

**Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning**

July - November 2024

**Submitted By,**

Harshini R,

B.Tech, Computer Science and Business Systems

(125018029)

**Submitted To,**

Swetha Varadarajan

# **TABLE OF CONTENTS**

| **S.No** | **Title** | **Page No.** |
| --- | --- | --- |
| 1. | Abstract | 3 |
| 2. | Introduction | 3 |
| 3. | About the Dataset | 4 |
| 4. | Background | 4 |
| 5. | Methodology Used | 5 |
| 6. | Discussion | 6 |
| 7. | Learning Outcomes | 7 |
| 8. | Results | 8 |
| 9. | Conclusion | 10 |

## **1.Abstract**

Grape leaf diseases such as black measles, black rot, and leaf blight cause significant agricultural losses. Early detection can help farmers manage these diseases effectively, but traditional methods are time-consuming and inefficient. This paper presents an automated system using machine vision and machine learning to classify grape leaf diseases.

A novel algorithm integrates K-means clustering for image segmentation and support vector machines (SVM) for classification. The model processes leaf images and extracts relevant features in various color models (RGB, HSV, and L*a*b). Compared to convolutional neural networks (CNN) and GoogleNet, this approach is more accurate and less computationally intensive.

Results demonstrate classification accuracies up to 98.97%, offering a practical tool for farmers to diagnose grape leaf diseases in real-time.

## **2.Introduction**

**Project Objectives**

The primary objective of this project was to develop an automated and accurate system for diagnosing grape leaf diseases based on visual symptoms. The system needed to be highly efficient, fast, and reliable to aid farmers in large-scale agricultural settings. Key diseases targeted include black measles, black rot, and leaf blight.

**Problem Formulation**

Manual visual inspection of plant diseases is time-consuming, costly, and prone to human error, especially over large areas. Additionally, many diseases share similar visual symptoms, making identification difficult. Existing machine learning methods, including deep learning, require large datasets and extensive computation time, making them less feasible for real-time applications in agricultural fields. Therefore, this study aims to develop a robust and faster alternative that can perform with limited data and lower computational power while maintaining high accuracy.

## 

## **3.About the Dataset**

PlantVillage Dataset: A publicly available dataset containing labeled images of healthy and diseased grape leaves. The dataset includes three grape diseases: black measles, black rot, and leaf blight, as well as healthy leaves.

## **4.Background**

Agriculture remains the backbone of many economies, particularly in regions like Iran, where crop yield and quality are essential. Plant diseases significantly impact agriculture by reducing production, making early and accurate diagnosis critical for effective management. Historically, plant disease diagnosis relied on visual inspection by experts, which is labor-intensive, time-consuming, and prone to human error, especially in large fields.

The introduction of machine vision and machine learning has revolutionized disease detection. By leveraging computer vision and AI, it is now possible to monitor vast fields and automate disease identification. Machine learning models can analyze disease symptoms based on images of plant leaves, enabling early intervention and reduced losses.

The proposed study focuses on grape diseases, which include black measles, black rot, and leaf blight. These diseases share similar symptoms, complicating diagnosis. Traditional techniques often fail to differentiate between them, underscoring the need for advanced algorithms that can effectively identify and classify multiple diseases.

Image segmentation is a crucial step in disease diagnosis. Techniques like K-means clustering help separate healthy leaf tissue from diseased areas, but they require precise parameter tuning and manual input. Moreover, deep learning models like CNNs offer automated feature extraction but are challenged by the need for vast datasets and computational resources.

The study also emphasizes the importance of feature selection to reduce model complexity and improve efficiency. Methods like PCA and the Relief algorithm are used to streamline the feature set, making machine learning models more robust and easier to deploy in real-world agricultural settings.

## **5.Methodology Used:**

#### **Image Preprocessing**

Image preprocessing plays a crucial role in improving the accuracy of disease detection. In this study, the preprocessing step involved removing the background to avoid distractions from irrelevant features like soil or shadows. This was accomplished using a combination of gray-level thresholding for pixel clustering and edge detection via the Canny algorithm. Once the leaf's edges were identified, morphological dilation was applied to connect the leaf boundaries, filling in the blanks to create a cleaner image.

The removal of shadows, which is a common challenge in agricultural image processing, was managed through thresholding techniques. The processed images were converted into the RGB, HSV, and L*a*b color spaces, enabling more detailed analysis based on different color properties. The removal of the background resulted in higher classification accuracy as it allowed the model to focus solely on the leaf's features without interference from the background.

