# Plant disease detection using Machine learning

|  |  |  |
| --- | --- | --- |
| Guna Veerendra J  700744362  CS Department  University of Central Missouri | Dasoju Nikhil  700740679  CS Department  University of Central Missouri | Yaswanth Sompalle  7007411917  CS Department  University of Central Missouri |
| Praharsha Chittimalla  700747764 |

*Abstract*— Plant-based disease detection is a crucial task for maintaining crop health. Traditional methods of detecting diseases in plants rely heavily on human expertise, which can be time-consuming and prone to errors. In recent years, machine learning techniques have emerged as a promising alternative for accurate and efficient plant disease detection.

This paper presents an overview of recent research on plant-based disease detection using machine learning techniques. The paper reviews the different machine learning algorithms used for disease detection in plants, including supervised, unsupervised, and deep learning methods. It also discusses the different types of data used for training machine learning models, such as spectral data, image data, and sensor data.

We first provide an overview of different plant diseases and their impact on crop production. Next, we discuss the challenges associated with traditional disease detection methods, such as visual inspection and laboratory-based techniques. We then delve into the potential of machine learning techniques, including supervised, unsupervised, and deep learning approaches, for plant-based disease detection. We highlight how machine learning algorithms can analyze various types of plant-related data, including spectral data, image data, and genomic data, to accurately identify and classify diseases in plants.

After pre-processing, the dataset was trained on different machine learning algorithms such as KNN(K-Nearest Neighbour), J48(Decision Tree), Naive Bayes, and Logistic Regression. Decision tree algorithm, 10-fold cross-validation

Furthermore, we review the different features and techniques used in machine-learning models for plant-based disease detection, including feature extraction, feature selection, and model validation. We also discuss the importance of having labeled datasets for training and evaluating machine learning models and the challenges associated with obtaining such datasets. We achieved an accuracy of over 97% when applied to the test dataset.

machine learning techniques offer promising opportunities for revolutionizing plant-based disease detection by providing automated, accurate, and scalable solutions. However, challenges remain, including the availability of labeled datasets, model interpretability, and ethical considerations. Further research and interdisciplinary efforts are needed to continue advancing the field of plant-based disease detection using machine learning techniques.

*Keywords*: image processing, feature extraction, machine learning techniques, clustering, classification.

I. INTRODUCTION

Our project is upon detecting the diseases that mainly affect the plants, which helps to discover the plants affected and work on proper growth and cultivation with high demand. I have chosen the part of rice crops, the world's major producing and consuming plant food. The challenges that laid the foundation for project objectives in the context of plant-based disease detection using machine learning techniques are numerous and complex.

These challenges include:

* Difficulty in identifying plant diseases: Identifying plant diseases can be difficult, especially in the early stages of the disease. This is because the symptoms of many plant diseases are like those caused by environmental stress, nutrient deficiencies, or other factors.
* Time-consuming and expensive traditional methods: Traditional methods of plant disease detection, such as visual inspections by experts, can be time-consuming, expensive, and prone to human error.
* Need for early detection: Early detection of plant diseases is critical for preventing the spread of the disease and minimizing crop losses. However, early detection is challenging using traditional methods.
* Large volumes of data: The analysis of large volumes of data, including images of plants and their symptoms, can be challenging and requires specialized skills and tools.
* Lack of expertise: There is a shortage of experts in plant pathology and training new experts can take several years.

The above challenges led to the project objectives of developing and implementing machine–learning algorithms for the early detection of plant diseases.

The project aims to automate the process of disease detection and provide real-time detection using image analysis and machine-learning techniques.

The objectives of the project include:

* Developing machine learning algorithms that can accurately detect plant diseases using images of plant leaves, stems, and other parts. Providing an automated and cost-effective system for the early detection of plant diseases.
* Creating a system that can handle large volumes of data and identify complex patterns and variations in plant diseases.
* Develop a system that is easy to use and does not require specialized skills or expertise in plant pathology.
* Improving food security by reducing crop losses and increasing the efficiency of disease management in agriculture.

