Detecting Grape Leaf Disease and Classification using Machine Learning

[*]https://github.com/mahathi444/ML-Final-Project-
https://github.com/pavanir2530/Final-project/upload/main
https://github.com/vamsi-mekala/ML_FINAL_PROJECT

https://github.com/jayanthranga/ML-FINAL-PROJECT

Vamsi Krishna Mekala Pavani Katkuri Mahathi Reddy Panyala Ranga Jayanth Kumar

Abstract—**Agriculture provides food to all human beings even in case of the rapid increase in the population and agriculture yield is a key factor in economic growth. Leaf diseases are important factors determining the yield and quality of plants. This is just one reason that plant disease diagnosis is crucial in the sector of agriculture, as the presence of illness in plants is extremely common. If necessary precautions are not followed in this region, plants suffer major consequences, which have a consequence on the quality, volume, or efficiency of the corresponding products. However different leaves of a single plant can be attacked by different types of diseases and hence it would become a difficult process for farmers to detect each disease manually and take action corresponding to that disease. The use of an automatic method for disease identification is advantageous because it lessens the amount of labor required to manage large crop fields and can identify disease symptoms at their earliest stage when they first emerge on leaf tissue. It examines the growth and overall health of the plant and ensures that the agricultural planting will operate as usual and provide a successful crop. This paper combines Computer Vision ad Machine Learning techniques to detect and classify the diseases of grape leaves. The Support Vector Machine, Logistic Regression, Random Forest Classifier, Decision Trees, and Naive Bayes Classifier machine learning algorithms have been evaluated in classifying the diseases of grape leaves. The proposed method uses image segmentation techniques to identify the leaf in the image and then classify the segmented leaves as healthy, black-rot, esca, or leaf blight. This paper used the Global Thresholding technique to separate leaves from the image. On comparing the results, the Support Vector Machine model gave the better performance out of all models developed.**

Index Terms—**Leaf Diseases, Computer Vision, Machine Learning, Classification, Global Thresholding**

I. Introduction

It is a wise endeavour to modify a portion of the Earth's crust through the cultivation of vegetation and other growing crops as well as the breeding of livestock for nourishment or other needs for human beings and economic gain. This activity is commonly referred to as agriculture or farming. It is a turning point and a catalyst for the early social evolution of humans. Agriculture has been developed thanks to the significant technological and scientific advances made possible by the wealth of human society. It is crucial to the economy and is seen as the foundation of the economic system in developing nations. Agriculture has long been linked to the production of essential food crops. Farming in the present day includes dairy, fruit, forestry, poultry, beekeeping, and other arbitrary things. A significant portion of the population is also given employment opportunities by it.

In addition to producing food for people and animals alike, agricultural systems must also include environmental protection considerations. Due to the potential increase in the cost and environmental effects, there is currently growing demand to limit the usage of pesticides. Crop monitoring makes it possible to spot potentially dangerous locations and treat each one separately, greatly improving the effectiveness of the management of the disease.

Plant diseases are quite expensive for farmers to deal with on an agricultural farm. According to IRJET study, crop losses caused by weeds, animals, infections, pests, and other factors contribute for 20 to 40% of the total production of the world's agriculture. It is not always effective to physically examine specific characteristics of leaf surface, such as texture, colour, and form, in order to detect illnesses. In order to diagnose problems in their plants on large farms, the majority of farmers all over the world hire professional agriculturists. However, the process is expensive and time-consuming. Modern methods for automating the detection and classification of plant diseases are absent from certain farmers' conventional practises. Large farms see a considerable decrease in the amount and efficiency of agricultural production as a result of farmers' failure to detect plant diseases. The ability to continuously monitor plant disease without a lot of labor-intensive work, particularly in remote farm locations, is made possible by smart agriculture, which is an essential digital asset for farmers.

Computer vision is a technology that teaches how to save and change images and videos as well as how to extract information from them. Real-time applications increasingly heavily rely on the OpenCV toolkit for computer vision, imaging analysis, and machine learning. Technologies for image processing are used to identify the most obvious aspects and edit photographs in order to improve them.

