**V SEMESTER B.Tech. (CCE)**

**ICT 3262: DATA MINING AND PREDICTIVE ANALYSIS LAB**

**REPORT**

**Classification of Apple Plant Leaves as Healthy or Diseased using Data Mining Algorithms and OpenCV**

**Tushaan Soi - 210953270**

**Lance Baretto - 210953276**

**CCE - B**

**Mr. Chethan Sharma**Assistant Professor - Senior Scale  
Department of Information & Communication Technology

**Mrs. Swathi B P**Assistant Professor -Senior Scale  
Department of Information & Communication Technology

**Index**

1. **Introduction ………………………………………………… 3**
2. **Literary Review ……………………………………………. 5**

1. **Methodology ………………………………………………… 6**
2. **Results and Discussion …………………………………… 13**
3. **Conclusion …………………………………………………...16**
4. **References ……………………………………………………19**
5. **Appendices …………………………………………………...20**

**1. Introduction**

**1.1 Background**

Early detection of plant diseases is critical for various reasons, including guaranteeing crop health and yield. These include but are not limited to:

1. Crop Loss Prevention - Early diagnosis enables for timely intervention and mitigation measures, limiting disease spread throughout the crop. As a result, crop loss is reduced and total output is maintained.

2. Cost Cutting - Early detection of plant diseases decreases the financial burden on farmers. When opposed to addressing widespread illnesses, which may need substantial resources, early intervention approaches are more cost-effective.

3. Food Security - Early detection supports global food security by ensuring a stable and reliable food supply. By preventing extensive crop damage, farmers can meet the demand for food production.

**1.2 Problem Statement**

The lack of an efficient and automated system for identifying diseased leaves poses a significant challenge, leading to delayed intervention, increased economic losses for farmers, and a heightened risk of disease spread throughout the orchards.

This project aims to develop a robust and automated system leveraging OpenCV and data mining algorithms for the classification of apple plant leaves as either healthy or infected. The primary objective is to enhance the efficiency of disease detection, allowing for timely intervention and reducing the impact of diseases on crop yield.

By addressing this challenge, the project seeks to contribute to the advancement of agricultural practices and the sustainable management of apple orchards, ultimately benefiting farmers and promoting food security.

**1.3 Objective**

The main objective of this project is to develop a system using data mining algorithms and OpenCV for automated classification of apple plant leaves. The main goals include:

1. Develop an Image Processing Pipeline

2. Build a Comprehensive Dataset

3. Preprocess Data

4. Feature Extraction Using Data Mining Algorithms

5. Select and Train Classification Model

6. Evaluate Model Performance

**1.4 Scope of the Study**

The scope of the study showcases the limitations and constraints that were set based on all the resources available to us:

1. Plant Species and Disease Focus - The study will focus specifically on the classification of apple plant leaves as healthy or infected. It does not extend to the identification of diseases in other plant species.

2. Image Data Source - The project will utilize a selected dataset of apple plant images for training and testing purposes. The study does not include the development of methods for capturing new images or extending the scope to other data sources.

3. Disease Types - The classification system will target a predefined set of common diseases affecting apple plants. It does not encompass the identification of all possible diseases or disorders that may affect apple plants.

4. Performance Metrics - The model's performance primarily relies on standard metrics which include accuracy, recall, precision and F1 score. The study does not extend to exploring additional metrics or customized evaluation methods.

**2. Literary Review**

In [1], 54306 images of plant leaves were analysed with 38 class labels assigned to them with each class label being a crop-disease pair. PlantVillage Dataset was used. The training testing ratio was 80:20. AlexNet and GoogLeNet were used as Deep Learning Architectures, with an accuracy of 98.36 and 96.21 respectively.

In [2]. K-Means Clustering Method was used alongside traditional Image Processing tools along with colour histograms and local binary patterns to classify 3 kinds of diseases found on apple leaves; 4 diseases on tomato leaf. 3-6 Layers of CNN were used and the classification accuracy reached was 93%.

In [3], Random Forest and Histogram of an Oriented Gradient were used collectively to classify different plant leaves from an original created dataset as diseased or healthy. An accuracy above 95% was achieved.

In [4], A Gaussian Filter was applied, Support Vector Machines and ANN were used in sugar beet fields, along with 4 types of common weed. Accuracy was 92.92% and 92.50% of weeds were correctly identified. It was a dataset of 600 images.

In [5], Digital Image Processing Techniques and Back Propagation Neural Networks (BPNN) were used. An accuracy of 83% was obtained using VNIR and 93% with full spectrum. F1 score obtained was 0.12. The plants used were Apple, Corns, Grapes, Potato and Tomato.

