

Analysing Machine Learning based Approaches for Detecting Late Blight Disease in Potato Crop

Shikha Choudhary and Bhawna Saxena

Department of Computer Science and Engineering & IT,
Jaypee Institute of Information Technology, Noida-201309, U.P. India
Email: connect.shikhachoudhary@outlook.com, bhawna.saxena@jiit.ac.in

(Received April 27, 2023, Accepted June 28, 2023)

Abstract: Agriculture is a significant contributor in the world economy. With the drastic change in the geographical conditions, the occurrence of extreme events like floods, droughts, heat waves, etc. are increasing, thereby harming crop yield. Additionally, crop yield is adversely impacted by crop diseases causing significant losses towards food production. Protecting against losses incurred by crop diseases can aid in improving food security as well as strengthening the economy. Traditional methods of crop disease detection are time and labor-intensive, whereas the use of machine learning (ML) based methods fastens up the process, thereby helping implement corrective actions at an early stage. Multiple ML algorithms find application in the field of crop disease detection. Nevertheless, there exists a need to investigate the accuracies of different ML algorithm with regard to disease detection for a specific crop and disease combination. The performance of three ML algorithms, namely Random Forest, Linear Discriminant Analysis, and k-Nearest Neighbors with respect to late blight disease in potatoes was investigated in this work.

Keywords: Machine Learning; Random Forest Classifier; k-Nearest Neighbors; Late Blight.

1. Introduction

Agricultural products play a vital role in the global economy whilst being the primary source of livelihood for almost 50% of the world's population. The status of food security at the global level is alarming. With an exponentially increasing population it is extremely important to enhance productivity for sustainable development. Crop diseases leads to huge yield loss as it impacts the normal growing cycle of a plant and interrupts the processes such as photosynthesis, transpiration, germination, fertilization,

etc. There are different categories of crop diseases like viral, fungal and bacterial diseases and there are various factors like weather, soil, bacteria etc. which may cause the disease in the crop. The traditional methods of disease detection can be tiresome and ineffective, and by the time a disease is actually identified, there is no way to reverse the damages done by the disease. Also, a bit of expert advice is a must for accurate detection which further adds up the cost for farmers. So, it is very important to use the latest technologies available to do rapid disease detection in order to take corrective measures on time and save huge losses due to crop disease.

Machine Learning (ML) is extremely popular these days as it is being implemented in various areas for making predictions, classification and forecasting etc. Agriculture is one of the areas where ML is being widely used for speeding up disease detection in crops by utilizing techniques of image processing, segmentation and feature extraction followed by classification using supervised/unsupervised or deep learning algorithms. Although deep learning-based classifiers like Convolutional Neural Networks (CNNs) are being used prevalently, nevertheless non-deep learning-based classifiers like Random Forest (RF), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) have also been reported to show promising and comparable results. All the classifiers come with their own strengths and weaknesses. Choosing a classifier is mostly dependent on the type of dataset being used. The objective of the proposed work is to investigate the accuracies of different ML algorithm with regard to disease detection for a specific crop and disease combination.

The key contributions of the presented work are (i) utilization of ML-based approaches for detecting late blight disease in potato crops, (ii) practiced three ML models for training and testing disease detection using healthy and diseased leaves, and (iii) examining the effectiveness of the three ML models for detecting late blight in potato crop.

The proposed work has been presented in six sections. A study of the existing literature survey is presented in Section 2. Section 3 presents the essential knowledge in this direction. Section 4 discusses the methodology used. Section 5 of the paper discusses the experimental design and findings. The work is concluded in Section 6.

