Tomato Leaf Disease Detection using Machine Learning Model

Lekha J[1], Saraswathi S[2], Suryaprabha D[3], Noel Mathew Thomas[4]

{lekha.j@christuniversity.in[1], nascsaraswathi@nehrucolleges.com[2], spayrus@gmail.com[3]}

Christ University, Lavasa, Pune[1], Nehru Arts and Science College, Coimbatore[2,3]

Abstract. Agriculture is the primary source of employment for over half of India's population, making it heavily dependent on this sector. Indian farmers encounter a plethora of challenges during their agricultural pursuits, which include but are not limited to droughts, pests, infertile land, lack of irrigation, and plant diseases. As per reliable reports, plant diseases and pests are accountable for crop losses amounting to 5000 crores annually in India, rendering them a significant apprehension for the farming community. Plant disease identification can be a cubersome task and this paper aims to develop a disease identification model for Tomato leaves using three different Machine Learning algorithms, namely Convolutional Neural Network (CNN), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The primary goal is to evaluate and compare the performance of each algorithm for the identification of Tomato leaf diseases.

Keywords: Machine Learning, Plant disease detection, KNN, CNN, SVM, Tomato leaf disease detection

1 Introduction

India is the world's second largest producer of tomatoes, with a production of over 20 million metric tons annually. The country has diverse agro-climatic conditions that make it suitable for tomato cultivation throughout the year. The central tomato-growing states in India include Maharashtra, Karnataka, Andhra Pradesh, Madhya Pradesh, and Gujarat. The tomato industry also provides employment opportunities for millions of people in the country.

Tomato plant diseases have a significant impact on crop yields and quality. They can cause leaf wilting, fruit rot, and decreased plant vigour. Plant diseases are typically manifested in the plant's leaves, stems, and fruit. Therefore, the plant leaf is often used to detect the presence of disease as it displays the most prominent symptoms

IACIDS 2023, November 23-25, Lavasa, India
Copyright © 2024 EAI
DOI 10.4108/eai.23-11-2023.2343189

This paper discusses various tomato leaf diseases, including but not limited to late blight, target spot, mosaic virus, yellow leaf curl virus, bacterial spot, septoria leaf spot, and spider mites The conventional method of detecting plant diseases with the naked eye can be both time-consuming and laborious. Farmers may need more knowledge about the various diseases that can impact their crops.Over the years, various methods have been utilized to design automatic or semi-automatic systems for detecting plant diseases. Detecting diseases through symptoms on plant leaves has made the process simpler and cost-effective. These systems have proven to be faster, cheaper and more precise compared to the traditional method of manual observation by farmers. Image processing can assist in bridging this knowledge gap and also save time and labor. This paper aims to implement various classification techniques and find the best model for image segmentation with the highest accuracy.

The different methods of image classification explored in this paper are K-nearest Neighbors(KNN), Support Vector Machines(SVM) and Convolutional Neural Networks (CNN). The dataset being used is sourced from the Plant Village dataset and contains 16,012 images (consisting of 10 different labels).

2 Literature Review

Lavika Goel et. al.[5] stated that the recent developments in deep learning and deep learning technology have enhanced agriculture by visualizing the soil properties or weather conditions for crop disease identification and classification, providing more accurate solutions than conventional methods. Plant diseases can be caused by infectious agents (caused by fungi, bacteria, viruses, nematodes and protozoa) and non-infectious agents(physiological factors such as temperature and weather conditions, Nutrition deficiency, etc). The diseases in plants can be recognized by observing the plant's root, leaf and stem portions. The study employed various segmentation and feature extraction techniques to identify regions of interest and extract meaningful information for disease classification. The analysis revealed that the most commonly used feature extraction methods were Grey Level Co-occurrence Matrix (GLCM), Histogram of Gradient (HOG), and Oriented FAST and Rotated BRIEF (ORB), while the most frequently used machine learning-based classifier was SVM due to its robustness to interdependent input features and non-Gaussian data distribution. The accuracy rates achieved by certain combinations of feature extraction and classification methods were 99.98% (ORB and Linear SVM), 99.62% (GLCM and MobileNet), and 99.74% (GLCM and InceptionV3).