In [6], Bacterial, Viral and Fungal Diseases were identified on various Plant Leaves. Gabor Filter and ANN based classifier got an accuracy score of 91%. It was utilised with Fuzzy-C combination and Kmeans Clustering. Support Vector Machine was also used to obtain an accuracy of 94.74%.

In [7], CNN was used with Popular Deep Learning Models such as DenseNet-121, ResNet-50 and Inception 14. 38 output classes were used. In the initial 10 epochs 78% accuracy was obtained. And the final accuracy result was 84.27%. A fine tuned pre trained model got an accuracy of 99.81%.

In [8], CNN was used with AlexNet Architecture. The accuracy achieved with this method on plant leaves was 96.30% on the Maize Plant.

In [9], Tomato plant leaves were used to identify plant diseases and pests. CNN was used with GoogLeNet, AlexNet, VGGNet, ResNet. With pests the accuracy achieved was over 88%, and for Plant Disease Detection it was 95.87%.

In [10], CCD and CMOS image conversion were compared to analyse which is a faster image procession technique. ANN and K Neighbour Cluster was utilised and compared for accuracy. The established conclusion was a multicamera input with the above gives the best yield.

**3. Methodology**

**3.1 Data Collection**

The data was obtained from the PlantVillage Dataset. The original source may be found at <https://github.com/spMohanty/PlantVillage-Dataset/tree/master/raw/color..> This dataset was created using apple leaves. The Dataset is divided into two files for training purposes: Diseased and Healthy, which contain photos of leaves with labels.

The Diseased Folder includes diseased/unhealthy apples that have been infected with Apple Scab, Black Rot, or Cedar Apple Rust. Green and healthy photos make up the Healthy Folder.

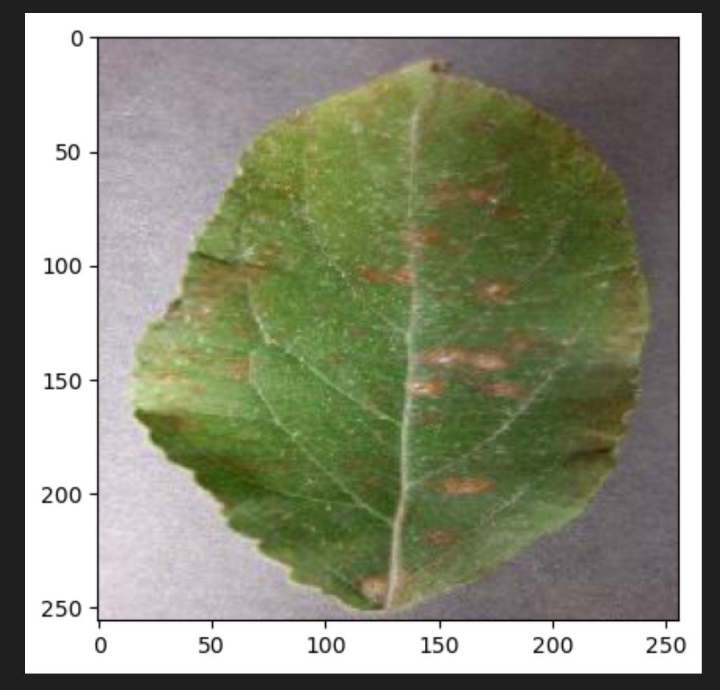
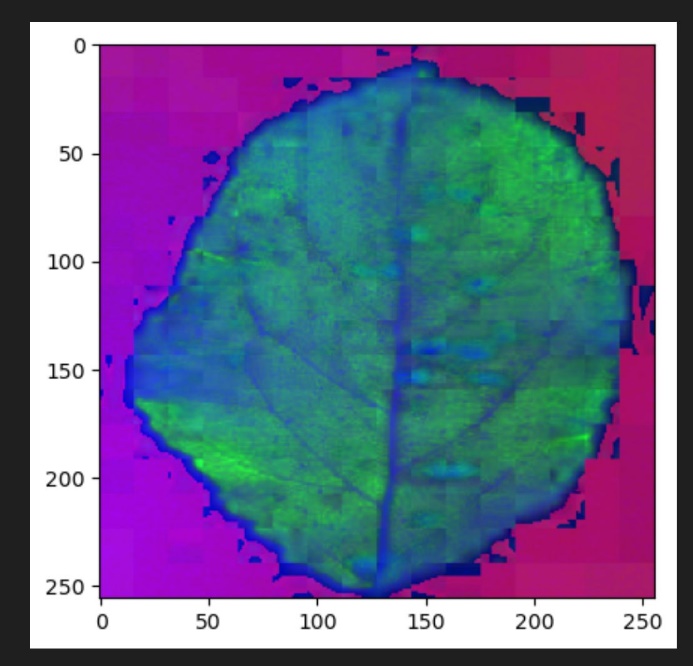
**3.2 Data Preprocessing**