#### **K-Means Clustering**

K-means clustering was utilized to segment the diseased portions of the leaf. This unsupervised learning algorithm divides the dataset into K predefined clusters, minimizing the sum of squared distances between each pixel and its corresponding cluster center. In previous research, selecting the correct number of clusters (K) required manual input, which introduced variability and errors. In this study, however, the process was automated to eliminate the need for manual intervention.

By automatically segmenting the diseased area, the K-means clustering algorithm allowed the system to focus solely on the regions of interest, improving classification accuracy. This technique was applied both on the images with and without background, and it was found that background removal led to even higher accuracy, as it eliminated noise and irrelevant data.

#### **Feature Extraction and Selection**

Feature extraction is a key process for translating raw pixel data into meaningful insights that machine learning models can use for classification. The extracted features were divided into texture-based features (such as GLCM), color-based features (from the RGB, HSV, and L*a*b spaces), and shape-based features. A variety of algorithms, including Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Harris corner detection, were used to gather rich information from the images.

The extracted features were categorized into four groups:

1. **Captured Image**
2. **K-means Clustering of Captured Image**
3. **Images Without Background**
4. **K-means Clustering of Disease Area Without Background**

To enhance the performance of the SVM model, a feature selection technique called Relief was employed. Relief works by randomly selecting samples from the dataset and adjusting the importance scores of each feature based on their contribution to classifying the sample. By focusing only on the most significant features, this method reduced the feature space from 117 features to around 30, significantly reducing computational time while maintaining high accuracy.

#### **Principal Component Analysis (PCA)**

PCA was applied for dimensionality reduction to reduce the complexity of the feature space. This technique converts high-dimensional data into fewer dimensions while preserving the variance that contributes most to the model's accuracy. By eliminating redundant or irrelevant features, PCA helped the model avoid overfitting, reduced the time required for training, and simplified visualization.

#### **Support Vector Machine (SVM) Classifier**

SVM was chosen as the primary classifier due to its efficiency in handling high-dimensional data. Several kernel functions—linear, polynomial, radial basis function (RBF), and sigmoid—were tested. The linear kernel consistently provided the best results in terms of accuracy. The use of multi-class SVM allowed the model to classify the four categories of leaves (black measles, black rot, leaf blight, and healthy leaves) effectively.

**6.Discussion**

The study highlights the challenges and limitations of current disease detection techniques. Traditional image processing methods, while simple, struggle with accuracy and processing time. Deep learning models offer superior accuracy but are computationally expensive and require large training datasets, making them less feasible for field applications.

The proposed method addresses these issues by using automatic K-means clustering for image segmentation and SVM for classification. This approach minimizes human input and improves efficiency. The use of feature dimension reduction, like PCA, further enhances performance by eliminating redundant information and reducing computational overhead.

Comparing the proposed algorithm to deep learning models reveals several advantages. While CNN and GoogleNet achieve respectable accuracies, they take significantly longer to process images. In contrast, the study's algorithm delivers comparable accuracy in a fraction of the time, making it suitable for real-time applications in agriculture.

The research also underscores the potential of hybrid models that combine traditional and deep learning techniques. These models offer a balanced approach, leveraging the strengths of both methodologies. However, the study emphasizes that the choice of features and the efficiency of the clustering algorithm are crucial for achieving high performance.

**7.Learning Outcomes**

From this study, readers learn the importance of efficient and accurate plant disease detection methods. The research demonstrates how combining traditional and modern image processing techniques can yield significant improvements in diagnosis speed and precision.

The study teaches the value of feature selection and dimensionality reduction in machine learning. By using PCA and Relief, the algorithm reduces computational complexity while maintaining high accuracy, offering insights into the optimization of AI models for agricultural use.

Readers gain an understanding of the challenges associated with deploying deep learning models in real-world settings. The study emphasizes the need for high-quality, large datasets for deep learning and shows how traditional methods can serve as a viable alternative in resource-limited environments.

The research also explores the practical applications of machine vision in agriculture, highlighting its potential to revolutionize disease monitoring and crop management. By automating disease detection, farmers can make more informed decisions, reducing losses and improving productivity.

Finally, the study underscores the importance of ongoing innovation in agricultural technology. It encourages further exploration into hybrid models and real-time systems, emphasizing that advances in AI and computer vision can have a profound impact on global food security.

**8.Results**

**Accuracy**

The highest classification accuracy was achieved using the SVM model with linear kernel, particularly when applied to images with background removed and K-means clustering of the disease area. The proposed method achieved a remarkable accuracy of 98.97%, outperforming CNN (86.82%) and GoogleNet (94.05%).

**Processing Time**

The proposed method significantly reduced processing time compared to deep learning models. The SVM model, particularly when applied to background-removed images, required far less time for disease detection and classification compared to CNN and GoogleNet, which took up to 156 minutes for training.