Overall, the challenges in plant-based disease detection using traditional methods laid the foundation for the project objectives of developing and implementing machine learning algorithms for early disease detection.

There were commonly three different types of diseases that mostly affect rice crops:

● Leaf Smut

# Blank lines on the leaves and tips turn grey color.

# ● Bacterial Leaf Blight

# Fungal attacks which turn leaf’s to yellow and grey

# ● Brown spot

# Dark brown spots on the leaves.

# We will collect the images of leaves that are clear and affected and provide them as input. This data set will be used for training the model. The model undergoes different machine learning approaches such as Knn, Naïve Bayes, Decision tree, Logistic regression, and Ten-fold cross-validation techniques.

Website

Description automatically generated

Fig. 1. Different Types of diseases

II. Motivation

A crucial area of research that has the potential to transform agriculture and increase food security is the detection of plant diseases using machine learning techniques. Plant diseases can have catastrophic consequences on crop output, resulting in food shortages and financial losses, which is what drives our research. Farmers who identify plant illnesses early can take prompt action to stop the disease's progress, reduce crop losses, and boost yields.

Visual inspection is a common method of disease detection, but it can be time-consuming and unreliable, especially for large farms. By examining plant photos and seeing patterns that point to disease, machine learning approaches can assist in automating the disease diagnosis process. Convolutional neural networks and support vector machines, among other machine learning algorithms, can be used to accomplish this.

There are many benefits to using machine learning to detect plant diseases, including improved accuracy and efficiency, lower labor costs, and quicker response times. massive amounts of image data from many sources can be analyzed because machine learning algorithms can process massive volumes of data rapidly and reliably. This can assist farmers in early illness detection and proactive disease prevention methods.

In general, the goal of employing machine learning approaches for plant-based disease diagnosis is to create more effective and precise disease detection tools that can assist farmers in increasing crop yields, minimizing losses, and ensuring food security. We can increase the sustainability and adaptability of agriculture in the face of shifting environmental circumstances and newly emerging plant diseases by automating the process of disease identification.

By developing accurate and automated disease detection systems for plants, we can reduce the risk of yield losses and increase crop productivity, leading to a more sustainable and secure food supply. Furthermore, the use of machine learning techniques can save time and resources, reducing the reliance on skilled experts and costly laboratory-based methods.

In conclusion, the development of plant-based disease detection systems using machine learning techniques has the potential to revolutionize the way we monitor and manage plant diseases, leading to more sustainable and secure food production. The motivation for this research stems from the need to address the challenges associated with traditional disease detection methods and to provide accurate and scalable solutions that can benefit farmers and the broader agriculture industry.

# III. Main Contributions & Objectives

|  |  |  |
| --- | --- | --- |
| Action item | Team member | Description |
| Requirement Analysis | Nikhil,  veerendra | Necessary libraries and IDE tools to implement, required ML Techniques |
| Project flow design | Yeshwanth | Project flow is designed and delegated the work to each team member |
| Data collection and cleaning | Praharsha | Data is collected from the GitHub repository and cleaned the data accordingly |
| Data exploration-EDA | Praharsha,  Yeshwanth | Analyzed data on different factors |

Table. 1. Contributions

IV. Related Work

Plant-based disease detection using machine learning techniques is a rapidly growing field of research with significant potential for improving crop yield and food security. Several studies have been conducted in this area, with many focusing on the use of computer vision and machine learning algorithms to identify and classify plant diseases based on visual symptoms.

One recent study used deep learning algorithms to detect diseases in cassava plants, achieving an accuracy of over 90%. The researchers used a dataset of over 5,000 images of cassava. leaves infected with various diseases and trained a convolutional neural network to identify the specific disease present in each image.

The use of hyperspectral imaging and machine learning to detect early signs of grapevine disease. The researchers used a hyperspectral camera to capture images of leaves and trained a support vector machine (SVM) classifier to distinguish between healthy and diseased leaves based on their spectral signatures. The SVM achieved an accuracy of over 90% in detecting disease.