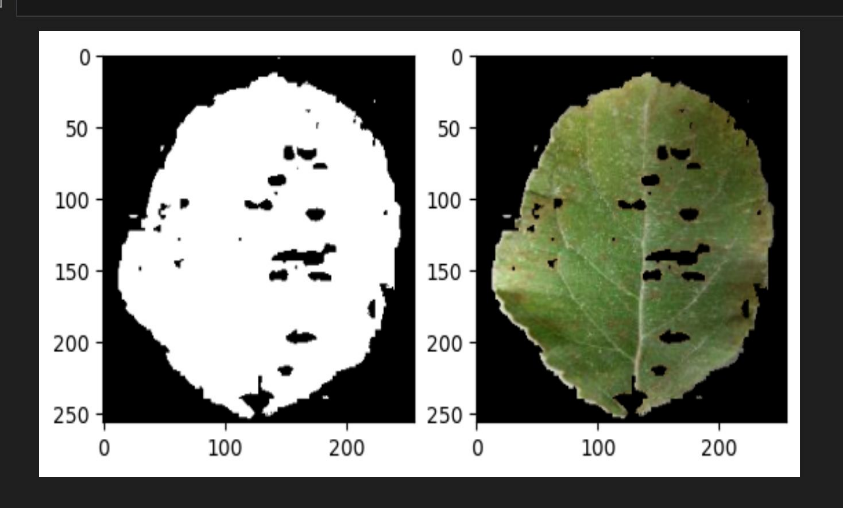
1. 800 images were used for each class: Diseased and Healthy.
2. The image was converted from RGB to BGR format. Since Open CV accepts images in their original form.
3. The images were then converted from BGR to HSV. HSV is used to distinguish luma (image intensity) and chroma (the colour information).

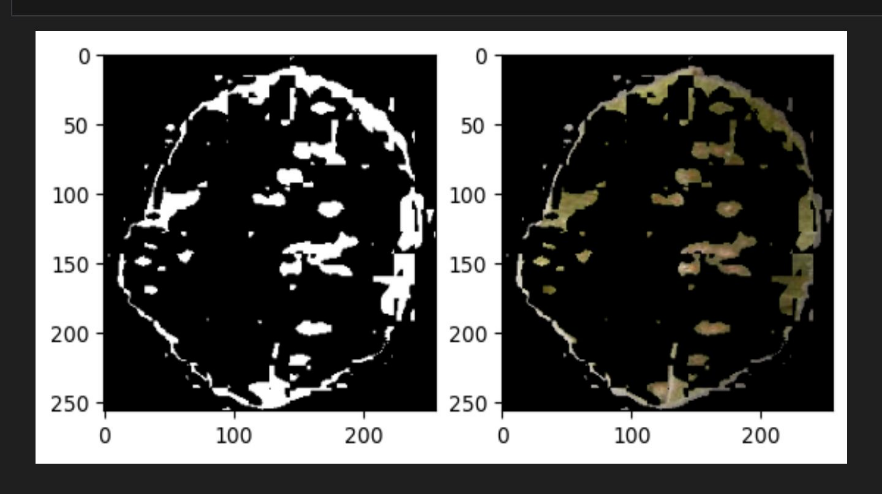
HSV is often used as the code for converting RGB to HSV is widely available and can be easily implemented.

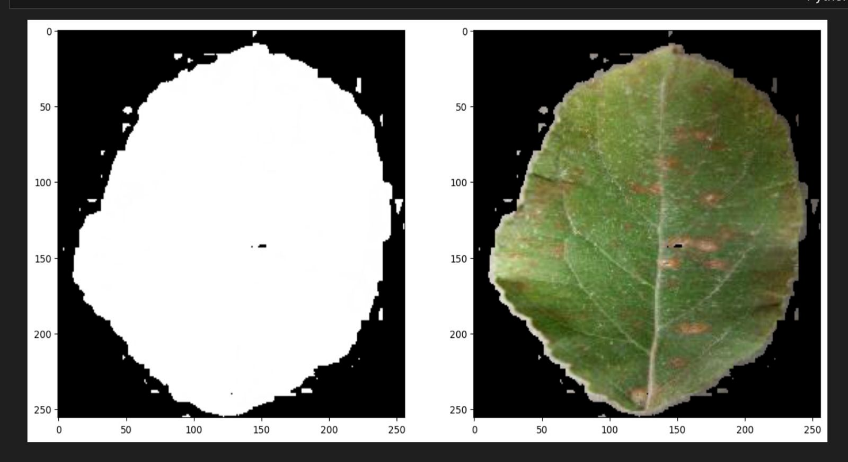
1. A close-up of a green leaf

   Description automatically generatedImage Segmentation for Colour Extraction - Image segmentation is required to separate the leaf image from the backdrop. The colour of the leaf is retrieved from the image.