2. Related Work

A.A. Joshi et al.¹ proposed two models one with KNN and the other one with Minimum Distance Classifiers (MDC). Firstly, features were extracted using the image processing technique and the extracted features were used as input for image classification. Then the available dataset was divided into

training and testing datasets and both were maintained separately for each disease. The proposed models were validated for four different diseases of rice. According to the results, KNN reported an accuracy of 87.02% whereas MDC reported an accuracy of 89.23%. S. Ramesh et al.² proposed a methodology using RF for classification. RF is reported to give the best accuracy when compared with KNN, Naïve Bayes, SVM, and Logistic Regression. All the mandatory steps of Image processing were implemented like dataset creation, image segmentation and feature extraction. Then, the model was trained with the acquired dataset and the images were classified into different categories. The proposed model reported an accuracy of up to 70.14%. A. Afaq et al.³ prepared a model which used Image processing and KNN, SVM and RF for the classification to predict the disease in Tomato Crop. The proposed methodology reported an accuracy of up to 97%. G. Geetha et al.⁴ also designed a model to identify Tomato crop diseases like TYLCV, bacterial spot and early blight using image processing technique and KNN algorithm for the classification. K. Tambe et al.⁵ implemented a model using Image processing for feature extraction and SVM, Decision Tree, KNN and Gradient Boosting for the classification. Among the four different classifiers implemented, Gradient Boosting depicted the highest accuracy of 98.54%. G.P.R. Kranth et al.⁶ have conducted an experimental analysis of different ML algorithms like RF, ANN, Naive Bayes, Decision Trees and K-Means for plant disease detection. Decision Trees reported the highest accuracy of 97.83% among all the classifiers. D. Tiwari et al.⁷ have designed a solution based on hybrid model using VGG19 and multiple classifiers like SVM, KNN, ANN, and Logistic Regression. among all the classifiers used, Logistic Regression achieved the highest accuracy of 97.8%.

S. J. Shiv Ram Dubey et al.⁸ have conducted the experiment by creating a model for disease detection in apple fruit. In the proposed model, a multi-class SVM classifier has been used for the classification which has reported an accuracy of 93%. H. Waghmare et al.⁹ have also presented a multi-class SVM classifier-based model for disease detection in grapes which reported an accuracy of 96.6%. A. Singh et al.¹⁰, P. B. Padol et al.¹¹ have also designed a model using multi-class SVM and used K-means algorithm for Image processing. The proposed model was able to report 95.99% accuracy for potato image dataset and 88.89% on grapes image dataset respectively.

Bo Liu et. al.¹² proposed a model using multi-class SVM and multi-class LDA for disease detection in the crops/weed. SVM reported an accuracy of more than 90% whereas LDA reported an accuracy of more than 85%. According to the comparison, SVM performed better as its capable of non-linear classification as well.

Aravind K.R et al¹³ presented an approach using image processing for image segmentation and LDA, Naive Bayes (NB), SVM, and neural networks as the classification algorithms. LDA, NB, SVM and neural networks reported an accuracy of 86.09%, 76.52%, 86.96% and 86.04% for eggplant and tomato leaves, respectively.

Based on the survey of existing literature in the addressed subject area, it is evident that ML-based techniques are being widely applied for crop disease detection.

3. Preliminaries

This section briefly discusses the working of the three supervised ML-based classifiers used in this work, namely RF, LDA, and KNN.

Random Forest (RF)

It utilizes the concepts of bagging and aggregation. RF is an ensemble of multiple decision trees which work individually on the problem and finally their result is combined to provide the final output. It works by taking best average or maximum value. Votes are taken from all the decision trees used to make the best prediction, as all the decision trees work on slightly different features. It can be applied for classification as well as regression problems.

Linear Discriminant Analysis (LDA)

It works on the principle of projecting a feature from higher dimension to a lower dimension and divides it into two or more different classes or groups. LDA is also referred as discriminant function analysis or normal discriminant analysis. It is mostly used to for classification purpose.

k-Nearest Neighbors (KNN)

In this the dataset is divided into groups or classes by identifying k nearest neighbors. Neighbors are selected using distance formulas, like Manhattan, Hamming, Minkowski and Euclidean. Hamming is used for categorical variables and Manhattan, Euclidean and Minkowski work best for continuous variables. It can be utilized for both classification and regression operations. Let us assume two points, such as (a1, b1) and (a2, b2) in the two-dimensional coordinate plane, then the Manhattan Distance (d1) between these two points is given by (1) and the Euclidean distance (d2) is given by (2)

$$(1) \quad |a1-a2|+|b1-b2|$$

$$(2) \quad d=\sqrt{((a2-a1)^2+(b2-b1)^2)}$$

In the presented work, the performance of the three aforementioned ML classifiers is analyzed with regards to the detection of late blight disease in potato leaves.