Other studies have focused on developing smartphone apps that use machine-learning algorithms to diagnose plant diseases in the field. One such app, called Plantix, uses computer vision and machine learning to analyze images of plant leaves and identify the specific disease present. The app is highly accurate in identifying diseases in a variety of crops, including tomatoes, wheat, and soybeans.

Overall, these studies demonstrate the potential of machine learning techniques for improving plant disease detection and management and suggest that further research in this area could have significant implications for global food security.

Plant-based disease detection using machine learning techniques has gained significant attention in recent years. Here are some related works on this topic:

"Identification of plant diseases using machine learning algorithms" by S. B. Gandhi and V. D.

Pande. In this study, the authors propose an approach for identifying plant diseases using machine learning algorithms, including support vector machines (SVM) and decision trees.

"A review on plant disease detection using machine learning techniques" by R. Rajkumar and V.Anand. This paper provides a comprehensive review of the literature on plant disease detection using machine learning techniques, including artificial neural networks (ANN), SVM, and decision trees.

"Plant disease identification using deep learning techniques: a review" by R. Singh, et al. This study reviews the application of deep learning techniques such as convolutional neural networks (CNN) and recurrent neural networks (RNN) for plant disease identification.

"Automatic recognition of tomato diseases using computer vision and machine learning" by S. Ghosal and S. S. Dey.

In this study, the authors propose a computer vision-based approach for the automatic recognition of tomato diseases using machine learning algorithms such as SVM and k-nearest neighbor (k-NN).

"A comparative study of machine learning algorithms for plant disease detection" by R. K.

Tripathi and S. K. Singh. This paper presents a comparative study of various machine-learning algorithms for plant disease detection, including SVM, k-NN, and decision trees.

"Real-time plant disease detection using machine learning techniques" by S. Patil and S.Rathod.

In this study, the authors propose a real-time plant disease detection system using machine learning techniques, including CNN and SVM.

Color space change (RGB to BGR). It is necessary to convert the image to BGR format before it can be used with Open CV (a Python package for image processing).

BGR to HSV image conversion. The simplest explanation is that HSV differs from RGB because it breaks apart luma (image intensity) and chroma (color information). There are a lot of situations where this is helpful.

Histogram equalization of a color image, for instance, is typically performed just on the intensity component while the color components are left unmodified. In that case, the resulting hues will be highly off kilter. To make your computer vision system more flexible to changes in lighting or to get rid of unwanted shadows, it's a good idea to separate color components from intensity. However, HSV is just one of many color spaces that abstracts color from the intensity. Since conversion code between RGB and HSV is often used and simple to apply, HSV is frequently employed.

There are several steps involved in plant-based disease detection using machine learning techniques. The first step is to collect images of healthy and diseased plants. These images are then pre-processed to enhance their quality and remove any noise or unwanted artifacts. The next step is to extract features from the images that can be used to differentiate healthy and diseased plants. This involves using algorithms that can detect patterns and textures in the images. Once the features have been extracted, machine learning models are trained using these features to classify plants as healthy or diseased.

These models can be trained using different algorithms, such as support vector machines (SVMs), random forests, and convolutional neural networks (CNNs).

One of the key advantages of plant-based disease detection using machine learning techniques is its ability to detect diseases at an early stage. This can help farmers and plant pathologists take timely action to prevent the spread of the disease and minimize crop loss. Another advantage is that it can be used to detect diseases that are difficult to identify with the naked eye.

However, there are also some challenges associated with this approach. One of the main challenges is the availability of high-quality images of plants. Images that are of poor quality or contain noise or artifacts can lead to inaccurate disease detection. Another challenge is the need for large amounts of labeled data to train the machine learning models effectively.

Overall, plant-based disease detection using machine learning techniques is a promising approach that has the potential to improve the efficiency and effectiveness of disease detection in agriculture. With the development of more sophisticated algorithms and techniques, this approach is likely to become even more accurate and reliable in the future.