**3.3 Feature Extraction**

3 features - Hu moments, Colour Histogram, and Haralick texture

Global Feature Descriptors were applied. They are taken out from the image using three main descriptors:

• Colour - Color Channel Statistics and Colour Histogram

• Shape - Hu Moments

• Texture - Haralick Texture

**Hu Moments**, also known as Hu's or Hu's invariant moments, are a set of seven scalar, rotation-invariant, and scale-invariant image moments used in image analysis and computer vision. Hu Moments are particularly useful for shape recognition and object classification tasks.

The calculation of Hu Moments involves complex mathematical operations based on the moments of the image. OpenCV and other image processing libraries provide functions to compute Hu Moments, making them accessible and practical for various computer vision tasks, including shape analysis, object recognition, and image-based quality control.

**Haralick texture** features, also known as Haralick texture descriptors or texture co-occurrence features, are a set of statistical measures used to characterize the texture or patterns in an image.

Haralick texture features are particularly useful for quantifying the spatial relationships between pixel intensity values in an image.

**Colour histogram** is a graphical representation of the distribution of colours in an image. It counts the frequency of different colour values or colour channels in an image and provides a quantitative description of the colour content within the image.

Colour histograms are widely used in image processing, computer vision, and graphics for various applications, including image retrieval, object recognition, and image analysis.

The purpose of this code is to organize the training dataset for supervised machine learning. It collects the class labels, sorts them for consistency, and creates two empty lists to accumulate the extracted features and their corresponding class labels.

These features and labels will later be populated as the code processes each image in the training dataset, extracting features and assigning the appropriate class labels. This organized data can then be used to train and evaluate machine learning models for image classification.

**3.4 Selection of Data Mining Algorithms**

The Model is trained over 7 machine learning models:

• Linear Discriminant Analysis

• Logistic Regression

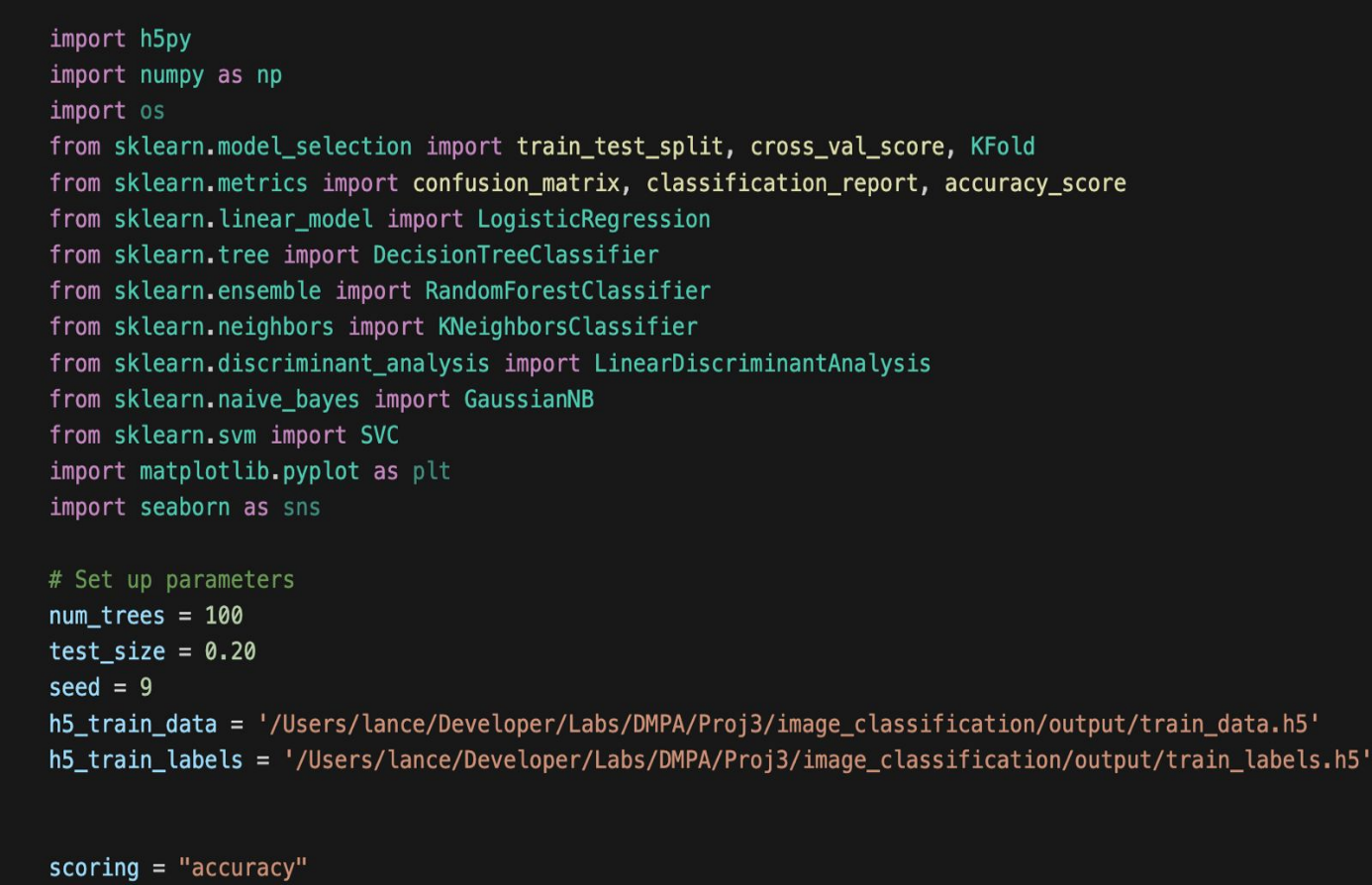
• Support Vector Machine

• K Nearest Neighbours

• Decision Trees

• GaussianNB

• Random Forest



**3.5 Model Training**

The Dataset is split into training and testing set with an 80:20 ratio respectively.

During training, the model learns to map input features (extracted from images) to the corresponding output labels (healthy or infected). Adjust model parameters to optimize its performance on the training set.

Furthermore, the model is verified using the 10k fold cross validation approach. The dataset is folded into 'k' equal portions. After that, the model is trained on 'k-1' folds and tested on the remaining fold. This procedure is repeated 'k' times using a new fold as the test set each time, and the performance is averaged over these iterations to provide a more robust assessment of the model's efficacy.

**A screen shot of a computer program

Description automatically generated**

**3.6 Evaluation Metrics**

The evaluation of the developed model for classifying apple plant leaves as healthy or infected relies on a comprehensive set of metrics to provide a nuanced understanding of its performance.

**Accuracy**, a fundamental measure, gauges the overall correctness of the classification. However, to gain deeper insights, precision, recall, and F1 score come into play.

**Precision** quantifies the model's ability to correctly identify infected leaves among those predicted as infected, while recall measures its effectiveness in capturing all truly infected leaves.

The **F1 score**, which harmonizes precision and recall, is especially valuable in scenarios where a balance between false positives and false negatives is critical.

The **confusion matrix** serves as a visual aid, breaking down the model's predictions into true positives, true negatives, false positives, and false negatives, offering a holistic view of its classification performance.

**Support**, indicating the number of occurrences of each class, contributes context to the precision and recall values.

Furthermore, **boxplot diagrams** complement these metrics by providing a graphical representation of the distribution of prediction scores, aiding in the identification of potential outliers or areas of uncertainty.

Collectively, these evaluation metrics and visualizations furnish a robust framework for assessing the reliability and effectiveness of the model in classifying apple plant leaves.

**4. Results and Discussion**

**4.1 Model Performance**

**Random Forest** (RF) emerges as the top performer with an accuracy of 97.81%, showcasing its robustness and effectiveness in handling complex patterns without overfitting.

Based on that, the visual results in the report would include results from the RF Algorithm.

Logistic Regression (LR), Classification and Regression Trees (CART), Support Vector Machine (SVM) and K-Nearest Neighbours (KNN), display commendable performances with accuracies above 92%, indicating their suitability for linear and non-linear relationships.

Linear Discriminant Analysis (LDA) follows closely, suitable for linear classifications. However, Naive Bayes (NB) lags behind, with an accuracy of 86.17%.

