Potato Plant Leaves Disease Detection and Classification using Machine Learning Methodologies

Aditi Singh¹, Harjeet Kaur¹
¹School of Computer Science Engineering, Lovely Professional University, Punjab, India
¹aditisingh4614@gmail.com, ²harjeet.kaur@lpu.co.in

Abstract. Agriculture is one of the essential sectors for the survival of humankind. At the same time, digitalization touching across all the fields that became easier to handle various difficult tasks. Adapting technology as well as digitalization is very crucial for the field of agriculture to benefit the farmer as well as the consumer. Due to adopting technology and regular monitoring, one can able to identify the diseases at the very initial stages and those can be eradicated to obtain a better yield of the crop. In this document, a methodology was proposed for the detection as well as the classification of diseases that occur for the potato plants. For this scenario, the openly accessible, standard, and reliable data set was considered which was popularly known as Plant Village Dataset. For the process of image segmentation, the K-means methodology was considered, for the feature extraction purpose, the gray level co-occurrence matrix concept was utilized, and for the classification purpose, the multi-class support vector machine methodology was utilized. The proposed methodology able to attain an accuracy of 95.99%.

Keywords: Image Processing, Image segmentation, Plant disease detection using image classification, Kmeans, SVM.

1. Introduction
Agriculture is an essential sector in countries like India as those countries’ economy directly or indirectly dependent on agriculture. It indicates the necessity of taking care of plants from seedling until the expected crop obtains. Through this process, the crop needs to cross a lot of phases to obtain the expected crop such as weather conditions, the survival of the crop from various diseases, and the survival of the crop from various animals. Of these major phases, the crops can be protected from the various animals by providing proper protection for the field and this issue can be solvable. The next major issue is weather conditions which will not be in the control of humans, humans can only pray for better weather conditions to obtain a better crop. Finally, The major issue which is very crucial to protect the crop from various diseases as these diseases can impact the complete growth and yield of the crop. If one can able to identify these diseases in time, then the crop can be protected using appropriate fertilizers. If this process of identification and classification of diseases able to digitalize which would be helpful for the agriculturists. It will decrease the time for the identification of disease and precision in classifying the diseases.

There are a lot of significant crops exist in India, one among them is Potato. More than three-fourths of the population of India consumes potato daily at the same time it is one of the popular yielding crops in India. Yet, the yield of the potato crop can be diminished due to various diseases such as late blight and early blight. These diseases are also known as Phytophthora Infestans and Alternaria Solani respectively in scientific terms. Timely identification and classification of these diseases will lead to avoid the yield as well as financial losses. The popular way of identification of these diseases through the utilization of
the human eye for decades. But, this methodology arises with certain infeasibilities such as overtime will be taken for processing and shortage of experts at fields in remote locations. Therefore, the image analysis turned out to be an efficient methodology that will play a vital role in monitoring as well as the identification of the plant disease conditions effectively. Because the visible patterns are available on the plant leaves and patterns will be identified using various image processing methodologies for obtaining a particular pattern corresponding to a disease which will create an impact in the identification of various diseases. Thus the obtained features or patterns will be compared with the historical data and able to classify the disease which can be done by using various machine learning methodologies. So, the combination of image processing and machine learning is very effective in the identification and classification of diseases.

The present document was structured into various sections such as section-1 introduces the necessity of plant leaves diseases detection using image processing and machine learning methodologies, section -2 deals with the study of previous research based on identification and classification of diseases, precisely, a literature review, section-3 discusses the various methodologies necessary for the detection and classification of plant leaves, section-4 discusses the generated results based on the proposed framework and evaluation metrics, lastly, section-5 generates the conclusion of the presented work and future work based on this framework.

2. Literature Review
Monzurul Islam et al. 2017[1] proposed an approach that combines the processing of images and machine learning to allow leaf image disease to be diagnosed. This has been an automated system that categorizes potato plant diseases as well as unaffected leaves from the public image, known as ‘Plant Village’. The mentioned segmentation process and classification process through support vector machine methodology displays classification of images about 300 and the accuracy of the proposed model about 95%. Thus, the proposed approach offers a way for the automatic diagnosis of plant diseases on a huge scale. The multiclass SVM image segmentation is used for designing a system that is automated and easy to use. For primary diseases in potatoes such as Late Blight and Early Blight, a little computational effort is identified. The approach would provide framers with viable, reliable, and successful methodology and time-saving processes for disease identification.