Object recognition and categorization are two common functions of machine learning algorithms. The process of recognizing, comprehending, and classifying things and concepts into predetermined groups, often known as "sub-populations," is known as classification. Machine learning programs use a range of techniques to classify upcoming datasets into appropriate and pertinent categories with the aid of these which was before training datasets. Machine learning classifiers use input training data to assess the chance or likelihood that the data that comes after will fit into one of the established categories. Classification is essentially a type of ”pattern recognition.” The same pattern is discovered in subsequent data sets by applying classification algorithms to the training data in this case. Machine learning relies heavily on classification since it teaches computers how to classify data according to specific criteria, such as specified traits. With the increased use of big data for decision-making across industries, classifications within machine learning are becoming a crucial technique. Researchers and data scientists can better understand data and identify trends with the aid of classification. Making more precise, data-driven decisions is made possible by using these data patterns.

Image segmentation is the division of a digital image into several image segments, often referred to as visual features or image objects, in the context of modern visual processing and computer vision. The intention of segmentation would be to make an image representation more understandable and straightforward to examine. Image segmentation is frequently used to identify boundaries and objects in photographs. Image segmentation, in more exact terms, is the method of assigning each pixel of an image a label so that pixels that share the same name have specific properties. Foreground and background can be distinguished in an image by segmenting it, or pixels can be grouped together based on their similarity in color or shape. There are different image segmentation techniques such as Threshold Based Segmentation, Edge Based Segmentation, Region-Based Segmentation, Clustering Based Segmentation, and Artificial Neural Network Based Segmentation among which our work used the Global Thresholding method which is Threshold based image segmentation technique.

A straightforward type of picture segmentation is image thresholding segmentation. It is a technique for converting an original image into a binary or cross image by applying a threshold value to the pixel intensity. We shall take into account the intensity distribution of each pixel in the image during thresholding. After that, we will choose a threshold to segment the image. We can segment an image into sections based on the brightness of the item and background in an image containing an object and backdrop. But to separate a picture into an entity and a backdrop, this threshold must be precisely calibrated.

II. Motivation

Due to the constantly changing climatic and environmental factors, illnesses are relatively common in crops. Crop diseases are typically difficult to control and have an impact on the growth and output of crops. The ability to accurately diagnose diseases and take timely preventative measures are essential for ensuring high production standards and a high standard of quality. The widely cultivated grape plant in India is susceptible to various diseases that can harm the fruit, stem, and leaves. Leaf diseases are the early signs of fungus, bacterium, and virus infection. The cost of treating plant diseases on agricultural property is relatively high. Physically examining specific leaf surface qualities, such as surface, color, and form can sometimes be ineffective for spotting diseases. In order to stop yield losses, it’s crucial to identify plant diseases. Manually observing plant diseases is really challenging. In addition to requiring a large amount of labor, it also necessitates knowledge of plant diseases and a lengthy period of time. In light of this, plant disease diagnosis can be carried out using image processing and machine learning models. With the use of images of the leaves, we have described a method for identifying plant illnesses in this research. A subset of signal processing called ”image processing” is able to extract from an image the picture’s attributes or other relevant data. A component of artificial intelligence called ”machine learning” operates automatically or provides guidance for carrying out certain tasks. This motivated us to evaluate the performance of various machine-learning classification algorithms to detect and classify the diseases of grape leaves.

III. Objectives

This paper is focused on evaluating different Machine Learning algorithms in classifying the diseases of grape leaves by segregating the leaves from the input image. This paper has three main objectives:

- Applying the Global Thresholding technique to segregate the grape leaves in an image from its background.
- To Extract the features from the grape leaves.
- To classify the disease of the grape leaf based on extracted features.

This paper explores the following Algorithms and analyses their performance on the dataset mentioned in section IV.

- Logistic Regression
- Support Vector Machine
- Random Forest
- Decision Tree
- Naive Bayes Classifier

IV. Related Work

Abdullah et al.[1] proposed an automated system for classifying rubber tree leaf diseases that makes use of the primary RGB colour model. The PCA technique was used to reduce the input dimension and the ANN was used to classify the three rubber leaf diseases.

H. Al-Hiary et al.[2] For the clustering and classification of diseases that impact plant leaves, developed applications of K-means clustering and neural networks. Five plant diseases—early scorch, cottony mold, ashen mold, late scorch, and small whiteness—were used to test their algorithm. Their experimental findings demonstrated the value of their suggested approach, which can considerably aid in the precise diagnosis of leaf diseases with minimal computational work. They applied image acquisition, processing, segmentation, feature extraction, statistical analysis, and classification techniques.

Shriroop C. Madiwalar et al.[3] suggested using color photos of mango leaves to identify plant diseases using machine vision. The diagnosis of two illnesses, anthracnose and leaf spot, was made on the leaves after disease identification. For the feature extraction process, the authors used three feature groups: the GLCM feature, the Color-based feature, and the Gabor filter feature. They came to the conclusion that the Gabor filter's boundary extraction effectively detected the smaller spots in the case of leaf spot, but the other two techniques are in charge of the denser texture analysis as in the case of anthracnose. The three feature sets taken combined produced the best results, but at the cost of greater computing complexity.