O. Kulkarni [10] employs computer vision techniques along with deep learning neural networks to train a model that can identify the type of disease based on the visual and textural attributes of leaves that share similarities. The dataset of 38 different classes contains 54306 images of 13 different types of crops and 26 types of diseases. The images are pre-processed and are divided into 80%-20% testing and training datasets before being fed into the Convolutional Neural Network(CNN) model. Transfer learning is used to build deep learning models using Mobile Net and Inception V3 pre-trained models. The segmented images for training the model resulted in better performance than color and grayscale images. The MobileNet and InceptionV3 models showed good performance with an accuracy of 99.62% and 99.74%, respectively, for crop type detection. The InceptionV3 model outperformed MobileNet in crop detection, while both models showed steady growth with 99.04% and 99.45% accuracy, respectively, for crop disease detection.

Panigrahi K.P et. al. [11] implements techniques such as image processing, feature extraction, and deep learning algorithms to detect and classify the diseases. CNN, SVM and Random Forests have been applied to achieve high accuracy rates. The Maize plant disease dataset from the plant village dataset contains 3823 images of four different class labels. The implementation is done using the python machine learning package along with pandas. The accuracy rates for the preprocessed images using different machine learning algorithms were: SVM at 77.56%, Naive Bayes at 77.46%, K-Nearest Neighbors at 76.16%, Decision Tree at 74.35%, and Random Forest at 79.23%.

Jain B et. al [4] used a pre-trained deep CNN based VGG-16(Visual Geometric Group), that has been trained on over 10 million images on detecting generic features from images, for feature extraction. Random forests have been used as a classifier because of its robustness and accuracy. The dataset contains 1003 images from three different classes of grape leaf diseases. The accuracy rates for the preprocessed images(resizing, normalization and conversion to arrays) of the model for: 80%-20% training – testing dataset is 91.66%, 70%-30% training – testing dataset is 90.68% while 60%-40% split provided an accuracy of 89.21%.

Costales H et.al [12] implemented a Convolution Neural Network(CNN) to design a rice leaf disease detection algorithm. The Feature-driven Development (FDD) under the family of Agile methodology was used for the software model to extract meaningful attributes from the images. The dataset by Huy Minh Do of 1260 labeled rice leaf images was downscaled to 500x500 pixels and are split into training and validation datasets(80%-20%) to train the model. The model provided an accuracy of 98% in its third iteration.

Ramesh S et. al [13] created Histogram of an Oriented Gradient (HOG) for extracting features of diseased leaf images while using Random forests for classification.The researcher utilized three feature descriptors, namely Hu Moments, to extract essential attributes of image pixels for object description. Before feature extraction, the RGB images were converted into grayscale. For distinguishing between healthy and diseased leaves based on texture, the researcher employed the Haralick texture feature, which uses an adjacency matrix to store the position of (I, J). Additionally, a color histogram was used to represent the color distribution in the image. The RGB format was initially transformed into HSV color space, then the histogram was calculated. The Random forest classifier was trained using 160 images of Papaya Leaves and provides an accuracy of 70%.

M Islam et. al.[14] proposes an automated potato leaf disease detection using Support Vector Machines. The algorithm is trained using 300 potato leaf images sourced from the plant village dataset. The images are masked and the region of interest(ROI) is extracted based on color and texture, which is performed using the Color Thresholder app in Matlab. To extract statistical texture features such as contrast, correlation, energy, and homogeneity, the researcher utilized the Gray Level Co-occurrence Matrix (GLCM).The Support Vector Machine model with a train-test split of 60%-40% achieved a testing accuracy of 95%. In order to enhance the model's robustness, 5-fold cross-validation was employed, resulting in an accuracy of 93.7%.

Khalili E et. al [15] focused on detecting and classifying soybean charcoal rot disease using various machine learning ,ethods. The disease caused by Macrophomina phaseoli (Tassi) Goid, is a significant contributor to reduced soybean productivity. Padol P B et. al [16] applies support vector machine for classifying the healthy and unhealthy leavevs.

3 Machine Learning Classification Methods

3.1 K-Nearest Neighbors (KNN)

KNN is a simpler and easier implementation algorithm used for data distribution based on distance metric. The accuracy of the algorithmic classification is based on the similarities among labels. The distance metric is used for measuring the similarities among labels and Minkowski distance is used for this purpose which calculates the k closet neighbors to a required data point. If the ‘k’ is a small value then it means overfitting and higher value means underfitting.

3.2 Convolution Neural Networks (CNN)

CNN is a collection of one or more layers which performs a particular task. Filters are applied in a specific layer to extract the required features from the image. Filters are nothing but a small matrix which slide over the entire image pixels and perform dot matrix operation. This results in a new matrix which is referred as feature map. The features of an image may include gradient values, colour values, and edge information. This layer can learn spatial hierarchies of patterns and extract low-level features such as edges, corners, and curves.