LR: 0.924193

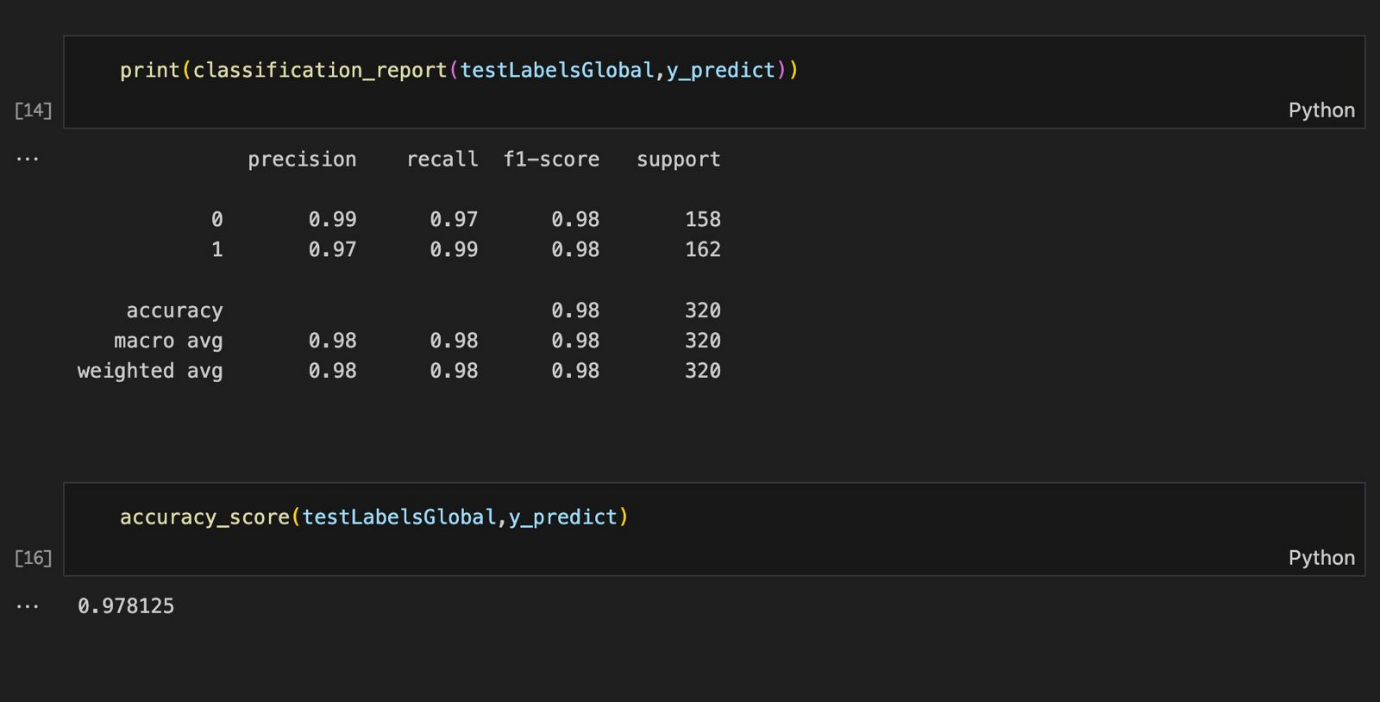
LDA: 0.901725

KNN: 0.934276

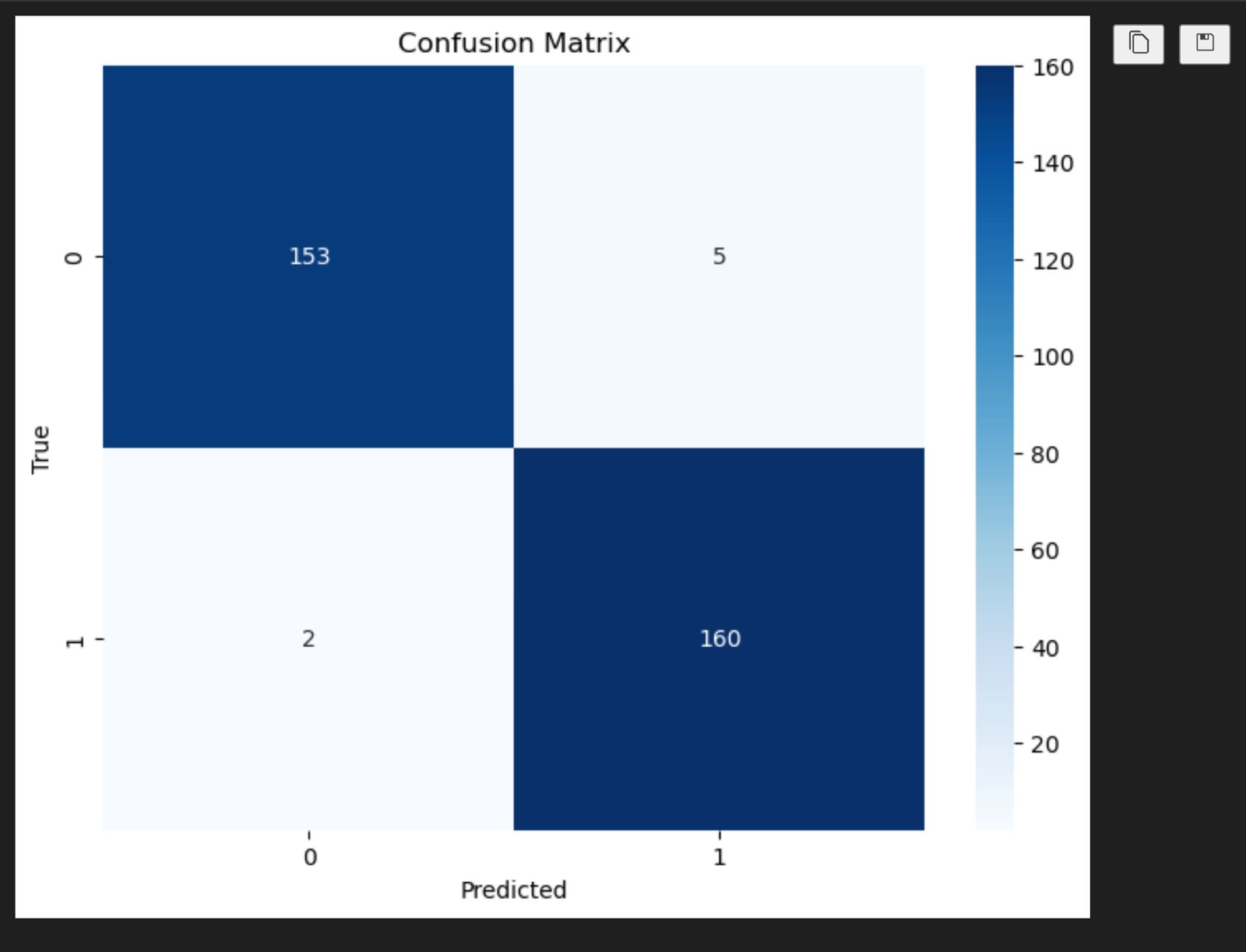
CART: 0.921494

**RF: 0.978125 (0.015149)**

NB: 0.861719

SVM: 0.927152

**4.2 Visualizations**



A **confusion matrix** was obtained showing the results of:

True Positive (153), True Negative (160), False Positives (5) and False Negatives (2).

A screen shot of a graph

Description automatically generated

A **boxplot chart** was obtained describing the result of each Algorithm in a candlestick diagram.

**4.3 Discussion of Results**

An accuracy of 97.81% and an F1 score of 0.98 for the Random Forest model in classifying apple plant leaves as healthy or infected are highly promising results. The accuracy metric indicates that the model correctly classified 97.81% of the instances in the dataset. This high accuracy suggests a robust overall performance in distinguishing between healthy and infected leaves.

The F1 score of 0.98 is particularly noteworthy. The F1 score is the harmonic mean of recall and precision, with a range between 0 and 1, with higher values indicating better model performance. In this case, a score of 0.98 signifies a balanced trade-off between precision and recall. The model effectively minimizes both false positives and false negatives, demonstrating its ability to accurately identify infected leaves while maintaining a low rate of misclassifying healthy leaves.

These results indicate that the Random Forest model is well-suited for the task of classifying apple plant leaves based on the features extracted using OpenCV and the selected data mining algorithms. The high accuracy and F1 score suggest that the model has learned intricate patterns and features associated with healthy and infected leaves, making it a reliable tool for automated disease detection in apple orchards.

However, it's crucial to conduct a comprehensive analysis of other evaluation metrics, such as precision, recall, and the confusion matrix, to gain a better insight on the model's strengths and potential areas for improvement. Additionally, considering the context of the specific application and the distribution of classes in the dataset will provide a more complete assessment of the model's practical utility. Overall, the reported accuracy and F1 score are promising indicators of the Random Forest model's success in the classification task.