H. Sabrol et al.[4] categorised photographs of healthy tomato plant leaf and stem, Septoria spot, bacterial spot, bacterial canker, tomato leaf curl, and tomato late blight. By separating colour, shape, and texture information from photos of healthy and ill tomato plants, the authors performed categorization. Six different types of tomato photos were categorised, and the overall classification accuracy was 97.3

V. Proposed methodology

A. Image Processing

The images are taken from the internet and come in a variety of sizes and sources. The photos also have noise because of poor lighting, weather occlusion, etc. To

simplify the computational process The images have been reduced in size to a typical width and height. The noise in this scaled image is then filtered using a Gaussian filter. We have utilised a 5*5 kernel size to filter the noise, which is a low pass filter that lowers the high frequency components of the signal.

B. Image Segmentation

The leaf portion of the preprocessed image is separated from the background image using the Grabcut segmentation technique. This approach uses the Gaussian Mixture Model (GMM) to classify pixels as foreground or background and also uses an initial rectangle to roughly separate the two. As the bounding box, we chose a rectangle with the dimensions (10, 10, w-30, and h-20), where w and h are the image's width and height. Figure shows the outcomes of the Grabcut technique. 3. The sick sections are removed from the foreground, or the leaf part, that was extracted. Lesions, coloured patches, and some yellowing leaf tissue make up the diseased portion. We have two distinct procedures for removing the infected area from the leaves.

1) Diseased Part Identification- Global Thresholding: This technique turns an RGB image into a greyscale image, which is subsequently transformed into a binary image via global thresholding. Connected component labelling is used to locate the contours on the thresholded image. Then, morphological techniques like dilation and erosion are used on the contour with the largest area. The original image is transformed into an HSV image, and thresholding is applied to the h channel. The contour detected picture and the HSV image are then both subjected to the binary AND operator.

2) Diseased Part Identification- Semisupervised Learning: In the BGR image, the damaged leaf tissue typically has a blue tint. By turning the RGB image into a BGR image, blue colour pixels are filtered off to segment the sick portion. We have utilised the training image to identify the lower and upper boundary of blue colour pixels in order to filter those pixels. The lower and upper border pixels are then filtered out of the input image as blue pixels. Thresholding is used on the filtered image to identify the sick areas.

3) Feature Extraction: Rich information about the substance of the image is provided by image features. These traits serve as some distinguishing qualities that can be utilised to distinguish between the various categories of input patterns. In this study, we classified the photos using the texture and colour characteristics of the photographs. Following the segmentation of the leaf's diseased area, colour features were looked at to determine if the provided input leaf was healthy or not. The segmented image won't have any regions and will only have black pixels if the input leaf image is one of a healthy leaf. While the feature vector for unhealthy leaves contains some colour information, it simply contains 0s for healthy leaves.

4) Classification using Different Classifiers:

- **Logistic Regression**

  By measuring each independent variable's distinct contribution, logistic regression is a quick and effective technique to examine the impact of a group of independent variables on a binary outcome. Logistic regression iteratively determines the strongest linear combination of variables with the highest likelihood of identifying the observed outcome using elements of linear regression indicated in the logit scale[12]. The "equation(1)" for logistic regression is as follows:

  $$ProbabilityOfOutcome(\hat{Y}_i) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i}} \quad (1)$$

  - Here Ŷi represents the estimated probability of being in one binary outcome category (i) versus the other
  - Here $e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_i X_i}$ represents the linear regression equation for independent variables expressed in the logit scale

  The logit scale transforms the original equation of linear regression to obtain natural log of the odds of being in one outcome category (Ŷ) versus the other category (1 –Ŷ) is given by "(2)"

  $$ln(\hat{Y}/1\text{–}\hat{Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \quad (2)$$

  Despite its name, logistic regression is more of a classification model than a regression model. For situations involving binary and linear classification, logistic regression is a straightforward and more effective approach. It's a classification model that's incredibly simple to implement and performs admirably with linearly separable classes.

