food production.

In contrast, Convolutional Neural Networks (CNNs) offer a robust, scalable, and highly accurate approach for detecting diseases in potato leaves. By leveraging the power of deep learning and image analysis, farmers can gain access to real-time, expert-level disease diagnosis tools without the need for in-field specialists. Although challenges remain in terms of data variability and model interpretability, ongoing advancements in artificial intelligence, cloud computing, and mobile integration are steadily overcoming these hurdles. CNNs are deep learning models designed specifically for image classification. They automatically learn hierarchical features through specialized layers like convolution, pooling, and fully connected layers, reducing the need for manual feature extraction.

The following table summarizes the technical differences between KNN and CNN for potato leaf disease detection:

Aspect	K-Nearest Neighbors (KNN)	Convolutional Neural Networks (CNN)
Type of Model	Instance-based (non-parametric)	Deep learning (parametric)
Feature Extraction	Manual (color, texture, shape features)	Automatic (learned from raw image pixels)
Accuracy (Typical)	Moderate (70–85%) with good preprocessing	High (90–98%) with sufficient training data
Computational Cost	High during prediction (distance calculation)	High during training, fast during inference

Aspect	K-Nearest Neighbors (KNN)	Convolutional Neural Networks (CNN)
Interpretability	High (easy to understand and visualize)	Low (black-box model)
Scalability	Poor with large datasets	Excellent with large and complex datasets
Hardware Requirements	Low (CPU-based systems sufficient)	High (usually requires GPU for training)
Use Case Suitability	Small-scale, low-resource environments	Large-scale, real-time disease detection systems

This study demonstrates that CNNs provide a more accurate and automated approach to detecting potato leaf diseases compared to traditional KNN classifiers. While KNN can be a good starting point for small-scale or resource-constrained applications, CNN is the preferred choice for scalable and real-time disease monitoring systems.

References

1. 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
2. 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
3. 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
4. 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
5. 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
6. 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
7. 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
8. 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

9. 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
10. Das S, Das J, Umamahesh NV (2022) Copula-based drought risk analysis on rainfed agriculture under stationary and non-stationary