Harshal Waghmare and Radha Kokare 2016[2] proposed technologies for plant disease detection analysis and pattern detection of the leaf texture. This work is based on the method of detection of grapes leaf disease. The device is used as input and segmentation on a single plant leaf after context removal is carried out. The image segmentation of the diseased component is then analyzed using a high-pass for the leaf. Special sectional leaf texture is obtained. Locally based fractal features nature invariant provides a good model of texture. The texture would be different for each independent illness. The texture pattern then extracted is graded by multiclass SVM. Multiclass SVM implementations are formulated to identify the diseases observed in grape plants for the processing of DSS (Decision Support Systems) automated and farmers, easily available. The Scheme performs segmentation and examination of a single leaf and the diseased portion of the leaf is observed by the high pass filter. A fractal-based retrieval of segmented leaf texture function that is invariant locally in nature and the strong texture module then provides. The texture removed pattern is then classified as an SVM designation for multiclass in groups who are healthy or ill classes respectively. The study concentrates on major widely encountered diseases grasses are downy mildew & black red. The recommendation approach quickly provides farmers with guidance from agricultural experts with 96.6 percent of accuracy.

Shima Ramesh and Mr. Ramachandra Hebbar et al. 2018[3] uses techniques in leaf-based image detection approaches that have impressive findings that have been shown. Random forest is included in this article for the detection of healthy and bad leaves for creating data sets. Our paper contains several phases Identifying dataset, function extraction, Identifying dataset, function extraction, classifier, and classification preparation. The data sets generated diseased and stable leaves are trained jointly under Random forest for the grouping of bad and healthy videos. Extracting image properties we use the histogram oriented gradient(HOG). In general, the use of machine learning to train wide available
data sets provides us with a simple way of detecting the occurring disease in plants. The model has been trained with Random Forest Classifier 160 papaya leaves images. The model can be categorized with an approximate 70 percent accuracy. Accuracy can be increased with a good number of images and using other local characteristics with global characteristics For example SIFT (Scale Invariant Feature Transform), SURF (Speed Up Robust Features), and DENSE along with BOVW (Bag Of Visual Word). Mrs. Shruthi U et al. 2019 [4] proposed machine learning methods to classify diseases and it applies mainly to data and It prioritizes itself and the performance of those tasks. This paper demonstrates the phases of general identification of plant diseases and the machine learning method on comparative research is for plant disease identification such as the acquisition of image-set, processing of the obtained data in the image-set, segmentation of the data in the image-set, feature extraction, and classification of images in image-set based on the extracted features and patterns identified. This survey shows that the Convolutional Neural Network high accuracy and more number of multiple disease crops. Rajleen Kaur and Dr. Sandeep Singh Kang 2015[5] proposed for automatic disease detection and disease part of the plant leaf images and also crop agriculture production. It’s achieved with computer advancement technology that allows agriculture to develop production. SVM is the latest classifier of the neural network approach and problem for the detection of accuracy. SVM is introduced in this article contains two datasets; one is a data collection for training and dataset of the train. The original image is first taken and used for processing. It offers, secondly, image pixels in black and background and hue section and saturation section is separated. Third, disease diagnosis and the unhealthy component are identified and a stable portion is segmented from it. This work also provides a percentage of the region where diseases arise and give us the disease name. As the image results the region impacted is 5.56%. This work gives accuracy which is stronger than the proposed results of the algorithm. Pooja V et al. 2017[6] proposed machine learning techniques for classification and disease detection and uses tools of image processing. Firstly, it captures the damaged region in the image and later it performs image processing. Segments in that image are generated and it recognized interested area and extraction of features is done. At last, through SVM results are sent and get new results. Disease classification is done by a support vector machine, methodology provides the best results than previously used techniques.