- **Support Vector Machine**
  Support vector machines are part of the category of supervised machine learning methods, which are typically applied to classification tasks. The Support Vector Machine's fundamental principle is to take data with relatively low dimensions and transform it into data with greater dimensions. The data is then divided using a hyperplane to classify it into various groups. By tolerating misclassifications, a support vector classifier is utilised to determine the appropriate decision boundary. The choice of the kernel function and its parameters is critical when using SVM to tackle real-world problems. By choosing the right kernel function and parameters, one may create an SVM classifier with strong generalisation capabilities. The most often used kernel function is the RBF. Finding the best hyper-parameter for an SVM model is a highly challenging issue. These hyper-parameters include things like what C or gamma values to utilise. But it can be discovered by simply attempting all possible combinations and observing which inputs are most effective.
- **Random Forest**
  Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead than depending on a single decision tree, the random forest uses estimates from each tree and predicts the result based on the votes of the majority of predictions. Higher accuracy and overfitting are prevented by the larger number of trees in the forest. The purpose of a criterion is to assess a split's quality. Here we considered both gini and entropy as criterion to obtain best outcome. The gini impurity counts the number of times a dataset piece will be incorrectly identified when it is randomly labelled. The Gini Index has a minimum value of 0. When a node is pure, which occurs when every element it contains belongs to a single, distinct class, this occurs. This node won't be split once more as a result. Therefore, the features with a lower Gini Index choose the best split. Additionally, it is at its highest value when the probabilities for the two classes are equal. Entropy is a unit of measurement for information that depicts the disorder of the target's features. The feature with the lowest entropy selects the optimal split, just like the Gini Index does. When the probability of the two classes is equal, it reaches its highest value, and a node is pure when the entropy is at its lowest point, which is zero.
- **Decision Tree**
  A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. While Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions. It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree.
  The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to construct a tree. A decision tree only poses a question and divides the tree into subtrees according to the response (Yes/No). The fundamental problem that emerges while developing a decision tree is how to choose the best attribute for the root node and for sub-nodes. So, a method known as attribute selection measure, or ASM, can be used to tackle these issues. By using this measurement, we can choose the idea attribute for the tree nodes with ease. There are two widely used ASM approaches, which are as follows: Information Gain/Entropy Gini Index The gini impurity counts the number of times a dataset piece will be incorrectly identified when it is randomly labelled. The Gini Index has a minimum value of 0. When a node is pure, which occurs when every element it contains belongs to a single, distinct class, this occurs. This node won't be split once more as a result. Therefore, the features with a lower Gini Index choose the best split. Additionally, it is at its highest value when the probabilities for the two classes are equal. Entropy is a unit of measurement for information that depicts the disorder of the target's features. The feature with the lowest entropy selects the optimal split, just like the Gini Index does. When the probability of the two classes is equal, it reaches its highest value, and a node is pure when the entropy is at its lowest point, which is zero.
  The DecisionTreeClassifier() is present in sklearn.tree library. It contains mainly two parameters. They are criterion and random state. Criterion is used to measure the quality of split, which can be estimated by information gain or gini index . Random state is used to generate the random states.
- **Ada Boost Classifier**
  A common method for solving binary classification issues is boosting. By transforming a number of weak learners into strong learners, these methods increase prediction ability.
  The basic idea behind boosting methods is that after creating a model using the training dataset, we create a second model to fix any mistakes in the original one. This process

is repeated until the mistakes are reduced and the dataset can be accurately forecasted. To get the final output, the boosting process combines several models. AdaBoost, also known as Adaptive Boosting, is a machine learning method used as an ensemble framework. Decision trees with one level, or Decision trees with only one split, are the most popular algorithm used with AdaBoost. Another name for these trees is Decision Stumps.
Adaboost must adhere to two requirements:

– On a variety of weighed training instances, the classifier should be trained interactively.
– It strives to minimise training error in order to offer the best match possible for these instances in each iteration.

This algorithm creates a model while assigning each data point an equal weight. Then, it gives points that were incorrectly categorised larger weights. The next model now gives more weight to all the points with higher weights. If no lower error is received, it will continue to train the models. The AdaBoost Classifier is present in sklearn library. It uses Decision Tree Classifier as default Classifier. The base_estimator, n_estimators, and learning_rate are the crucial parameters. The base_estimator is a weak learner with which the model was trained. For training purposes, DecisionTreeClassifier is used by default as a weak learner. Additionally, we can select other machine learning algorithms. The n_estimators tells the number of weak learners to train iteratively. The learning_rate is what affects how weak learners are weighted. As a default state, it uses 1.