4. Proposed Method

Proposed system for rapid disease detection in plants involves six stages as shown in Fig.1. The first phase includes acquiring potato images. The second phase prepares the image by rotating, resizing or cropping it. The third phase segments the image into various groups using multiple segmentation techniques. The fourth phase extracts the features using feature extraction methods. The last phase includes building the classifiers and identifying the leaf images as healthy or diseased leaf.

Image Acquisition

Image dataset was prepared for Potato Late Blight diseased leaf and Potato Healthy leaf. 200 images (including healthy and diseased) were downloaded from the PlantVillage dataset [14]. The downloaded dataset was separated into two categories: 80% images were put into training dataset and 20% images were put into test dataset.

- Total number of the Healthy Leaves = 100
- Total number of the Diseased Leaves = 100

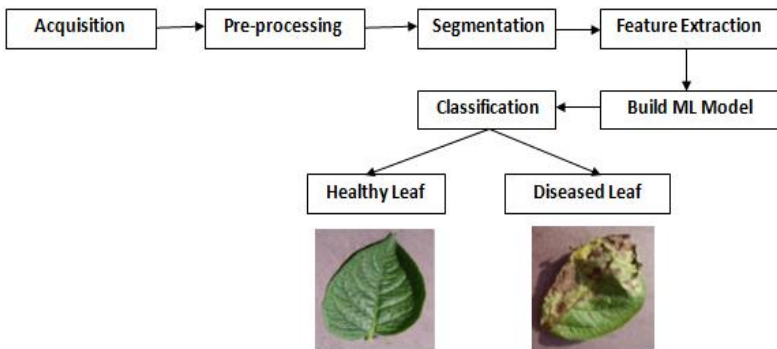


Fig. 1. Stages of proposed method

Image Pre-processing

Image Pre-processing prepares the input image by cropping, resizing or rotating the image, then segmenting the image for feature extraction. These extracted features are then used by the model for the classification.

Image Segmentation

Image segmentation divides the image into sub groups based on the different shapes of the object and edges in the image. The leaf image was segmented based on the masking of Lower Green & Upper Green in the image for Healthy and Lower Brown & Upper Brown for Disease and after combining the result of healthy and disease masking a final multi-dimensional image was generated.

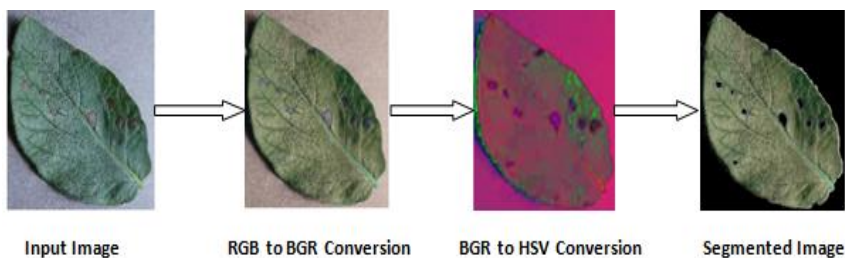


Fig. 2. Image segmentation

Feature Extraction

Features extraction involves extracting features from an image on the basis of texture, color and shape. The image is converted to grayscale and blurred to reduce the noise, so that the features can be extracted with more accuracy. The mean and standard deviation are calculated for color-based extraction and Haralik algorithm is used for texture-based extraction. Haralick algorithm works on correlation, skewness, contrast, inverse difference moments and entropy of the grayscale images.

ML-based Classification

Models were built using the three classifiers under consideration, namely RF, LDA and KNN. The models were then trained with 160 images (training dataset) and their performance was analyzed using the test dataset comprising of 40 images.

5. Experimental Outcomes

Google Colab has been used for experiments. Precision, Accuracy, Recall, and F1 score were used as the performance evaluation metrics.

