MACHINE LEARNING FOR DETECTION AND PREDICTION OF TOMATO LEAF DISEASES

A REVIEW PAPER

Sandra Saji
Information Technology
Amal Jyothi College Of Engineering
Kanjirappally, India
sandrasaji0716@gmail.com

Melbin Mathew
Information Technology
Amal Jyothi College Of Engineering
Kanjirappally, India
melbinnelliyaniyil107@gmail.com

Angel Maria S
Information Technology
Amal Jyothi College Of Engineering
Kanjirappally, India
angel3008mary@gmail.com

Amrutha Mugesh
Information Technology
Amal Jyothi College Of Engineering
Kanjirappally, India
amruthamugesh777@gmail.com

Jincy Lukose
Asst. Professor,Information Technology
Amal Jyothi College Of Engineering
Kanjirappally,India
jincy.lukose@amaljyothi.ac.in

***Abstract*—Tomato, which is scientifically known as Solanum lycopersicum, is a widely cultivated plant in the nightshade family, Solanaceae. It is an important source of food, both fresh and in processed form, and is grown in many parts of the world. However, tomato plants are prone to various diseases, which can significantly reduce their yield and quality. Early detection and prediction of these diseases can help in timely treatment and management which can ultimately lead to higher crop productivity. Machine learning techniques have shown promise in detecting and predicting plant diseases. This approach can be used to improve the efficiency and effectiveness of tomato cultivation and can have a significant impact on the agricultural industry. The use of machine learning algorithms can increase the efficiency of tomato cultivation. In this study, we present a machine learning-based approach for the detection and prediction of tomato leaf diseases. We use a dataset of imagesof tomato leaves infected with different diseases such as tomato yellow curl virus, bacterial spot, and late blight along with healthyleaves, to train a Random Forest model. The model is then tested on a separate dataset to evaluate its performance.**

***Keywords*—Random forest, feature extraction, training data,testing data, and tomato leaf disease detection**

I. INTRODUCTION

Tomato plants are susceptible to a variety of diseases, which can cause significant losses in yield and quality. Early detection and prediction of these diseases are crucial to prevent their spread and minimize their impact. Random Forest is a popular machine learning algorithm that has been widely used in various fields, including disease detection and prediction in plants. In this context, the Random Forest model can be trained on a dataset of images of tomato leaves affected by various diseases, along with healthy leaves. The model can then be used to detect and classify the disease in real-time, by analyzing images of the tomato leaves. Random Forest model-based disease detection and prediction techniques can provide efficient solutions for tomato leaf diseases. Using Random Forest for tomato leaf disease detection can help improve crop yields, reduce crop losses due to disease, and contribute to more efficient agricultural practices.

A. Random Forest

Random Forest is a machine learning algorithm that is used for both classification and regression tasks. It is a type of ensemble learning method. It builds a large number of decision trees and combines their output to make final decisions. The algorithm works by creating a set of decision trees, each trained on a random subset of the data and a random subset of the features. This randomness helps to reduce overfitting and makes the algorithm more robust. During the training phase, the algorithm selects a random subset of the available features to determine the best split for

DOI: 10.5281/zenodo.8013583

DOI:

each node in each decision tree. In the prediction phase, each decision tree in the forest makes a prediction and the final prediction is the average of all the individual predictions. Random Forest can handle a large number of features andis relatively insensitive to noisy data. Random Forest has several advantages over other machine learning algorithms, such as high accuracy, robustness to outliers, and the abilityto handle both categorical and continuous data.

B. Feature Extraction

Feature extraction of leaves is the process of identifying and extracting relevant information from leaf images that can be used for various applications such as plant species identification, leaf disease detection, and leaf classification. The process of feature extraction involves converting raw leaf image data into a set of meaningful features that can be used totrain machine learning models for classification or clustering. There are various techniques for extracting features from leaf images, including color-based features, texture-based features, shape-based features, and vein-based features.

Color-based features involve extracting color information from the leaf image. This can be done using techniques such as color histogram analysis, where the distribution of colors inthe image is analyzed to extract features such as color intensity,color contrast, and color homogeneity.