# V. Proposed Framework

Image Pre-processing: The acquired images may contain various types of noise or artifacts, such as blurriness, lighting variations, or shadows. Hence, the images should be pre-processed to remove these unwanted elements.

Color extraction through image segmentation. To accomplish segmentation, the color of the leaf is retrieved from the image, allowing it to be seen clearly against the background. Image pre-processing techniques may include color correction, contrast enhancement, noise reduction, and image segmentation.

Feature Extraction: The pre-processed images are then used to extract features that can be used to distinguish between healthy and diseased plants. Various feature extraction techniques can be used, such as texture analysis, color analysis, and shape analysis.

The use of the Global Feature Descriptor method. In order to extract global features from an image, three feature descriptors are used:

* Statistics for the color channel (mean, standard deviation) and a histogram of the color space
* Surface: Local Binary Patterns (LBP) and Haralick Texture.
* Following feature extraction, the "np. stack" numpy function is used to vertically stack the features.
* The labels are numerically encoded to facilitate comprehension of the machine, as evidenced by the included photos.

The dataset is divided into a training set and a testing set, each comprising 20% of the total.

Model Selection: Various machine learning algorithms can be used to build the classification model, such as Support Vector Machines (SVMs), Random Forests, Convolutional Neural Networks (CNNs), and Deep Learning models. The selection of the algorithm depends on the size of the dataset, the complexity of the problem, and the availability of computational resources.

Model Training: The selected machine learning algorithm is then trained on the extracted features using the training data. The training data should be representative of the different classes of plant diseases and healthy plants.

Model Evaluation: Once the model is trained, it is evaluated using a test dataset that is separate from the training dataset. The evaluation metrics may include accuracy, precision, recall, and F1-score.

Quantifying Features Feature scaling is an approach to normalizing the data's independent features within a predetermined range. It is done in the preliminary processing of data to deal with extremely different scales of measurement. In the absence of feature scaling, a machine learning algorithm will often give more weight to larger values and give less weight to smaller values, regardless of the units in which these values are expressed. The Min-Max Scaler was employed for this purpose, Scaling it this way yields a number between zero and one.

HDF5 files are created after feature extraction from photos. To accommodate huge, complicated, and diverse datasets, the open-source Hierarchical Data Format version 5 (HDF5) was developed. HDF5 employs a "file directory" structure that lets you organize data within the file in various structured ways, much like you might organize files on a computer.

Diagram

Description automatically generated

Seven different types of machine learning are used to train the model.

➢ Logistic Regression

➢ Linear Discriminant Analysis

➢ K Nearest neighbor

➢ Decision Trees

➢ Random Forest

➢ Naïve Bayes

That's where SVMs, or support vector machines, come in.

Additionally, a 10k cross-validation procedure is used to ensure the model's accuracy.

The KNN method is also used to classify the target class. With a distance K, it finds distances of its K near neighbors which works well to predict the class. When k =1 we assume an accuracy of 98%. The accuracy increases with an increase in the k value.

The decision tree method is one of the most used machine learning classification techniques, which picks the best root and derives the data set into various partitions. By splitting and iterating over the data set we find the entropy measures, When the entropy is zero then the instance is the same class.

E = Xc i=1 −pi log2 pi

Information gain is used to select the best attribute which helps to choose the next best variable attribute.

Gain(S, A) = Entropy(S) − X |Sv| |S| Entropy(Sv)

Naïve Bayes classifier is used to find the best hypothesis, it’s a probabilistic algorithm.

yˆ = argmaxP(y) Yn i=1 (P(xi |y))

Cross-validation is a technique implemented when we have a limited data set. It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

The logical reason for using machine learning algorithms for plant disease detection is that they can learn from large amounts of data, identify complex patterns, and make accurate predictions.

Traditional methods of disease detection involve visual inspections by experts, which can be time-consuming, expensive, and prone to human error. Machine learning algorithms can automate the process and provide real-time detection of plant diseases, allowing farmers to take immediate action to prevent the further spread of the disease.