Evaluation Metrics

The most common and understandable measures for evaluating the effectiveness of ML-based categorization are recall, precision accuracy, and F1 score. Based on the Confusion Matrix (Table 1), these performance measures have been computed.

Table 1. Confusion Matrix

		Ground Truth	
		Positive	Negative
Detected	Positive	True Healthy (TH)	False Healthy (FH)
	Negative	False Diseased (FD)	True Diseased (TD)

Based on the confusion matrix following metrics, (3)-(6), have been used:

$$(3) \quad \text{Accuracy} = (\text{TH} + \text{TD}) / (\text{TH} + \text{TD} + \text{FH} + \text{FD})$$

$$(4) \quad \text{Precision} = \text{TH} / (\text{TH} + \text{FH})$$

$$(5) \quad \text{Recall} = \text{TH} / (\text{TH} + \text{FD})$$

$$(6) \quad \text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Experimental Results and Observations

All the above-mentioned classifiers were trained and evaluated for classifying healthy and diseased leave images. Figures 3(a), (b) and (c) show the heatmaps generated for RF, LDA and KNN respectively.

The accuracy of each of the three classifiers is presented in Fig. 4. It can be observed that accuracy of RF is 100% on the test data which is the best accuracy and KNN has reported the accuracy of 82% which is the least among all. RF reported the highest accuracy among all three classifiers because RF consists of multiple decisions trees which work individually on different features. RF is an ensemble approach and uses the concept of bagging and randomness of features when it builds each and every individual tree to create an uncorrelated forest of trees. RF's prediction is based on maximum voting and is thus more accurate than that of any individual tree. RF also deals with the overfitting issues because all the decision trees are trained on slightly different features so they tend to overfit differently.

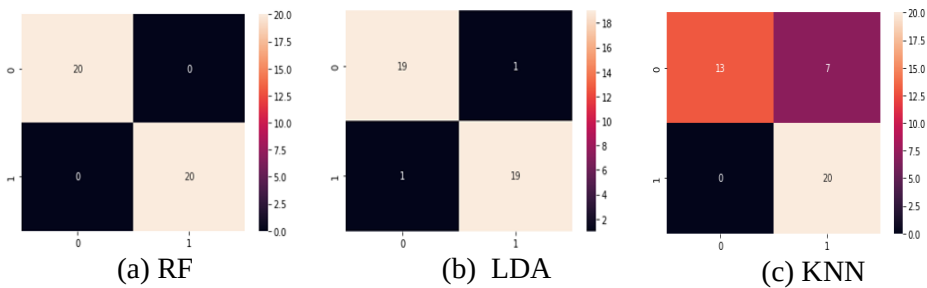


Fig. 3. Heatmap plots

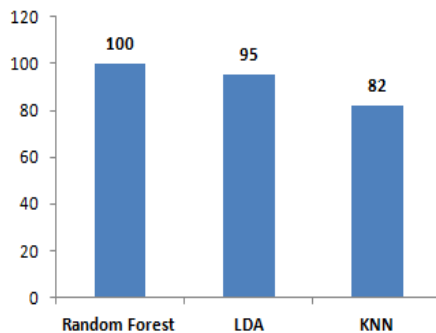


Fig. 4. Accuracy Comparison (in percentage)

Table II depicts the overall performance like precision, recall and F1-score of all the classifiers used for the classification. It can be clearly observed that RF outperforms both LDA and KNN in terms of overall performance.

Table 2. Table captions should be placed above the tables.

Classifier	Labels	Precision	Recall	F1-score
RF	Diseased	1.00	1.00	1.00
	Healthy	1.00	1.00	1.00
LDA	Diseased	0.95	0.95	0.95
	Healthy	0.95	0.95	0.95
KNN	Diseased	1.00	0.65	0.79
	Healthy	0.74	1.00	0.85

6. Conclusion

This work presents a quantitative comparison of the performance of three ML-based models, namely RF, LDA and KNN when utilized for late blight disease detection in potato crop, using Haralick feature extraction technique. Features such as moments, contrast, RGB color standard deviation, entropy, correlation etc. are extracted for this work. Among the given models, RF produced the highest level of accuracy (100%) trailed by LDA at 95% and KNN at 82%. So, the present work successfully investigates the accuracies of different ML algorithm for disease detection.