Texture-based features involve analyzing the texture of the leaf surface. This can be done using techniques such as Gabor filters, where the texture of the image is analyzed to extract features such as texture contrast, texture entropy, and texture homogeneity.

Shape-based features involve analyzing the shape of the leaf. This can be done using techniques such as edge detection, where the edges of the leaf are detected and used to extract features such as leaf perimeter, leaf area, and leaf circularity. Once the features have been extracted, they can be used to train machine learning models for various applications such as leaf classification, plant species identification, and leaf disease detection.

C. Training Data

Training data is a set of examples used to train a machine learning model. It typically consists of a large dataset of inputs and corresponding outputs, where the inputs represent features or attributes of the data, and the outputs represent the target variable or the result that the model is intended to predict.

The quality and quantity of training data are critical factors in determining the accuracy and generalizability of the machine learning model. The training data should be representative of the problem that the model is intended to solve and should be large enough to capture the full range of variationin the data.

Before using training data, it is important to pre-process and clean it to remove any inconsistencies or outliers that could negatively impact the performance of the model. Once the training data has been pre-processed, it is split into a training set and a validation set, which is used to evaluate the performance of the model.

D. Testing Data

Testing data, also known as a test set or validation set, is separate set of data used to evaluate the performance of a machine learning model after it has been trained on a training dataset. The purpose of the testing data is to measure the model’s ability to generalize to new, unseen data.

The testing data is typically selected randomly from the available dataset, and it should be representative of the data distribution that the model is expected to encounter in real world scenarios. The testing data should also be kept completely separate from the training data so that the model does not learn any information from the testing data during the training process.

The testing data is used to evaluate the performance of the machine learning model by comparing its predicted outputs to the actual outputs for the test data. This process is often called model evaluation or testing. The evaluation metrics used to measure the performance of the model will depend on the specific problem and the type of machine learning algorithm used.

By evaluating the model on a testing dataset, we can estimate its ability to generalize to new, unseen data and determine if it has overfitted or underfit to the training dataset. If the model performs well on the testing data, it is likely to perform well in real-world scenario.

II. LITERATURE SURVEY

- The proposed system [1] uses RGB to grayscale conversion, thresholding, GLCM, and random forest classifier for the detection of tomato leaves diseases. The main objective of the system is to provide an accurate, fast, efficient, and inexpensive solution for the detection of tomato leaves.
- Random Forest Machine Learning Algorithm is used in this proposed model [2] to detect tomato leaf diseases accurately. The model identifies 7 tomato leaf diseases along with healthy leaves with high accuracy.
- The developed system in this paper [3] can forecast the attack of diseases on Mango fruit. It uses past weather data and crop production. Random forest algorithm is used for detecting the presence of diseases in mango fruitcrops.
- The proposed system [4] uses Convolutional Neural Network (CNN) for the detection of tomato leaves diseases. The model identifies the diseases with a high accuracy rate.
- Machine learning techniques are used by the proposed

DOI:

system [5] to detect crop diseases. It also suggests pesticides as a remedy to cure crop diseases.

- The paper [6] proposed a model which identifies rice leaf diseases using machine learning techniques. It identifies the most common 3 diseases of rice leaves namely leaf smut, bacterial leaf blight, and brown spot diseases.
- The proposed system [7] detects crop insects using different machine learning algorithms such as ANN, SVM, KNN, NB, and CNN.9-fold cross-validation is applied to improve the classification model performance.
- Random forest algorithm is used by the proposed system [8] to detect crop diseases. A Histogram of an Oriented Gradient(HOG) is used to extract the features of the image.
- The paper [9] proposes methods for the detection of rice plant diseases using machine learning algorithms and also recommends remedies to cure them.
- Convolutional Neural Network (CNN) is used in the proposed system [10] for the detection of plant diseases.
- The proposed system [11] detects papaya disease andalso compares some machine learning algorithms such as random forest, k-means clustering, SVM, and CNN for the highest accuracy rate.
- The paper [12] proposes a review of the advancement of machine learning technologies for plant disease detection. SVM and CNN algorithms are used in the proposed system.
- The proposed system [13] uses SVM and K-means clustering for the detection of plant diseases and a comparison is done. Image processing steps are done for the feature extraction.
- The paper [14] proposes a method to detect diseases of crops using a Deep Learning algorithm i.e. Convolutional Neural Network (CNN).
- The proposed system [15] uses a machine learning algorithm SVM to detect Betel vine disease. Bacterial Leaf Spots and Stem Leaf are detected using the proposed model.

