An Empirical Study to Predict Potato Diseases using Machine Learning Algorithms

Akanksha Kaushik[1], Manvi Breja[2]

akankshakaushik@ncuindia.edu[1], manvibreja@ncuindia.edu[2]

Department of CSE, The NorthCap University, Gurugram, India[1,2]

Abstract. Potato is one of the common and most prevalent crop in multiple countries, but it is prone to numerous diseases that can result in severe productivity losses. The diseases on potatoes need to be early recognized and categorized so that appropriate treatment and control is done. Using a variety of machine learning techniques, we looked at how effectively five generic potato diseases such as `early-blight`, `late-blight`, `common-scab`, `black-scurf`, and `soft-rot` are classified. Our research indicates that potato infections can be effectively classified by applying machine learning algorithms, which can help with disease monitoring and control in real time. In this research, we offer a system that classifies two distinct diseases in potato plants according to the health of their leaves using machine learning to produce an accurate classification system. The proposed system utilizes supervised and unsupervised classification models. This experiment exhibited a 95% accuracy rate after implementing Convolutional Neural Network which highlights the effectiveness and accuracy in detecting potato diseases early.

Keywords: Potato, Plant Disease, Support Vector Machine, K-Nearest Neighbor, Random Forest, Convolutional Neural Network, Prediction.

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362546

1 Introduction

Potato is amongst world's highly cultivated and consumed crops, providing an important source of food and money for millions of people. Unfortunately, potato plants are prone to a variety of diseases which causes degradation of crop quality and ultimately result in loss. Early diagnosis and categorization of these diseases are critical for successful management and control, which can assist decrease the use of pesticides and other expensive control methods. Conventional disease diagnosis and classification procedures are time-consuming and sometimes need specialist expertise and equipment. In recent years, to automate the classification of potato disease, machine learning (ML) algorithms are utilized. Machine learning algorithms analyze enormous amount of data which help in uncovering patterns and characteristics that older approaches find difficult to discern [1], In our study, we determine that PVY-infected potato plants display distinctive spectral reflectance patterns when compared to surrounding non-infected potato plants. By utilizing infrared wavelengths, SVM machine learning classifier achieved 89.8% accuracy by distinguishing between spectral reflectance curves. The curves of both PVY-infected as well as non-infected potato plants are input to classifiers by employing image processing and machine learning techniques.

This research demonstrates that integrating machine learning classifiers with remote sensing outperforms current industry benchmarks helps in enhancing PVY detection efforts which ultimately leads to improve the quality of potato seeds and maximize the yields. Fungal infections on potato leaves, specifically early bright and late blight during different stages have a significant impact on the yield of potato. These infections appear as brown spots and black marks on the leaves. To early detect these diseases, images of potato leaves are taken which are examined for any sign/sympoms of diseases, and then categorized and labeled as healthy or infected. For this, we applied image processing techniques integrating with machine learning algorithms. The performance of SVM classifiers, Random Forest and CNN are evaluated on the same test dataset of potato leaves.

2 Literature Survey

There are different methods or models available in machine learning which are used to find the accuracy. Several researchers have used several models, and they have achieved accuracy as described in the below mentioned table 1:

Table 1. Literature Survey

Authors	Methods/Models	Key Findings/Results	Accuracy (%)
Tiwari et al. [2]	VGG16, VGG19 CNN architectures	Distinguished 4 types of potato plant diseases based on leaf conditions. VGG19 achieved highest accuracy of 91%.	91
Sharma et al. [3]	SVM, KNN, Naive Bayes, Decision Tree, Gaussian filter, K-means clustering	Categorized early bright and late blight infections on potato leaves. SVM had highest accuracy of 92.9% after preprocessing and clustering	92.9
Ji et al. [4]	SVM, Hyperspectral Imaging, LDA, K-means	Identified damaged potatoes using SVM and hyperspectral imaging.	90