Jitesh P. shah et al. 2016 [7] describes an examination of various image processing and machine learning methodologies for recognition of affected plants by various diseases using images. Not only the document examined the various methodologies, but also concisely discussed crucial concepts of image processing and machine learning for the recognition of plant diseases as well as classifying them. This study was taken to the depth of considering 19 document-based experiments on various diseases related to rice, fruit, and various other plants. This includes the image data collection scale, no groups, pre-processing, technological segmentation, styles, classifier accuracy, etc. It also utilized the survey for further research and upgrade the detection and classification of various diseases related to rice plants.

Asma Akhtar et al. 2013[8] compared the results of various machine learning methodologies for the recognition and classification of disease affected plant leaves using their images. A three-phase methodology was introduced. The region of the leaf affected by the disease will be recognized through the segmentation of the leaf images. Thus, the obtained region will be passed on to a methodology that extracts the features and these features will be utilized for the classification of various diseases as well as healthy leaves. For the process of extraction of features the popular methodologies such as DCT and DWT were utilized and for the classification process SVM methodology was utilized. The proposed methodology able to attain an accuracy of 94.45%.

ZarreenNaowal Reza et al.2016[9] discuss stem diseases of jute plants that were detected which is one of the most valuable cash crop countries of Asia. An Android-based automated framework Request to take images of the disease was introduced affected stems and send to the dedicated Jute plants assay server. The impacted section on the server-side. Using the custom thresholding to segment the image on Hue-based segmentation formulation. The Relevant function values from the segmented section are derived for color co-occurrence technique texture analysis. The value derived is compared to the stored sample value in the predefined database that contributes to the disease classified, defined by using
Multiclass SVM. At the end stage, the results of the classification and the requisite control, steps are returned to the farmer in 3 seconds on their phone.

Sukhvir Kaur et al. 2018[10] explain about various parts of a plant the signs of plant diseases are clear, but the most often found leaves are for infection detection. Researchers tried to simplify the method of diagnosis of plant diseases and leaf image classification. Many researchers successfully employed and made a major contribution to computer vision technology. This manuscript sums up the benefits and drawbacks of all these experiments to illustrate some critical topics. A conversation on common infections in various stages of a disease and a research situation is presented. To classify those who appear to be, the success of state-of-the-art techniques is analyzed to fit well through several types of crops. The manuscript highlights a variety of suitable strategies along with future research paths, some areas of concern. The survey will help scientists accomplish computer vision understanding systems for the identification of plant diseases.

S.Ramesh and D.vydeki 2018[11] offers an algorithm for machine learning to find symptoms in rice plant about the disease. Automated plant identification, the Machine learning algorithm is used to carry out the disease. Images of safe leaves and blasting diseases damaged leaves are taken for the suggested method. The characteristics are removed safely and disease-prone portions of the rice leaf. The whole thing. Collection of 300 images for training and testing. The images are store under the proposed system and the leaf suggested were graded as either good or infected. The effects of the simulation give a 99 percent accuracy for bursts and 100 percent is during the training phase under normal images. The exams accuracy of the process for infected and healthy persons is 90 and 86 percent respectively. The above approach helps farmers in crops protection from disease. This approach gives you a free of disease and quality development of the crop. Finally, it would be proposed that the Indian farmers should use this approach to prevent the disease spread and make choices in the crop at all times for improving crop production and improving profitability.

V. Suresh et al. in 2020 [12] proposed a website based image processing and machine learning technological methodology. In this framework, firstly, the images will be gathered and then provided to the trained classifier through the website. Once the image is processed and extracted the features then the disease will be detected. Once the disease is detected, the website redirects the page into the page containing appropriate pesticides and chemicals with their usage and MRPs as well. It is one of the great work which identified so far as every researcher looking to detect and classify the diseases whereas this proposed methodology not only detecting the disease but also suggesting the appropriate pesticides suitable for it.

The above-discussed literature review indicating the importance of agriculture and the detection of disease at right time will not affect the very essential yield. The detection and classification of diseases in various plants will be followed with various phases such as data acquisition in the form of images, pre-processing of the obtained images using image processing methodologies, image segmentation for the identification of the region of interest, extraction of features, and finally, depending on the obtained patterns classifying the images. The methodologies are mainly discussed based on machine learning and image processing methodologies as the proposed framework is also based on those methodologies only.