III. METHODOLOGY

The main steps for the detection of leaf diseases are:

- Image Acquisition: In this section, pictures of plant leaves are gathered with the help of digital media, and cameras, with desired resolution and size. The images will even be taken from the net. The image information is responsible for the higher potency of the classifier at intervals in the last phase of the detection system.
- Image segmentation: This section aims at simplifying the illustration associate Image such that it becomes a lot of meaning and is easier to analyze. As a result of the premise of feature extraction, this phase is also the fundamental approach to the image process.
- Feature Extraction: After segmentation, the after-effect so far achieved is the realm of interest. Hence, throughout this step, the options from this space of interest have to be compelled to be extracted. These options are demanded to determine the means of a sample image. Options are usually based on color, shape, and texture.
- Classification: The classification section implies working out if the input image is healthy or pathologic. If the image is found to be diseased, some existing works have further classified it into a number of diseases.

The methodology implemented in the project is the feature extraction of the leaves using the random forest to accurately predict the presence of the disease in the tomato leaf. A publicly available dataset of diseased and healthy leaves is used at first. Random forest is used to classify healthy and diseased plant leaves. The classification of the healthy and diseased leaves of tomato plants is the main aim of the project. The system is trained with 4 classes of tomato leaf images obtained from the PlantVillage dataset. The same samplesare divided into training, valid, and testing sets. The project is divided into two phases. In the first phase, the feature extraction is done on all images and stored in a feature array. During the second phase, the Random Forest algorithm isused for training data using extracted features as input and the corresponding disease labels as output is obtained. Then the trained model is tested on the testing set to evaluate its performance. Once the model is trained and tested, it can be used to predict the disease type of new tomato leaf images. The preprocessed image is passed through the feature extraction pipeline to extract relevant features, and the Random Forest classifier is used to predict the disease type based on these features. If the model’s performance is not satisfactory, the methodology can be refined by using more advanced feature extraction techniques or exploring other machine learning algorithms.

IV. IMPLEMENTATION

The project is implemented in Python using Visual Studio Code. The dataset consists of healthy and diseased leaf images of tomato leaves. The dataset consists of 4 classes with 8788 images in total. The dataset classes are as below:

- Tomato yellow leaf curl virus
- Tomato Bacterial Spot
- Tomato Late Blight
- Tomato healthy

Each folder in the training set is named with the class names as above. The features from the leaves are extracted during the training phase and stored in the feature array, at the same time the folder names (labels) are stored in the Training labels array. The feature array and the training labels array are used for training the Machine Learning algorithms. The dataset is

DOI:

separated into 80 training set and 20 test set for evaluating the performance of the Machine learning algorithm.

V. RESULT AND DISCUSSION

The number of tomato leaf images used for train, testing, and valid are as given below:

- Tomato yellow leaf curl virus- 2399, 858, 1029 imagesfor train, test, and valid respectively.
- Tomato Bacterial Spot- 952, 341, 409 images for train,test, and valid respectively.
- Tomato Late Blight- 854, 306, 367 images for train, test,and valid respectively.
- Tomato healthy- 712, 255, 306 images for train, test, andvalid respectively.

The Machine Learning algorithm is trained with the dataset, once this is done the Machine Learning algorithm is tested by inputting a new image unseen during the training phase. The ML algorithm predicts the class of image and the same is displayed on the Python Console.

The features extracted during the training phase are storedin the training feature array and the labels are stored in the training labels array. The training features and labels are split into train data and test data, out of which 80 is the trainingdata and 20 is the test data. However, this ratio can be altered by changing the split ratio.

The test folder contains images unseen during the training phase. The dataset was trained using the Random Forest algorithm which resulted in high accuracy .The images along with the prediction are displayed in the Python Console.