The Pooling layer reduces the spatial size of the feature maps by down-sampling them, thereby reducing the computational power required to process the images by decreasing dimensionality. This layer operates by taking the maximum (max pooling) or average (average pooling) value from the region of the image covered by the kernel. Max pooling is preferred over average pooling as it performs de-noising of images and dimensionality reduction more effectively.

The Activation layer applies a nonlinear function (e.g. ReLU) to introduce non-linearity into the network, which is then fed to the fully connected layer. Backpropagation is applied to each iteration of the process, and the "epochs" help the model differentiate the features and classify the images using the Softmax classification technique.

3.3 Support Vector Machine (SVM)

Support Vector Machines use decision boundaries or hyperplanes to classify data points into distinct classes by mapping lower-dimensional data points into higher-dimensional space using kernel functions for linear regression. It is particularly effective in dealing with high-dimensional datasets and has proven to be a popular choice in the classification of diseases. It works by identifying the support vectors, which are the data points closest to the decision boundaries, and optimizing the margin between these support vectors to minimize classification errors.

The advantage of using SVM is its ability to handle non-linearly separable and imbalanced datasets.

4 Methodology

The methodology adopted in this proposed work is as follows

4.1 Dataset Description

The dataset contains 16,102 images of tomato leaves spread across 10 labels, collected under controlled conditions. The images have been sourced from the Plant Village dataset. The dataset was split into training (80% of images) and validation (20% of images) sets to build the disease detection model.

4.2 Image Pre-processing

Image pre-processing is crucial, while building a Machine learning model, as it enhances the quality of image, standardizes the data and reduces computational complexity. . In this study, the images were downscaled to 32x32 pixels, which effectively reduced the image size and processing time required by the model. Additionally, the images were converted from RGB to grayscale, resulting in a decrease in data size due to the reduction of color channels from three to one. The application of histogram equalisation enhanced the features of the data, while normalization prevented the larger pixels from dominating the training process.

4.3 Feature Extraction

VGG16 is a neural network object detection and classification algorithm. The researcher has used VGG16 consisting of sixteen layers, including thirteen convolutional layers and three fully connected ones. The model consists of three convolutional and max pooling layers of 3x3 pixels and 2x2 pixels, respectively. The model has two fully connected layers with 4096 neurons and a softmax activation layer with ten output neurons. The model is trained using the Adam optimizer with a learning rate of 0.001, and the loss function is categorical cross-entropy. It is designed to process input images of size 224x224x3 (RGB channels) and predict the class of the image from 10 different possible classes.

4.4 Model Evaluation

The machine learning classification techniques mentioned in this paper were implemented on the dataset after feature extraction using VGG16.CNN achieved an accuracy of 79.14% after 40 epochs of training (Figure 1). KNN and SVM achieved 74.56% and 68.22% of accuracy respectively

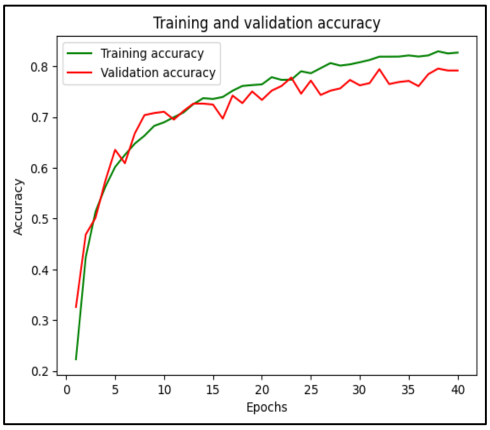

Fig. 1. Accuracy of CNN model over the epochs

Table 1. Accuracy of each model

Algorithm	Accuracy
CNN	79.14%
KNN	74.56%
SVM	68.22%

5 Conclusion

This research study centered on the development of a disease detection model for tomato leaves using machine learning techniques. The dataset utilized in this study comprised of 16,012 images of ten labels, obtained from the Plant Village dataset. The images were subjected to pre-processing and augmentation and were used to train CNN, KNN, and SVM models. Among the models, CNN achieved the highest accuracy levels, as evidenced by the results of previous studies reviewed in the paper. The study underscores the importance of feature extraction and pre-processing in the development of machine learning models for disease detection. The accuracy levels attained by the models are noteworthy and suggest the potential of developing a practical application for the detection of tomato leaf diseases. Further research can explore alternative extraction and optimization techniques to achieve even higher accuracy levels.