3. Methodologies

Machine learning playing a vital role in the automation of various systems. Having that intention in mind, the proposed framework was planned based on the machine learning methodologies. Particularly in the scenario of detection and classifying the images into various categories of diseases. This section was organized such that firstly, the discussion will be done based on the system requirements and data acquisition of the data that utilized for the present proposed methodology. Secondly, the discussion will be based on image segmentation. Thirdly, the discussion will be based on feature extraction. Fourthly, the discussion will be on the process of classification. Finally, the discussion will be based on the proposed methodology.
3.1. Dataset and System Description
The collected data from an openly accessible database of images: ‘Plant Village’ that consists of diseased as well as healthy images about 54,306 from the total crop species about 14. From that particular database, potato species-related data was considered to implement the proposed framework. The obtained dataset consists of 300 potato plant leaves which were categorized into 3. They are as follows:

- The leaves having a disease called Late Blight
- The leaves having a disease called Early Blight
- The leaves are in a healthy state

The database of images consists of healthy leaves about 100 and disease affected leaves about 200. The database of images was divided into two databases such as the training database and the testing database. The training database consists of 70% of the image database i.e, 210 images and the testing database consists of the remaining 30% of the image database i.e, 90 images.

The proposed framework was implemented on the operating system windows 10 with the processor of Intel®Core™ i3-8130U CPU @ 2.20 GHz – 2.21GHz with RAM of 8 GB. The framework was implemented using Python.

3.2. Image Segmentation
The data considered are colored images and this phase is essential to separate the infected regions from the existing plant leaves. The sample of three images was mentioned as shown in figure-1. In image segmentation, the portion which will be segmented is referred to as the region of interest or ROI. Obtaining ROI is very crucial for the affected leaves to grab the patterns available for each of the diseases. It was implemented by using the K-means algorithm. The algorithm was implemented through the following steps:

Step-1: Determine the number of clusters
Step-2: The centroid of each cluster will be randomly initiated, and the number of centroids considered is equal to the number of clusters determined.
Step-3: Distance metric will be applied to each cluster between each of the points and the centroid of that cluster.
   - Once all the distances calculated in a cluster, consider the mean of those distances
   - Update the centroid with the mean value.
Step-4: Repeat the above step until the updation of the centroid of each cluster is very minimal or no change.

Various methods are used for distance metrics such as Minkowski distance, Manhattan distance, and Euclidean distance as mentioned in equations- 1 to 3 respectively.

\[
\text{Dist}(X, Y) = \left( \sum |x_i - y_i|^r \right)^{\frac{1}{r}} \quad (1)
\]

\[
\text{Dist}(X, Y) = \left( \sum |x_i - y_i| \right) \quad (2)
\]

\[
\text{Dist}(X, Y) = \left( \sum |x_i - y_i|^2 \right)^{\frac{1}{2}} \quad (3)
\]

Minkowski distance method is a highly generalized equation in terms of distance metrics. Other distance metrics can be derived from the Minkowski distance.
Figure 1. The three Sample leaves of potato are (a): leaf affected by Light Blight (b): leaf affected by Early Blight (c): leaf unaffected (Healthy)

3.3. Feature Extraction
When image segmentation was done and obtained the region of interest. From this region of interest, the features will be obtained. The feature set for the complete data set will be very huge. So, the important features need to be extracted for the scenario of classification. Feature extraction can also be considered as a dimensionality reduction process as it extracts the important information from a complete feature set such that no information misses out. It also speeds up the learning and generalization of the process. The features extracted using the concept of the Gray Level Co-occurrence Matrix using various statistical metrics.