VI. CONCLUSION

The proposed system identifies and classifies the tomato leaf diseases using the Random Forest algorithm. The proposed approach involves extracting features from images of tomato leaves, followed by training a Random Forest model to classify the leaves into healthy or diseased categories. The tomato diseases such as the Yellow leaf curl virus, Early Blight, and Bacterial Spot along with Healthy leaves are diagnosed by the system and predicts the disease of tomato leaves accurately. This approach can be used to improve the efficiency and effectiveness of tomato cultivation and can have a significant impact on the growth of tomatoes in the agricultural industry. The use of machine learning algorithms such as Random Forest can increase the efficiency of tomato cultivation.

References

[1] Sherly Puspha Annabel, V. Muthulakshmi, “AI-Powered Image-Based Tomato Leaf Disease Detection”, Third IEEE International Conference on I- L SMAC, 2019

[2] Meghana Govardhan, Veena M B, “Diagnosis of Tomato Plant Diseases using Random Forest”, Global Conference for Advancement in Technology (GCAT), 2019

[3] P. B. Jawade, Dattatray Chaugule, Devashri Patil, and Hemendra Shinde, “Disease Prediction of Mango Crop Using Machine Learning and IoT”, Published by Springer,2020

[4] Tahmina Tashrif Mim, Md. Helal Sheikh, Roksana Akter Shampa, Md.Shamim Reza and Md. Sanzidul Islam, “Leaves Diseases Detection of Tomato Using Image Processing”, IEEE Eighth International Conference on System Modeling Advancement in Research Trends, 2019

[5] Anuradha Badage, “Crop Disease Detection using Machine Learning: Indian Agriculture”, International Research Journal of Engineering and Technology (IRJET), 2018

[6] Kawcher Ahmed, Tasmia Rahman Shahidi, Syed Md. Irfanul Alam and Sifat Momen, “Rice Leaf Disease Detection Using Machine Learning Techniques”, International Conference on Sustainable Technologies for Industry, 2019

[7] Thenmozhi Kasinathan, Dakshayani Singaraju, Srinivasulu Reddy Uyyala, “Insect classification and detection in field crops using modern machine learning techniques”, Published by Sciencedirect, 2020

[8] S. Ramesh et al., ”Plant Disease Detection Using Machine Learning,” International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C), 2018

[9] Shital Patil, Rupali Saha, Aasha Sangole, ”Rice Plant Disease Detection and Remedies Recommendation Using Machine Learning”, International Research Journal of Engineering and Technology (IRJET), 2022

[10] Prof. Shailesh Kurzadkar, Achal Meshram, Aman Barve, Kajal Dhargave, Mruganayani Alone, Vijaya Bhongale, ”Plant Leaves Disease Detection System Using Machine Learning”, International Journal of Computer Science and Mobile Computing,2022

[11] Md. Ashiqul Islam, Md. Shahriar Islam, Md. Sagar Hossen, Minhaz Uddin Emon, ”Machine Learning based Image Classification of Papaya Disease Recognition”, IEEE Fourth International Conference on Electronics, Communication and Aerospace Technology,2020

[12] Hilman F. Pardede, Endang Suryawati, Dikdik Krisnandi, R. Sandra Yuwana, Vicky Zilvan, ”Machine Learning Based Plant Diseases Detection: A Review”, International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications, 2020

[13] Dr. Shaik Asif Hussain, Raza Hasan, Dr. Shaik Javeed Hussain, “Classification and Detection of Plant Disease using Feature Extraction Methods”, International Journal of Applied Engineering Research, 2018

[14] Ms. Nilam Bhise, Ms. Shreya Kathet, Mast. Sagar Jaiswar, Prof. Amarja Adgaonkar, Plant Disease Detection using Machine Learning”, International Research Journal of Engineering and Technology (IRJET), 2020

[15] Md Zahid Hasan, Nahid Zeba, Md. Abdul Malek and Sanjida Sultana Reya, “A Leaf Disease Classification Model in Betel Vine Using Machine Learning Techniques”, 2nd International Conference on Robotics, Electrical and Signal Processing Techniques, 2021