References

[1] D. V. Lindberg and H. K. H. Lee, “Optimization under constraints by applying an asymmetric entropy measure,” J. Comput. Graph. Statist., vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.

[2] B. Rieder, Engines of Order: A Mechanology of Algorithmic Techniques. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[3] I. Boglaev, “A numerical method for solving nonlinear integro-differential equations of Fredholm type,” J. Comput. Math., vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.

[4] CSE Department, Dr.NGP Institute Of Technology, D Gopinath, Hemavarthini, M., Jayanthan, K., & Krishnan, M. (2020). Plant Disease Detection using Image Processing. International Journal of Engineering Research & Technology (IJERT), 9(03), 5. http://www.ijert.org

[5] Zhou, C., Zhou, S., Xing, J., & Song, J. (2021). Tomato leaf disease identification by restructured deep residual dense network. IEEE Access, 9, 28822-28831.

[6] Peng Jiang , Yuehan Chen ,Bin Liu , Dongjian He , Chunquan Liang ,’ Real-Time Detection of Apple Leaf Diseases Using Deep Learning Approach Based on Improved Convolutional Neural Networks’, (Volume: 7), pp. 06 May 2019

[7] Jain, B., & Periyasamy, S. (2022). Grapes disease detection using transfer learning. arXiv preprint arXiv:2208.07647.

[8] Goel, L., & Nagpal, J. (2022). A Systematic Review of Recent Machine Learning Techniques for Plant Disease Identification and Classification. IETE Technical Review, 1-17

[9] Costales, H., Callejo-Arruejo, A., & Rafanan, N. (2023). Development of a Prototype Application for Rice Disease Detection Using Convolutional Neural Networks. arXiv preprint arXiv:2301.05528.

[10] India Agroportal. (2021).Tomato https://agroportal.lk/knowledge/indian-crop-overview/tomato/

[11] Rathore, S. (2021). Tomato Diseases: Symptoms, Prevention, and Treatment. Agriculture Review. https://www.agriculturereview.in/tomato-diseases-symptoms-prevention-and-treatment/

[12] Singh, V., & Singh, B. (2020). Automatic detection of plant diseases using image processing techniques: A review. Computers and Electronics in Agriculture, 169, 105153. https://doi.org/10.1016/j.compag.2020.105153

[13] O. Kulkarni, "Crop Disease Detection Using Deep Learning," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697390.

[14] Panigrahi, K. P., Das, H., Sahoo, A. K., & Moharana, S. C. (2020). Maize leaf disease detection and classification using machine learning algorithms. In Progress in Computing, Analytics and Networking: Proceedings of ICCAN 2019 (pp. 659-669). Springer Singapore.

[15] Costales, H., Callejo-Arruejo, A., & Rafanan, N. (2023). Development of a Prototype Application for Rice Disease Detection Using Convolutional Neural Networks. arXiv preprint arXiv:2301.05528.

[16] Ramesh, S., Hebbar, R., Niveditha, M., Pooja, R., Shashank, N., & Vinod, P. V. (2018, April). Plant disease detection using machine learning. In 2018 International conference on design innovations for 3Cs compute communicate control (ICDI3C) (pp. 41-45). IEEE.

[17] M. Islam, Anh Dinh, K. Wahid and P. Bhowmik, "Detection of potato diseases using image segmentation and multiclass support vector machine," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, Canada, 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946594.

[18] Khalili, E., Kouchaki, S., Ramazi, S., & Ghanati, F. (2020). Machine learning techniques for soybean charcoal rot disease prediction. Frontiers in plant science, 11, 590529.

[19] P. B. Padol and A. A. Yadav, "SVM classifier based grape leaf disease detection," 2016 Conference on Advances in Signal Processing (CASP), Pune, India, 2016, pp. 175-179, doi: 10.1109/CASP.2016.7746160.

[20] Bhimte, N. R., & Thool, V. R. (2018, June). Diseases detection of cotton leaf spot using image processing and SVM classifier. In 2018 Second international conference on intelligent computing and control systems (ICICCS) (pp. 340-344). IEEE.

[21] PlantVillage dataset. (n.d.). Retrieved February 27, 2023, from https://github.com/spMohanty/PlantVillage-Dataset/