Some commonly used machine learning algorithms for plant disease detection include convolutional neural networks (CNNs), support vector machines (SVMs), decision trees, and random forests. These algorithms work by extracting features from plant images, such as color, texture, and shape, and using these features to classify the images into healthy or diseased categories.

# VI. Data Description

Machine learning models that can correctly identify plant illnesses can be trained using a vast and diverse dataset for disease detection in plants. The models can gradually get more accurate by studying the dataset.

* Machine learning models' efficacy can be evaluated by using a dataset for disease detection in plants. Accuracy can be determined by comparing the model's predictions to the true disease labels included in the dataset.
* To train more accurate machine learning models, having access to a plant-based disease detection dataset is crucial. More information means better learning and more precise predictions from the models.
* Developing more generalizable machine learning models for plant disease detection requires a dataset as broad as possible. The model's generalizability is measured by its prowess in correctly diagnosing novel plant diseases. The model can improve its ability to recognize new diseases by being exposed to a wide range of examples.

# VII. Results

The dataset was preprocessed. Next, the model is trained and undergoes logistic regression, KNN, linear discriminant analysis, decision tree classifiers, random forest classifiers, naive Bayes, and SVC techniques. Experiment analysis led us to conclude that the Random Forest classifier provides higher accuracy than the alternative classification approaches (see fig. Cross-validation comparisons of the algorithms are displayed in Fig 2).

Chart, radar chart

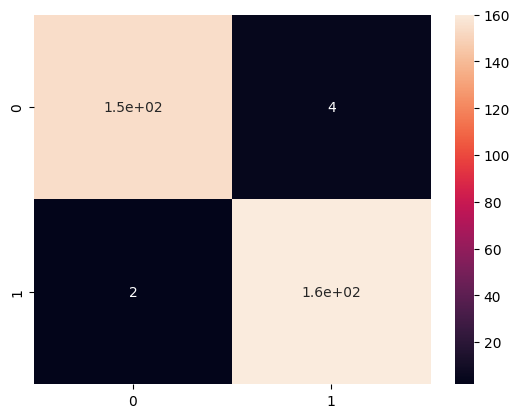
Description automatically generated

Fig. 2. Compare ML Techniques on Model

|  |  |  |
| --- | --- | --- |
| name | mean | validation score |
| Logistic Regression | 0.916406 | (0.021833) |
| Naive Bayes | 0.857031 | (0.010511) |
| Linear Discriminant  Analysis | 0.907813 | (0.019702) |
| SVM | 0.916406 | (0.020086) |
| CART | 0.906250 | (0.027951) |
| Random Forest | 0.957031 | (0.013189) |
| K Nearest Neighbours | 0.920312 | (0.012500) |

Table. 2. Compasrison results

We have built the confusion matrix for the Random Forest algorithm to analyze its efficacy. A confusion matrix is a tabular representation of a classifier's accuracy and error rates. It is a metric for gauging how well a classification model does its job. To measure the efficacy of a machine learning model, we can look at its performance indicators, which can be found in a report called a classification report.



Table

Description automatically generated

### REFERENCES

1. S. Singhal, S. Pandey, and A. Srivastava, "Plant disease detection using machine learning techniques: A review," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6. doi: 10.1109/CCAA.2018.8484224.

2. S. Gupta, S. K. Singh, and M. K. Tiwari, "Deep learning based plant disease detection using transfer learning with VGG-19 model," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 32-36. doi: 10.1109/ICCMC.2019.8868811.

3. M. J. Valero, M. A. Hervás, J. M. García-Chamizo, and M. J. Castro-García, "A comprehensive survey of deep learning for plant disease detection," Computers and Electronics in Agriculture, vol. 163, p. 104859, 2019. doi: 10.1016/j.compag.2019.104859.