3.4. Classification Process
Depending on the features extracted, the images need to be classified into various categories such as Late Blight, Early Blight, and healthy leaves. For the classification purpose, the support vector machine algorithm was utilized. It is a supervised learning methodology which utilizes hyperplanes for the classification purpose. It is a part of an optimization problem and it is defined as mentioned in the equation-4. The training set can be represented as $(X_k, Y_k), K = 1, 2, ..., i$ where $X_k \in R^n$ and $Y_k \in \{1, -1\}^i$.

$$\min_{w, f, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^{i} \xi_k \text{ subject to } Y_k = (w^T \phi(X_k) + e) \geq 1 - \xi_k, \xi_k \geq 0$$

(4)

The above-mentioned principle is utilized for binary classification. But, in this particular scenario, the classification is not binary, so the existing process will be generalized to obtain the multiclass classification. The generalization will be done in such a way that one category is considered into one class and the other two categories are considered as another class and it will be repeated for the number of categories.

3.5 Proposed Methodology
The process of detection of diseases from plant leaf images involves various steps and each of those steps can be discussed as follows.
Step 1: Image Acquisition: It deals with the acquisition of data from reliable sources to maintain the standard and stability so that it can be compared or extended for future studies.

Step 2: Image Pre-processing: It is a very essential phase of the framework. In this phase, mainly deals with the denoising of the image, enhancement of the image, and maintaining standard image size for all the images. Denoising and enhancement of images are essential to get a better result while segmenting the images.

Step 3: Image Segmentation: In this phase, the image will be segmented according to the region of interest. Here, in this case, the region of interest is the regions on the leaf which are affected by various diseases that need to be separated from the existing images.

Step 4: Extraction of Features: Depending on the obtained region of interest need to identify the patterns that exist. A different region of interests will have different patterns, from that scenario, one can able to extract features that are crucial in deciding the detection as well as classification.

Step 5: Evaluate the Affected Region: By comparing the region of interests and extraction features, one can able to evaluate the affected regions to obtain better accuracy in the model, otherwise there exist higher deviations.

Step 6: Processed Data: All the information related to processed image data by using the steps 1 to 5 will be gathered into a single location.

Step 7: Training Data: The training data will be obtained from the processed data. About 75% of the data with random indexing was considered to train the classifier model.

Step 8: Testing Data: The testing data will also be obtained from the processed data. About 25% of the data with random indexing was considered to test the classifier model.

Step 9: Classification: Test data will be provided to the trained classifier to classify the images into various categories such as Late Blight, Early Blight, and Healthy.

Step 10: Evaluation Metrics: Depending on the obtained results from the classifier model, the evaluation metrics such as precision, recall, F1-score, and accuracy will be obtained.

**Figure 2. Flow Chart of the Proposed Methodology**

**4. Results and Discussion**

The dataset considered was openly accessed standard dataset and it was divided into the training dataset consists of 210 images and the testing dataset consists of 90 images. The K-means algorithm was utilized
for the image segmentation, the Gray Level Co-occurrence Matrix utilized for feature extraction, and multi-class support vector machine methodology along with linear kernel function was utilized for the classification of potato leaves. The proposed framework evaluated using certain evaluation metrics such as precision, recall, F1-score, and accuracy. The proposed model achieved an overall accuracy of about 95.99%, the precision of about 96.12%, the recall of about 96.25%, and the F1-score of about 96.16%.

Table 1. Evaluation Metrics of Proposed Methodology

| Category Name  | Precision (%) | Recall (%) | F1-score (%) | Accuracy (%) |
|----------------|---------------|------------|--------------|--------------|
| Late Blight    | 91.07         | 95.41      | 93.29        | 94.71        |
| Early Blight   | 98.36         | 94.71      | 96.43        | 96.84        |
| Healthy        | 98.93         | 98.62      | 98.76        | 96.43        |
| Overall        | 96.12         | 96.25      | 96.16        | 95.99        |

The comparison of evaluation metrics can be seen as mentioned in the figure-3. One can identify the accuracy of the Late Blight category as low when compared to the other two categories whereas the accuracy of the healthy leaves category is higher than the other two categories. The overall accuracy of the model is satisfactory concerning the precision, recall, and F1-score obtained for the proposed model.

5. Conclusion
Digitalization increasing across all the fields and it is high time to adopt digitalization into the field of agriculture as well to obtain better protection in terms of growth and yield. Keeping this intention as the motivation for the proposed model to detect and classify the affected and unaffected leaves of potato. The proposed framework able to achieve an accuracy of 95.99%. Yet, this accuracy needs to be improved. The existing work further can be extended by using artificial neural networks, particularly, convolutional neural networks. These days, a lot of research related to images is happening based on CNN methodologies to obtain better and reliable accuracy. The concept of activation functions, batch normalizations, convolutional layers, and fully connected layers are playing a key role in CNN architectures to attain better accuracy.
References.