**Dimensionality Reduction**

PCA further enhanced the efficiency of the model by reducing feature dimensions. After PCA, the SVM model still maintained a high classification accuracy of 98.97%, while reducing processing time to under 100 seconds.

**Feature Selection**

The Relief feature selection method identified critical features from the extracted GLCM, HOG, and LBP feature sets, reducing the number of necessary features from 117 to 30 while maintaining classification accuracy. This not only improved computational efficiency but also reduced the time for real-time detection.

precision recall f1-score support

Blight 0.98 1.00 0.99 247

HL 1.00 1.00 1.00 81

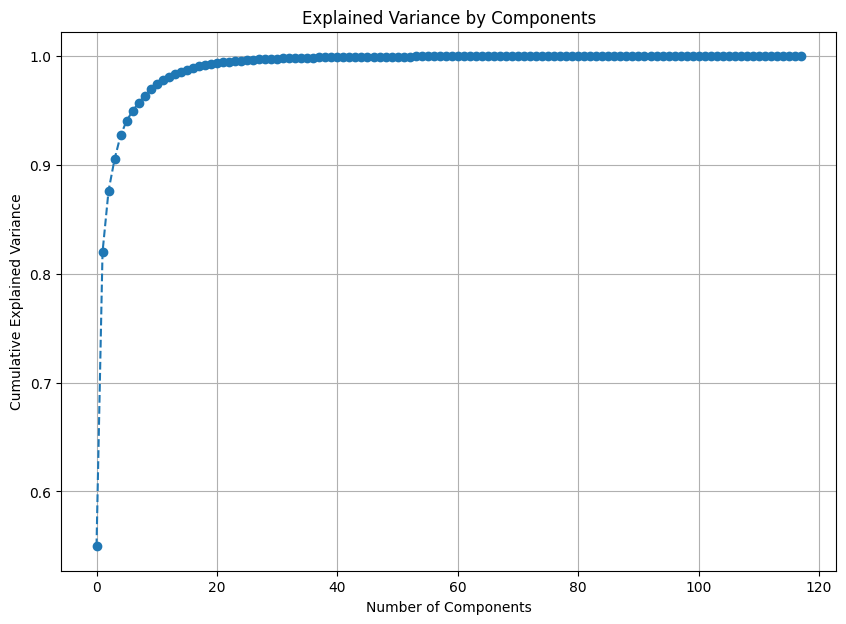
Msls 0.97 0.96 0.97 276

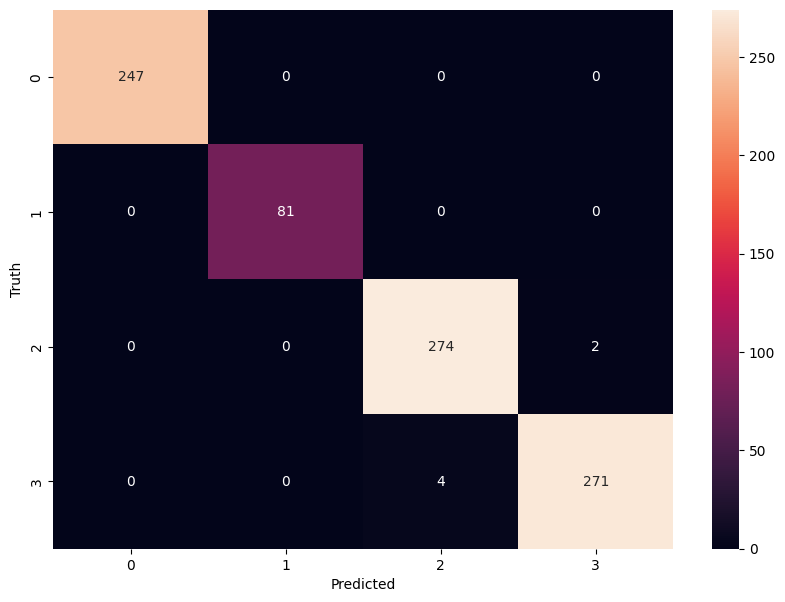
Rot 0.98 0.97 0.98 275

accuracy 0.98 879

macro avg 0.98 0.98 0.98 879

weighted avg 0.98 0.98 0.98 879





## **9.Conclusion**

This study demonstrates that a machine learning approach using SVM, K-means clustering, and PCA is highly effective for diagnosing grape leaf diseases. The proposed method outperforms deep learning models in both accuracy and processing time, making it a viable solution for large-scale, real-time agricultural disease monitoring. Additionally, the feature selection and dimensionality reduction techniques employed further optimize the system’s performance, ensuring that the model is both robust and computationally efficient.