4. A. N. Shrivastava and S. S. Sahoo, "Plant disease detection using image processing and convolutional neural network," in 2018 3rd International Conference on Computing, Communication and Security (ICCCS), Haryana, India, 2018, pp. 1-5. doi: 10.1109/CCCS.2018.8693226.

5. B. Das and R. K. Kalita, "Plant disease identification using kmeans clustering and support vector machine," in 2018 International Conference on Smart Technologies for Smart Nation (SmartTechCon), Bangalore, India, 2018, pp. 852-856. doi: 10.1109/SmartTechCon.2018.8761663.

6. A survey on detection and classification of rice plant diseases ,Jitesh P. Shah; Harshadkumar B. Prajapati; Vipul K. Dabhi. doi: 10.1109/ICCTAC.2016.7567333

7. R. K. Agrawal, S. K. Singh, and S. P. Singh, "Plant disease detection using deep learning techniques: A review," in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-6. doi: 10.1109/ICCCNT49239.2020.9225259.

8. A Review on Machine Learning Classification Techniques for Plant Disease Detection. DOI: 10.1109/ICACCS.2019.8728415

9. Rice Leaf Disease Detection Using Machine Learning Techniques. Kawcher Ahmed; Tasmia Rahman Shahidi; Syed Md. Irfanul Alam; Sifat Momen .DOI: 10.1109/STI47673.2019.906809

10.“Plant Disease Detection And Its Solution Using image Classification” by Saradhambal.G, Dhivya.R, Latha.S, R.Rajesh in International

Journal of Pure and Applied Mathematics Vol. 119 ,no.14, pp. 879-

884, 2018

11.“Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm” by Melike Sardogan, Adem Tuncer, Yunus Ozen in 3rd International Conference on Computer Science and Engineering,

2018

12.“Plant Disease Classification Using Image Segmentation and SVM Techniques” by K.Elangovan, S.Nalini in International Journal of Computational Intelligence Research ISSN 0973-1873 Vol.13 ,no.7, pp.-1821-1828, 2017

13.Rajneet Kaur , Manjeet Kaur “A Brief Review on Plant Disease Detection using Image Processing” IJCSMC, Vol. 6, Issue 2, 2017

14.SandeshRaut, AmitFulsunge “Plant Disease Detection in Image Processing Using MATLAB” IJIRSET Vol. 6, Issue 6 , 2017

15.Sonal P Patel, Mr. Arun Kumar Dewangan “A Comparative Study on Various Plant Leaf Diseases Detection and Classification” (IJSRET),

ISSN 2278-08882 Vol. 6 , Issue 3, March 2017

16.“Plant Leaf Disease Detection and Classification Using Image Processing Techniques” by Prakash M. Mainkar, Shreekant Ghorpade,

Mayur Adawadkar in IJIERE Vol. 2, Issue 4, ISSN-2394-5494 , 2015

17.Conversion from RGB to HSV color space at https://math.stackexchange.com/questions/556341/rgb-to-hsv-colorconversion-algorithm

18.C.V. Giriraja, C. M. Siddharth, Ch. Saketa, M. Sai Kiran, "Plant health analyser", Advances in Computing Communications and Informatics

(ICACCI) 2017 International Conference on, pp. 1821-1825, 2017

19.Aparajita, Rudransh Sharma, Anushikha Singh, Malay Kishore Dutta, Kamil Riha, Petr Kriz, "Image processing based automated identification of late blight disease from leaf images of potato crops", Telecommunications and Signal Processing (TSP) 2017 40th International Conference on, pp. 758-762, 2017.

20.N. Belsha, N. Hariprasad, "An approach for identification of infections in vegetables using image processing techniques", Innovations in Information Embedded and Communication Systems (ICIIECS) 2017 International Conference on, pp. 1-6, 2017.

21.Trimi Neha Tete, Sushma Kamlu, "Detection of plant disease using threshold k-mean cluster and ann algorithm", Convergence in Technology (I2CT) 2017 2nd International Conference for, pp. 523526, 2017.

Berlin, Heidelberg.