[1] M. Islam, A. Dinh, K. Wahid, and P. Bhowmik, Detection of potato diseases using image segmentation and multiclass support vector machine 2017, Can. Conf. Electr. Comput. Eng., 8–11.

[2] H. Waghmare, R. Kokare, and Y. Dandawate, Detection and classification of diseases of Grape plant using opposite color Local Binary Pattern feature and machine learning for automated Decision Support System 2016, 3rd Int. Conf. Signal Process. Integ. Networks, 513–8.

[3] S. R. Maniyath et al., Plant disease detection using machine learning 2018, Proc. 2018 Int. Conf. Des. Innov. 3Cs Comput. Commun. Control., 41–5.

[4] U. Shruthi, V. Nagaveni, and B. K. Raghavendra, A Review on Machine Learning Classification Techniques for Plant Disease Detection 2019, 2019 5th Int. Conf. Adv. Comput. Commun. Syst., 281–4.

[5] R. Kaur and S. S. Kang, An enhancement in classifier support vector machine to improve plant disease detection, Proc. 2015 IEEE 3rd Int. Conf. MOOCs, Innov. Technol. Educ.,135–40.

[6] V. Pooja, R. Das, and V. Kanchana, Identification of plant leaf diseases using image processing techniques 2018, Proc. - 2017 IEEE Technol. Innov. ICT Agric. Rural Dev., 1, 130–3.

[7] J. P. Shah, H. B. Prajapati, and V. K. Dabhi, A survey on detection and classification of rice plant diseases 2016, 2016 IEEE Int. Conf. Curr. Trends Adv. Comput.

[8] A. Akhtar, A. Khanum, S. A. Khan, and A. Shaukat, Automated plant disease analysis (APDA): Performance comparison of machine learning techniques 2013, Proc - 11th Int. Conf. Front. Inf. Technol. FIT 2013, 60–5.

[9] Z. N. Reza, F. Nuzhat, N. A. Mahsa, and M. H. Ali, Detecting jute plant disease using image processing and machine learning 2017, 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol.

[10] S. Kaur, S. Pandey, and S. Goel, Plants Disease Identification and Classification Through Leaf Images: A Survey 2019, Arch. Comput. Methods Eng., 26, 507–30.

[11] S. Ramesh and D. Vydeki, Rice blast disease detection and classification using machine learning algorithm 2018, Proc.- 2nd Int. Conf. Micro-Electronics Telecommun. Eng., 255–9.

[12] Suresh V, Gopinath D, Hemavarthini M, Jayanthan K, and Mohana Krishnan, Plant disease using image processing 2020, IJERT, 9, 78-82.

[13] Kiani, E., & Mamedov, T. (2017). Identification of plant disease infection using soft-computing: Application to modern botany. Procedia Computer Science, 120, 893–900. https://doi.org/10.1016/j.procs.2017.11.323

[14] 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. https://doi.org/10.1016/j.inpa.2016.10.005

[15] Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145(January), 311–318. https://doi.org/10.1016/j.compag.2018.01.009

[16] Golhani, K., Balasundram, S. K., Vadamalai, G., & Pradhan, B. (2018). A review of neural networks in plant disease detection using hyperspectral data. Information Processing in Agriculture, 5(3), 354–371. https://doi.org/10.1016/j.inpa.2018.05.002

[17] Maththalli, C., Gedefaye, E., Endalamaw, F., & Necho, A. (2018). Plant leaf diseases detection and auto-medicine. Internet of Things, 1–2, 67–73. https://doi.org/10.1016/j.iot.2018.08.007

[18] Pantazi, X. E., Moshou, D., & Tamouridou, A. A. (2019). Automated leaf disease detection in different crop species through image features analysis and One Class Classifiers. Computers and Electronics in Agriculture, 156(July 2018), 96–104. https://doi.org/10.1016/j.compag.2018.11.005