Future works might include different crops with different diseases and features. Enhancements in the prediction using the given models can be done by unique feature extraction from the plants' leaf images. Amalgamation of different algorithms can be used to get the most optimal value of the hyperparameters in the models for better predictions.

References

1. A. A. Joshi and B. D. Jadhav; Monitoring and controlling rice diseases using image processing techniques, *International Conference on Computing, Analytics and Security Trends (CAST) Proceedings*, (2016) 471-476.
2. S. Ramesh, R. Hebbar, M. Niveditha, R. Pooja, B. N. Prasad, N. Shashank, P. V. Vinod; Plant disease detection using machine learning, *International Conference on Design Innovations for 3Cs Compute Communicate Control Proceedings*, (2018) 41-45. doi: 10.1109/ICDI3C.2018.00017.
3. A. Afaq, J. Mukhtar, K. Vanishree; Leaf disease detection using ML and IOT, *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* 8(5) (2020) 1704-1710. doi: 10.22214/ijraset.2020.5276.

4. G. Geetha, S. Samundeswari, G. Saranya, K. Meenakshi, M. Nithya; Plant leaf disease classification and detection system using machine learning, *Journal of Physics: Conference Series* 1712 (2020) 012012. doi:10.1088/1742-6596/1712/1/012012
5. K. Tambe, D. Lohakare, M. Shinde, M. M. Deshpande; Detection of plant leaf disease using machine learning, *International Research Journal of Modernization in Engineering Technology and Science* 02(07) (2020).
6. G. P. R. Kranth, M. H. Lalitha, B. A. Mathur; Plant disease prediction using machine learning algorithms, *International Journal of Computer Applications*, 182(25) pp. (2018) 1-7. doi: 10.5120/ijca2018918049.
7. D. Tiwari, M. Ashish, N. Gangwar, A. Sharma, S. Patel, S. Bhardwaj; Potato leaf diseases detection using deep learning, *International Conference on Intelligent Computing and Control Systems (ICICCS 2020) Proceedings*, IEEE Xplore Part Number: CFP20K74-ART, (2020) 461-466. doi.org/10.1109/ICICCS48265.2020.9121067
8. S.R. Dubey, A.S. Jalal; Detection and classification of apple fruit diseases using complete local binary patterns, *IEEE 2012 3rd International Conference on Computer and Communication Technology*, (2012) 346-361. doi:10.1109/ICCCT.2012.76
9. H. Waghmare, R. Kokare, Y. Dandawate; Detection and classification of diseases of grape plant using opposite colour local binary pattern feature and machine learning for automated decision support system, *3rd International Conference on Signal Processing and Integrated Networks (SPIN) Proceedings* (2016) 513-518. doi:10.1109/SPIN.2016.7566749.
10. A. Singh and H. Kaur; Potato plant leaves disease detection and classification using machine learning methodologies, *IOP Conf. Series: Materials Science and Engineering Proceedings* 1022, (2021) 012121.
11. P. B. Padol, A. A. Yadav; SVM classifier based grape leaf disease detection, *IEEE International Conference on Advances in Signal Processing (CASP) Proceedings*, (2016) 175-179. doi: 10.1109/casp.2016.7746160.
12. B. Liu, R. Li, H. Li, G. You, S. Yan, Q. Tong; Crop/Weed discrimination using a field imaging spectrometer system, *Sensors* 19(23) (2019) 5154. doi:10.3390/s19235154.
13. A. K. Rangarajan, R. Purushothaman, M. Prabhakar, C. Szczepański; Crop identification and disease classification using traditional machine learning and deep learning approaches, *Journal of Engg. Research* (2021). doi:10.36909/jer.11941
14. D. P. Hughes and M. Salathe; An open access repository of images on plant health to enable the development of mobile disease diagnostics, *arXiv:1511.08060 [cs.CY]*, (2015). doi.org/10.48550/arXiv.1511.08060.