- **Naive Bayes Classifier**

A probabilistic machine learning model called the Naive Bayes classifier is utilised for classification tasks. The Bayes theorem is the cornerstone of the classifier.
The formula for Bayes' theorem is given as: "(3)"

$$P(A/B) = P(B/A)P(A)/P(B) \tag{3}$$

Where,
P(A/B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B/A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.
When B has already happened, we may use the Bayes theorem to calculate the likelihood that A will also occur. Here, A is the hypothesis and B is the supporting evidence. Here, it is assumed that the predictors and features are independent. That is, the presence of one feature does not change the behaviour of another.

VI. Results and Discussion

5675 grape leaves that were downloaded from the Plant Village website and the internet were used to evaluate the suggested system. Eighty percent of the photos have been used for testing and training. The global thresholding strategy, which was utilised to segment the specific sick section of the leaves and improve classification outcomes, was shown to be more appropriate for training the model. The training accuracy results from various machine learning methods are compiled. The results clearly show that when the features are retrieved from the diseased area of the images using global thresholding and trained using an SVM classifier with tuned parameters—SVM performs well when the data is very non-linear—good training accuracy is attained. For 1135 test photos, an overall accuracy of 93.035% was achieved.

Conclusion

In this paper, we suggest a machine learning-based automatic leaf recognition system that can detect illnesses in grape leaves. The suggested method uses the grab cut segmentation technique to first separate the leaf component from the background. Two distinct techniques are used to identify the sick region in the segmented leaves. Global thresholding is used in the first method, whereas semisupervised learning is used in the second. Texture and colour features are retrieved from the diagnosed diseased portion, trained using several classifiers, and the results are compared. For classification, we have SVM, Random Forest, and Adaboost algorithms. By utilising global thresholding and SVM, we were able to reach a superior result of 93.035% for testing accuracy.

References

[1] Abdullah, N. E., Rahim, A. A., Hashim, H., Kamal, M. M. (2007, December). Classification of rubber tree leaf diseases using multilayer perceptron neural network. In 2007 5th student conference on research and development (pp. 1-6). IEEE.

[2] Al-Hiary, H., Bani-Ahmad, S., Reyalat, M., Braik, M., Alrahamneh, Z. (2011). Fast and accurate detection and classification of plant diseases. International Journal of Computer Applications, 17(1), 31-38.

[3] Madiwalar, S. C., Wyawahare, M. V. (2017, February). Plant disease identification: a comparative study. In 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) (pp. 13-18). IEEE.

[4] Sabrol, H., Satish, K. (2016, April). Tomato plant disease classification in digital images using classification tree. In 2016 International Conference on Communication and Signal Processing (ICCSP) (pp. 1242-1246). IEEE.

[5] Bhange, M., Hingoliwala, H. A. (2015). Smart farming: Pomegranate disease detection using image processing. Procedia computer science, 58, 280-288.

[6] Mokhtar, U., Ali, M. A., Hassenian, A. E., Hefny, H. (2015, December). Tomato leaves diseases detection approach based on support vector machines. In 2015 11th International computer engineering conference (ICENCO) (pp. 246-250). IEEE.

[7] Owomugisha, G., Mwebaze, E. (2016, December). Machine learning for plant disease incidence and severity measurements from leaf images. In 2016 15th IEEE international conference on machine learning and applications (ICMLA) (pp. 158-163). IEEE.

[8] Aasha Nandhini, S., Hemalatha, R., Radha, S., Indumathi, K. (2018). Web enabled plant disease detection system for agricultural applications using WMSN. Wireless Personal Communications, 102(2), 725-740.

[9] Kakade, N. R., Ahire, D. D. (2015). Real time grape leaf disease detection. Int J Adv Res Innov Ideas Educ (IJARIIE), 1(04), 1.

[10] Kaur, P., Singla, S., Singh, S. (2017). Detection and classification of leaf diseases using integrated approach of support vector machine and particle swarm optimization. International Journal of Advanced and Applied Sciences, 4(8), 79-83..

[11] Sannakki, S. S., Rajpurohit, V. S., Nargund, V. B., Kulkarni, P. (2013, July). Diagnosis and classification of grape leaf diseases using neural networks. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[12] 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.

[13] Panchal, S. S., Sonar, R. (2016). Pomegranate leaf disease detection using support vector machine. International Journal of engineering and computer science, 5(6), 16815-16818.

[14] An, T. K., Kim, M. H. (2010, October). A new diverse AdaBoost classifier. In 2010 International conference on artificial intelligence and computational intelligence (Vol. 1, pp. 359-363). IEEE.

[15] Safavian, S. R., Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[16] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., Klein, M. (2002). Logistic regression (p. 536). New York: Springer-Verlag.