**5. Conclusion**

**5.1 Summary of Findings**

In summary, the findings of the study on classifying apple plant leaves as healthy or infected using OpenCV, and Data Mining algorithms are highly encouraging. With Random Forest, the model achieved an impressive accuracy of 97.8%, indicating its ability to correctly classify nearly 98% of the instances in the dataset. This high accuracy is indicative of the model's overall robust performance in distinguishing between healthy and infected leaves.

Moreover, the F1 score, a metric that balances precision and recall, reached an outstanding value of 0.98. This suggests that the model effectively minimized both false positives and false negatives, showcasing its proficiency in accurate disease identification while maintaining a low rate of misclassifying healthy leaves.

These results highlight the effectiveness of leveraging Random Forest, an ensemble learning algorithm, in combination with image processing techniques from OpenCV for the automated classification of plant diseases. The model demonstrates a strong capability to discern intricate patterns and features associated with healthy and infected apple plant leaves.

In conclusion, the study's findings suggest that the developed model holds significant promise as an efficient and reliable tool for early disease detection in apple orchards.

**5.2 Contributions**

The study on classifying apple plant leaves as healthy or infected using OpenCV, and data mining algorithms contributes significantly to the field of agricultural technology and plant disease detection.

1. Automation of Disease Detection - The development of an automated system for classifying plant diseases contributes to the ongoing efforts in automating agricultural processes. The study demonstrates the feasibility and effectiveness of machine learning and image processing to replace or complement manual inspection.