	clustering	Achieved 90% accuracy in non-destructive classification	
Coulibali et al. [5]	DBSCAN, KNN, Random Forest, SVM, pre-processed ionomers, nutrient concentrations (CLR)	Evaluated nutritional effects on high- and low-yield cultivars. Machine learning models achieved average accuracy of 70%.	70
Aldhyani et al. [6]	CNN, SVM, KNN, Random Forest, Decision Tree, Plant Village Dataset	Applied CNN algorithm to classify the diseases of plants which achieves best accuracy of 91.28% as compared to other models.	91.28
Kumar and Paswan [7]	ANN, KNN, Image processing, Color, Texture, Area properties	Used ANN and KNN to classify potato tuber diseases. ANN had higher accuracy of 84.76% compared to KNN (63.33%).	84.76 (ANN)

3 Implementation Methods

3.1 Dataset Description

This research utilizes a dataset related to potato leaf light was used that was provided in the 2016 Kaggle competition [8]. The datasets include 2152 images that can be further classified into three categories: (1) 235 images of leaves belonging to healthy potato, (2) 980 images of leaves from early blight potato and 937 images of leaves belonging to late blight potato. Fig.1 shows the sample image of early and late blight potato leaves.

- **Healthy Blight Leaf:** A healthy potato leaf is usually green and smooth. There should be no apparent damage or discoloration, such as spots, blotches, or yellowing. The leaf's margins should be smooth and symmetrical, with no curling or twisting.
- **Early Blight Leaf:** It is a fungal infection which damages the leaves of potato plants, creating dark patches and early leaf loss.
- **Late Blight Leaf:** This is a fungal disease which damages the leaves of potato plants and tubers, creating black, water-soaked lesions and rotting.

3.2 Dataset Pre-Processing

In the beginning, we simply augmented the data to increase its volume. Data augmentation is an approach to expand the dataset by incorporating more variations to each training sample [9]. This helps to significantly increase the size of dataset and further reduce the over-fitting. During data augmentation, each image in the training set is gently moved, rotated, and scaled by different percentages before being added to the training set. This allows to accommodate

variations in the size, orientation, position and location of the object in their image. The images can also be edited for contrast and lighting and can be flipped both vertically and horizontally. By combining all these alterations, we have expanded our training set. We then generated images of size 256 with a batch size of 32, each with three channels and 50 epochs as shown in Figure 1.

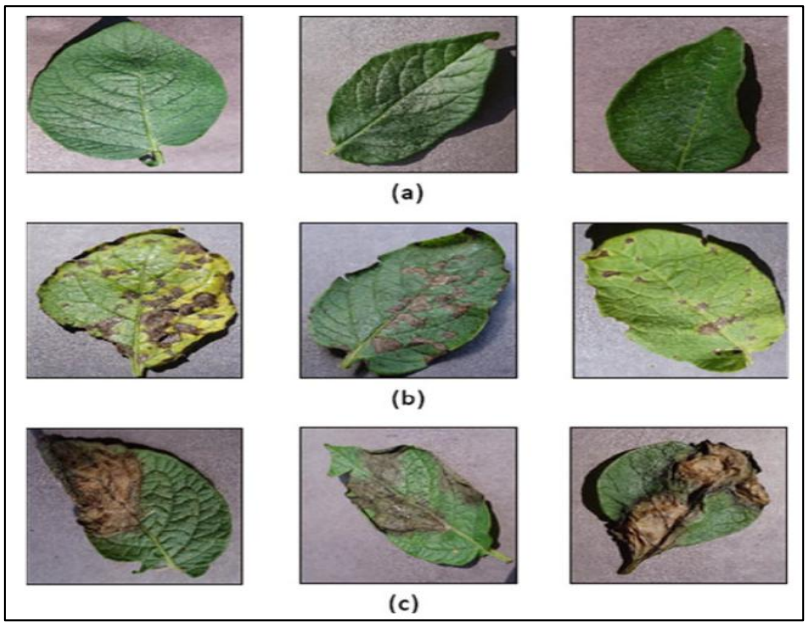


Fig. 1. Leaf blight sample photos, (a) healthy leaf category, (b) early blight leaf category, (c) late blight leaf category

3.3 Classification Techniques

Five machine learning classification approaches are utilized which are discussed in detail. Numerous studies are focused on machine learning classification to identify the diseases of potato plants. Potato diseases were identified using numerous machine learning algorithms. The following is a detailed explanation of these classifiers:

Convolutional Neural Network (CNN). CNN [10] is a deep learning model that was created to handle data with a structure resembling a grid, like photographs. It employs convolutional layers to retrieve features from input data with the pooling layers to lessen the spatial data size. Lastly, fully connected layers are used to make the classification decision.

Support Vector Machine (SVM). The SVM [11]algorithm is used to classify both linear and nonlinear data by selecting the best hyperplane that divides the classes with the largest margin. When the data is nonlinear, it is transformed into higher-dimensional space where a hyperplane can be used to distinguish the classes.

Random Forest (RF). RF [12] is collaborative learning approach that make use of multiple decision trees for improving the accuracy of the model. The absolute model prediction is determined by the class with the most votes, where each tree in the random forest predicts the class membership.

k-Nearest Neighbor (KNN). KNN [13] is a non-parametric approach that does not require training. In KNN algorithm during making a prediction, the test data point is classified into the most close data points. To achieve this, training set is searched for the k data points that are closest to the testing data point.

Logistic Regression. Logistic regression [14] is a statistical method which addresses binary classification problem where there are only two potential values for the response variable, which are commonly represented by the numbers 0 and 1. Logistic regression aims to ascertain the probability that a specific observation belongs to one of two groups by utilizing one or more predictor variables.

4 Proposed Approach

This section describes the research methodology used in this research work. The implementation of this work was done using Python on Google Colab Pro, with a GPU environment with 147.15 GB of disk space and 25.46 GB RAM. The system configuration has been shown in Table 2. Figure 2 illustrates the research methodology. In this study, the Potato leaf disease dataset was initially pre-processed using a data augmentation technique while developing different prediction models. The dataset is split into training and testing data comprising 80% training sets with 20% testing sets which is used to estimate the performance of machine learning algorithms in identifying potato leaf blight disease. While training the model, five machine learning classification algorithms such as SVM, KNN, RF, CNN, and LR are utilized. The trained models are further validated using 20% of test data. The performance of these prediction models is compared using the commonly used evaluation metric known as Accuracy which finds the ratio of total accurate predictions made by the classification model to the total number of predictions.

Table 2. Configuration of the System

Device name	DESKTOP-O61SD1M
Processor	Intel(R) Core (TM) i5-7200U CPU @ 2.50GHz 2.70 GHz
Installed RAM	4.00 GB
System type	64-bit operating system, x64-based processor
OS Name	Microsoft Windows 10 Home Single Language
System Model	HP Laptop 15-bs0xx

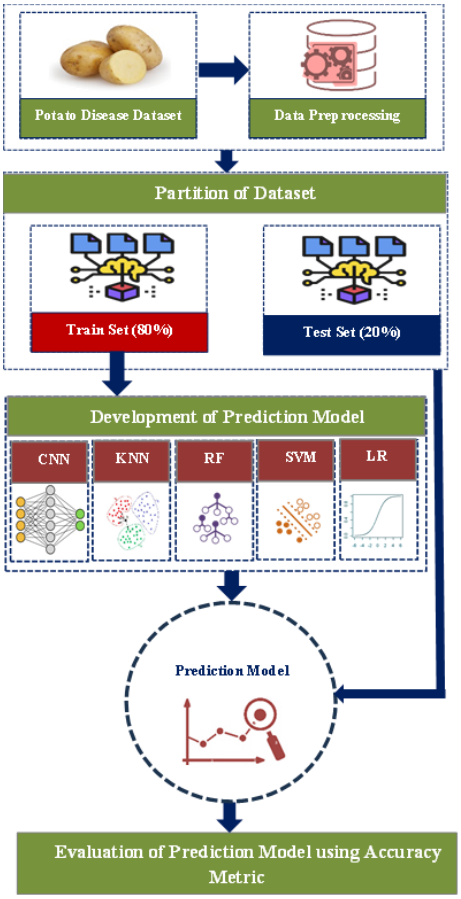


Fig. 2. Block Diagram of Proposed Methodology

5 Results and Discussion

In this study, a disease prediction model was developed using the Potato disease dataset to identify leaf blight diseases in potato plants. Due to the little size of the dataset, a data augmentation technique was applied. The dataset is split into 80-20 ratio where 80% corresponds to training set and 20% testing set. The trainset was applied to develop prediction models aimed at forecasting potato leaf blight disease using CNN, RF, KNN, LR, and SVM algorithms. All five trained models were then tested using the Test-Set. A comparison of these models was built using the "Accuracy" metric to select the most appropriate algorithm for this work. Results showed that the forecasting system based on CNN performed the best among all the five models with an accuracy of 95%, while KNN model performed the worst with 60% accuracy. The results are depicted in detail in Table 3. Therefore, it is depicted that CNN based prediction model can be effectively used in the farming industry for early prediction of leaf blight disease in potato plants.

Table 3. Experimental Results

Model		Accuracy
CNN		95%
RF		86%
LR		76%
SVM		94%
KNN		60%

6 Conclusions and Future Scope

Potato disease categorization is critical for guaranteeing potato crop health and output. Deep learning systems have shown promises for correctly and effectively diagnosing potato illnesses from leaf pictures. Ensemble approaches, transfer learning, hybrid CNN models, and active learning are some of the strategies that may be employed to enhance the performance and efficiency of classifying potato diseases. To execute these strategies, researchers and developers can use a variety of services like Kaggle, GitHub, TensorFlow, and PyTorch.

References

[1] Griffel, L.M., Delparte, D., Edwards, J.: Using Support Vector Machines classification to differentiate spectral signatures of potato plants infected with Potato Virus Y. Computers and Electronics in Agriculture. 153, 318–324 (2018)

[2] Tiwari, D., Ashish, M., Gangwar, N., Sharma, A., Patel, S., Bhardwaj, S.: Potato leaf diseases detection using deep learning. In: 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 461–466 (2020)

[3] Sharma, S., Anand, V., Singh, S.: Classification of diseased potato leaves using machine learning. In: 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT). pp. 554–559 (2021)

[4] Ji, Y., Sun, L., Li, Y., Li, J., Liu, S., Xie, X., Xu, Y.: Non-destructive classification of defective potatoes based on hyperspectral imaging and support vector machine. Infrared Physics & Technology. 99, 71–79 (2019)

[5] Coulibali, Z., Cambouris, A.N., Parent, S.-É.: Cultivar-specific nutritional status of potato (Solanum tuberosum L.) crops. PLOS One. 15, e0230458 (2020)

[6] Aldhyani, T.H.H., Alkahtani, H., Eunice, R.J., Hemanth, D.J.: Leaf pathology detection in potato and pepper bell plant using convolutional neural networks. In: 2022 7th International Conference on Communication and Electronics Systems (ICCES). pp. 1289–1294 (2022)

[7] Kumar, S., Paswan, S.: Comparative study of ANN and KNN classifiers performance for detection of potato tuber diseases. 10, 327–330 (2022)

[8] Mishra, A.: Potato Dataset. Kaggle. https://www.kaggle.com/datasets/arajmishra/potato-dataset. Accessed 11 Aug 2023 (2022)

[9] Van Dyk, D.A., Meng, X.-L.: The art of data augmentation. Journal of Computational and Graphical Statistics. 10, 1–50 (2001)

[10] Kim, P., Kim, P.: Convolutional neural network. In: MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence. pp. 121–147 (2017)

[11] Suthaharan, S., Suthaharan, S.: Support vector machine. In: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. pp. 207–235 (2016)
[12] Cutler, A., Cutler, D.R., Stevens, J.R.: Random forests. In: Ensemble Machine Learning: Methods and Applications. pp. 157–175 (2012)
[13] Peterson, L.E.: K-nearest neighbor. Scholarpedia. 4, 1883 (2009)
[14] LaValley, M.P.: Logistic regression. Circulation. 117, 2395–2399 (2008)