2. Early Disease Identification - By achieving a high accuracy and F1 score, the study contributes to the early identification of plant diseases. Early detection is crucial for implementing timely interventions, preventing the spread of diseases, and minimizing crop losses. The model's capability to accurately identify infected leaves supports proactive and targeted disease management practices.

3. Integration of Data Mining and Image Processing - The study showcases the successful integration of data mining algorithms, specifically the Random Forest model, with image processing techniques from OpenCV. This interdisciplinary approach demonstrates the synergy between machine learning and computer vision in solving complex agricultural challenges, setting a precedent for future research in precision agriculture.

4. Practical Applicability in Agriculture - The high accuracy and F1 score suggest that the developed model has practical applicability in agricultural settings. Farmers and orchard managers can potentially use this automated system as a tool for routine monitoring, enabling them to make informed decisions about disease management strategies and optimize resource allocation.

In summary, the study's contributions extend beyond the specific task of classifying apple plant leaves and offer valuable insights and tools that can positively impact agricultural practices, particularly in the realm of precision agriculture and automated disease management.

**5.3 Limitations**

Despite the promising results and contributions, it's essential to acknowledge and discuss the limitations encountered during the study. Understanding these limitations provides context for interpreting the findings and suggests areas for potential future research and improvement:

1. Limited Dataset Size - Expanding the dataset with more diverse examples could enhance the model's robustness.

2. Sensitivity to Environmental Conditions - The study may not have accounted for all possible environmental factors such as lighting, humidity, or background clutter in the images, affecting the model's generalizability to diverse real-world scenarios.

3. Algorithm Sensitivity – Fine-tuning hyperparameters and exploring alternative algorithms could be avenues for further improvement.

4. Single Crop Focus - The study specifically focuses on classifying apple plant leaves. Extending the study to multiple crops could enhance its applicability.

Recognizing these limitations underscores the need for ongoing refinement and expansion of the study's methodologies. Future research could address these challenges, providing a more comprehensive and robust solution for automated plant disease detection in diverse agricultural contexts.

**5.4 Future Work**

There are many things upon which we can improve the current model. It includes but is not limited to the following:

1. Dataset Expansion - Increase the dataset size to improve the model's generalization, encompassing a wider range of environmental conditions and disease variations.

2. Environmental Sensitivity Analysis - Conduct a thorough investigation into the model's sensitivity to environmental conditions, exploring variations in lighting, humidity, and background scenarios to enhance robustness.

3. Algorithmic Exploration - Explore alternative data mining algorithms beyond Random Forest, assessing their suitability for plant disease classification and comparing their performance.

4. Multi-crop Analysis - Extend the study to include multiple crop species, evaluating the model's performance in diverse agricultural contexts and enhancing its versatility.

**6. References**

[1] Sharada P. Mohanty, David P. Hughes, Marcel Salathe, “Using Deep Learning for Image Based Plant Disease Detection”, \*Front Plant Sci\*, 2016

[2] Lili Li, Shijuan Zhang, Bin Wang, “Plant Disease Detection and Classification by Deep Learning”, \*IEEE Volume 9\*, 2020

[3] Shima Ramesh, Ramachandra Hebbar, Niveditha R, Pooja R, “Plant Disease Detection using Machine Learning”, \*IATMSI IEEE conference\*, 2018

[4] C Jackulin, S. Murugavalli, “A comprehensive review on detection of plant diseases using Machine Learning and Deep Learning Approaches”, \*ElseV Measurements: Sensor\*, Volume 24, December 2020

[5] Pranesh Kulkarni, Atharva Karwande, Tejas Kolhe, “Plant Disease detection using Image Processing and Machine Learning”, \* ET, Vishwakarma Institute Conference\*, 2019

[6] Kiran R. Gavhale, “An overview on the research on Plant Leaves Detection using image processing Techniques”, \*IOSR Journal of Computer Engineering\*, January 2014

[7] Andrew J., Jennifer Eunice, Daniela Elena Popescu, M. Kalpana Chowdary, Jude Hemanth, “Deep Learning Based Leaf Disease Detection in Crops using Images for Agricultural Applications, \*MDPI: Agronomy\*, 2022

[8] Muhammad Hammad Saleem, Johan Potgieter, Khalid Mahmood Arif, “Plant Disease Detection and Classification by Deep Learning”, \*National Library of Medicine: Plants (Basel)\*, October 2019

[9] Jun Liu, Xuewei Wang, “Plant Disease and Pests Detection based on Deep Learning”, \*BioMedCentral\*, February 2021

[10] Mbulelo S. P. Ngongoma, Musasa Kabeya, Katleho Moloi, “A review of plant disease Detection Systems for Farming Applications”, \*MDPI: Artificial Intelligence Tools and Models\*, 2022

**7. Appendices**

* Final Result of Image PreProcessing:

