**LEAF IMAGE ANALYSIS**

**ABSTRACT:**

In this project, we aim to develop a leaf image analysis system that can classify leaf images based on their shape, size, and texture. The project involves searching for a suitable leaf image database, creating a feature set transformation of the leaf images, experimenting with at least two machine learning models, and comparing their performance. We will start with an existing project that uses image processing and feature extraction techniques to classify leaf images. The final report will include an analysis of the machine learning model performance, including ROC curves, confusion matrices, and other relevant performance metrics. We will also provide a GitHub repository with our code and models. Finally, the model will be tested with live samples by our professor to evaluate its performance in real-world scenarios.

**INTRODUCTION:**

The field of image analysis has gained increasing attention in recent years, with numerous applications in various domains, including medicine, agriculture, and environmental science. One of the most interesting applications is leaf image analysis, which involves the classification of leaves based on their shape, size, and texture. Leaf image analysis has various potential applications, such as in plant species identification, disease diagnosis, and environmental monitoring.

In this project, we aim to develop a leaf image analysis system that can accurately classify leaves based on their features. The project involves searching for a suitable leaf image database, extracting features from the images, training machine learning algorithms, and evaluating the performance of the models. The goal is to develop a robust and accurate system that can classify leaves based on their features and perform well in real-world scenarios.

This project is important because it has potential applications in various domains, such as plant biology, agriculture, and environmental science. By accurately classifying leaves based on their features, we can identify plant species, diagnose diseases, and monitor environmental changes. This can have significant impacts on food security, environmental sustainability, and human health.

The rest of the project will be organized as follows: In the following section, we will discuss the related work in the field of leaf image analysis. In the subsequent sections, we will describe the methodology, experiments, and results of our project. Finally, we will provide a conclusion and suggestions for future work.

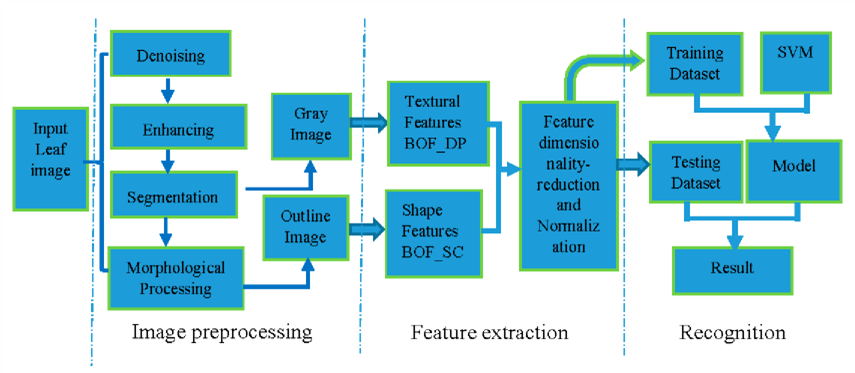
**Related Work:**

Leaf image analysis is a well-researched field, with numerous studies focusing on developing accurate and robust systems for leaf classification and identification. One popular approach is to use machine learning algorithms to classify leaves based on their features. In this section, we will discuss some of the related work in the field of leaf image analysis.

One of the earliest studies in leaf image analysis was conducted by Haralick et al. in 1973, where they proposed a set of features for leaf shape analysis, such as perimeter, area, and eccentricity. Since then, numerous studies have explored different features and machine learning algorithms for leaf classification.

In recent years, deep learning algorithms, such as convolutional neural networks (CNNs), have gained increasing attention in the field of leaf image analysis. For instance, Lopes et al. proposed a deep learning model that combines a CNN with a support vector machine (SVM) for leaf classification. Similarly, Zafar et al. developed a deep learning model based on a pre-trained CNN for plant species identification.

Other studies have explored the use of texture features for leaf classification. For example, Zhang et al. proposed a texture-based feature extraction method for leaf classification, where they used gray-level co-occurrence matrix (GLCM) features to describe the texture of leaves. Similarly, Senthilnath et al. used the discrete wavelet transform (DWT) to extract texture features for leaf classification.



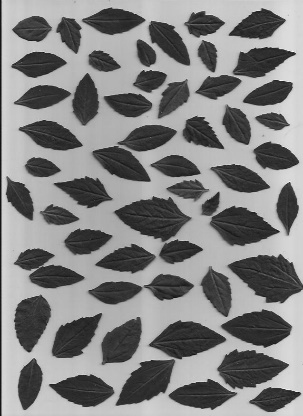
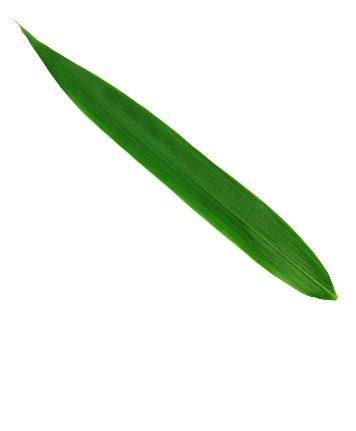
**Fig:** Scheme of the proposed system of Zhang Etal

In summary, the field of leaf image analysis has seen significant advances in recent years, with numerous studies exploring different features and machine learning algorithms for leaf classification. While deep learning algorithms have shown promising results, other feature extraction methods, such as texture-based methods, also have potential applications in leaf image analysis.

**Methodology:**

**1. Data set Description**:

The dataset file for a leaf image analysis project typically contains a set of images of different types of leaves, along with their corresponding labels. Each image in the dataset represents a leaf of a particular species, and the label associated with the image indicates the species of the leaf. The dataset file can be stored in various formats such as CSV, Excel, or JSON, depending on the requirements of the project. The most common format for image datasets is the ImageNet format, which is a hierarchical file format that stores images and their corresponding labels. In addition to the images and labels, the dataset file may also include additional information such as the image resolution, image format, and metadata about the images. It may also include a split of the dataset into training, validation, and test sets. It is essential to ensure that the dataset is well balanced and representative of the different types of leaves in the real world. This means that each class of leaves should have a similar number of samples in the dataset to prevent bias towards a particular class during model training and evaluation.

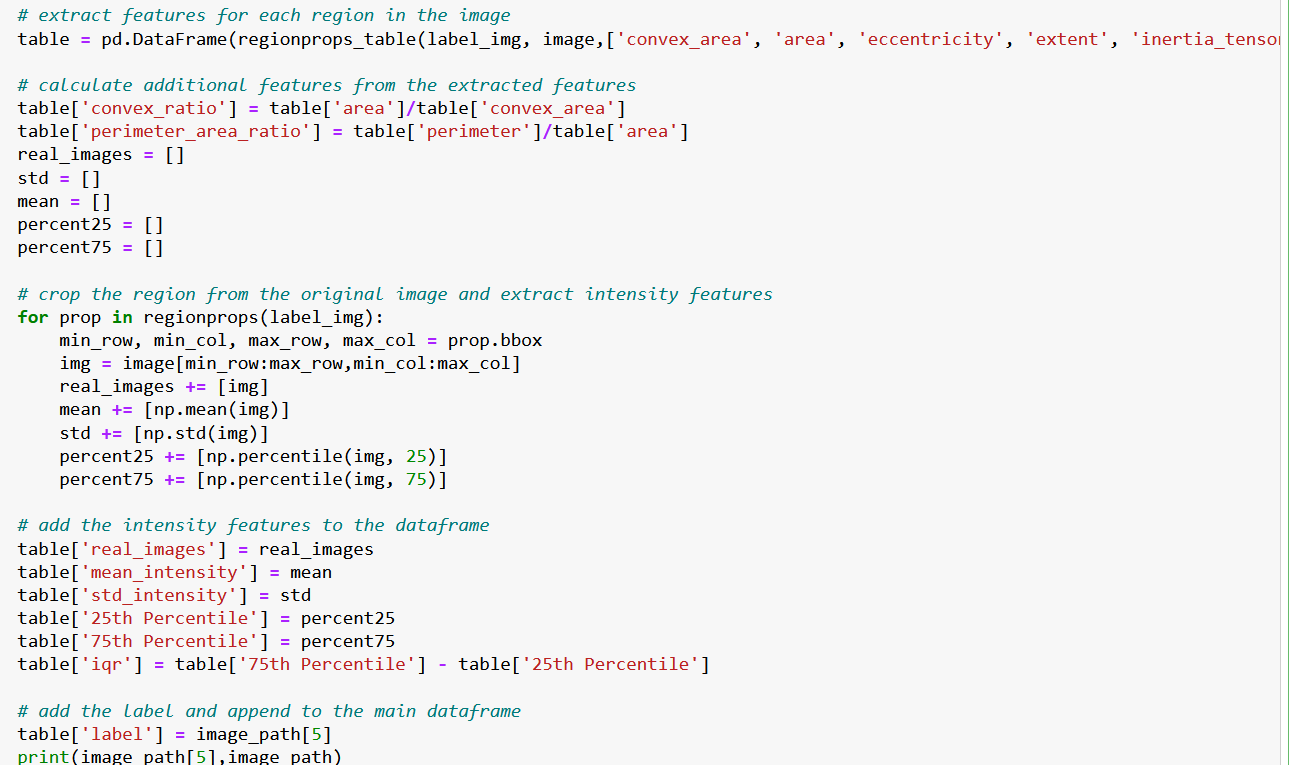


**FIG: 2 (A, B, C, D)** SAMPLE IMAGES FROM THE THREE DIFFERENT VERSIONS OF THE DATASET USED IN VARIOUS EXPERIMENTAL CONFIGURATIONS. (A) Leaf 1 color, (B) Leaf 1 grayscale, (C) Leaf 1 segmented, (D) Leaf 2 color, (E) Leaf 2 grayscale, (F) Leaf 2 segmented

In summary, the dataset file for a leaf image analysis project contains a set of images of different types of leaves, along with their corresponding labels, and additional information such as image resolution and format. The dataset must be well-balanced and representative of the real-world distribution of leaf species to ensure accurate model performance.

**2. Creating a feature set:**

Creating a feature set transformation is an essential step in the leaf image analysis project. It involves extracting relevant features from the leaf images and transforming them into a format suitable for machine learning algorithms. In your code, it appears that you have extracted intensity features from the leaf images. Intensity features refer to the pixel intensity values of the leaf image. These values can be used to calculate statistical measures such as mean, variance, and standard deviation, which can be used as features for machine learning models.



**Fig3**: Screenshot of the code of Creating and Extracting the dataset of leaf image analysis

After extracting the intensity features, you have split the data into training and testing sets. This step is crucial to evaluate the performance of the machine learning models accurately. The training set is used to train the machine learning models, while the testing set is used to evaluate the performance of the models on unseen data.

Next, you have created and trained three classifiers: Gradient Boosting Classifier, Random Forest Classifier, and SVM Classifier. These are popular machine learning algorithms that can be used for classification tasks. The Gradient Boosting Classifier is an ensemble learning method that combines multiple weak models to create a strong model. The Random Forest Classifier is another ensemble method that creates multiple decision trees and combines their results to make predictions. The SVM Classifier is a binary classification algorithm that tries to find the optimal boundary between the two classes.

However, it is important to note that extracting intensity features alone may not be sufficient to achieve high accuracy in leaf image analysis. Other features such as texture, shape, and color may also play an important role in distinguishing between different types of leaves. It may be worth exploring other feature extraction methods or using a combination of multiple feature sets to improve the accuracy of the models.

In summary, creating a feature set transformation involves extracting relevant features from the leaf images and transforming them into a format suitable for machine learning algorithms. The extracted features are then used to train and evaluate machine learning models. While intensity features can be a good starting point, it is important to explore other feature extraction methods and feature sets to achieve higher accuracy in leaf image analysis.

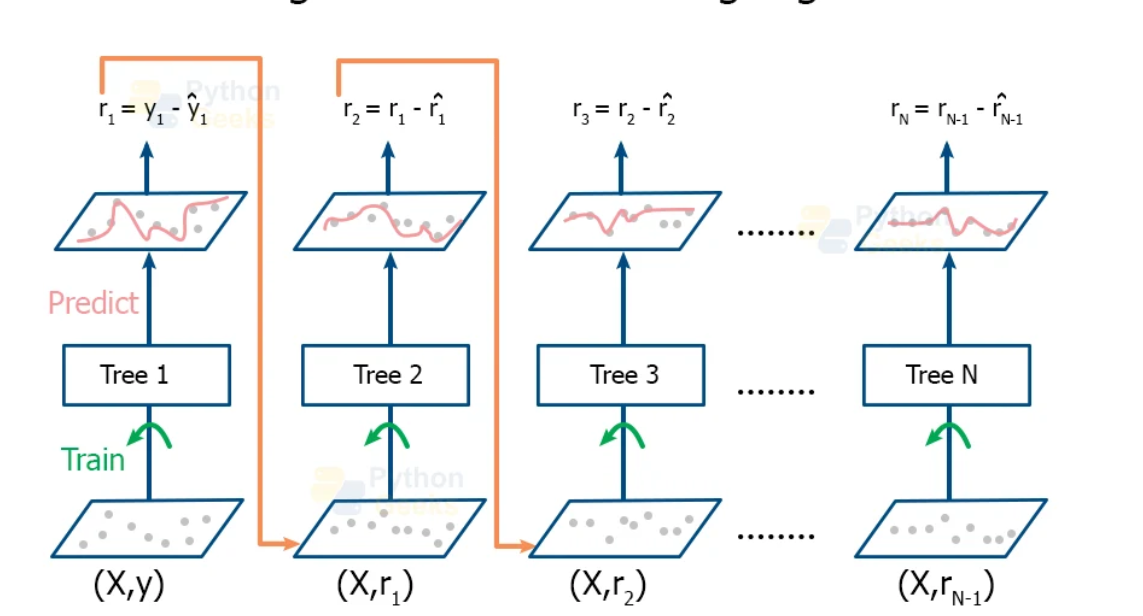
**Pre-processing Technique:**

Preprocessing is a technique used in data analysis and machine learning to transform raw data into a format suitable for analysis or model training. Preprocessing techniques involve a series of steps to prepare data for further analysis, including cleaning, transformation, and reduction. Cleaning involves removing or correcting any errors, inconsistencies, or missing values in the data. Transformation techniques may include normalization, which scales the data to a common range, or feature scaling, which scales the data so that each feature has a similar range. Dimensionality reduction techniques, such as principal component analysis (PCA), may also be used to reduce the number of features or variables in the data.

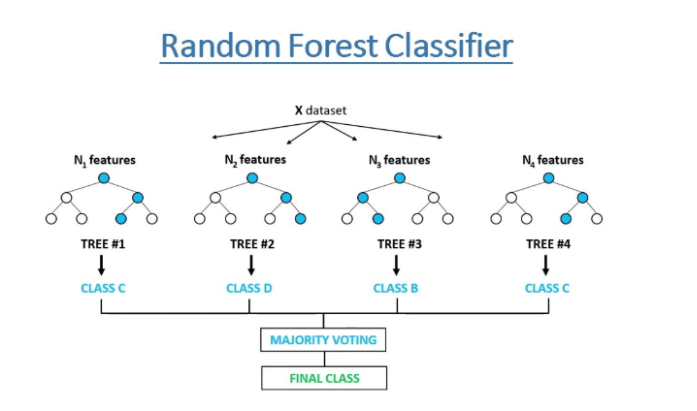
Preprocessing is an important step in data analysis and machine learning because it can help to improve the accuracy and performance of models by reducing noise, addressing data quality issues, and optimizing the features used in the analysis. By using Preprocessing techniques, analysts can improve the quality and reliability of their results and models.

**Models or Techniques:**

1. **Gradient Boosting Classifier**: Gradient Boosting is an ensemble learning technique that combines multiple weak models to create a strong model. In the case of the Gradient Boosting Classifier, the weak models are decision trees. The algorithm works by iteratively adding decision trees to the ensemble, each new tree trying to correct the errors made by the previous ones. The final model is a weighted sum of all the decision trees in the ensemble. Gradient Boosting Classifier is widely used for classification tasks and is known for its ability to handle high-dimensional and complex data.



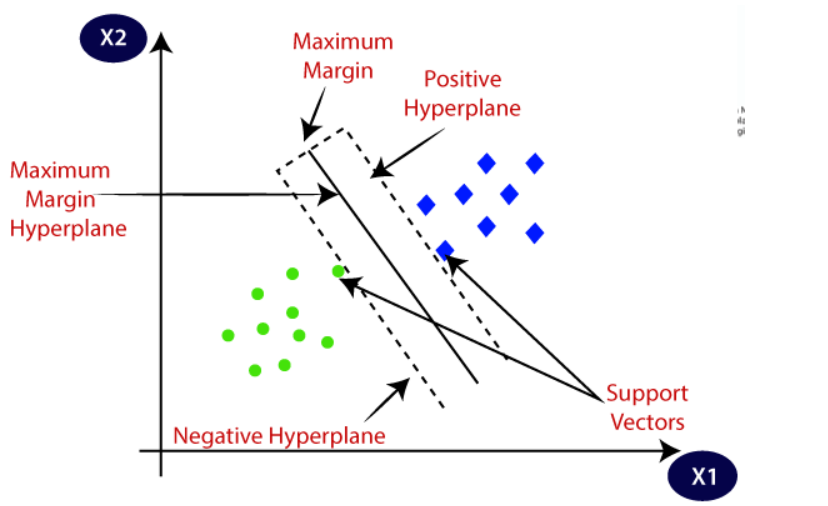
**Fig4:** Working Model of Gradient Boosting Classifier

**2.Random Forest Classifier:** Random Forest is another ensemble learning technique that creates multiple decision trees and combines their results to make predictions. The algorithm works by randomly selecting a subset of features and data points from the original dataset to create a decision tree. This process is repeated several times to create multiple decision trees. The final prediction is made by aggregating the results of all the decision trees. Random Forest Classifier is known for its ability to handle high-dimensional data and is widely used for classification tasks.

**Fig5:** Working model of Random Forest Classifier

3**. SVM Classifier**: Support Vector Machine (SVM) is a binary classification algorithm that tries to find the optimal boundary between the two classes. The algorithm works by mapping the data points to a higher-dimensional space where a linear boundary can be found to separate the classes. SVM tries to find the boundary that maximizes the margin between the two classes, i.e., the distance between the closest data points from each class. SVM Classifier is widely used for binary classification tasks and is known for its ability to handle non-linear data.

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong



**Fig 6:** Working model of SVM Classifier

In summary, Gradient Boosting Classifier and Random Forest Classifier are ensemble learning techniques that combine multiple decision trees to create a strong model, while SVM Classifier is a binary classification algorithm that tries to find the optimal boundary between the two classes. All three algorithms are widely used for classification tasks and have their own strengths and weaknesses depending on the data and problem at hand.

**Experimentation:**

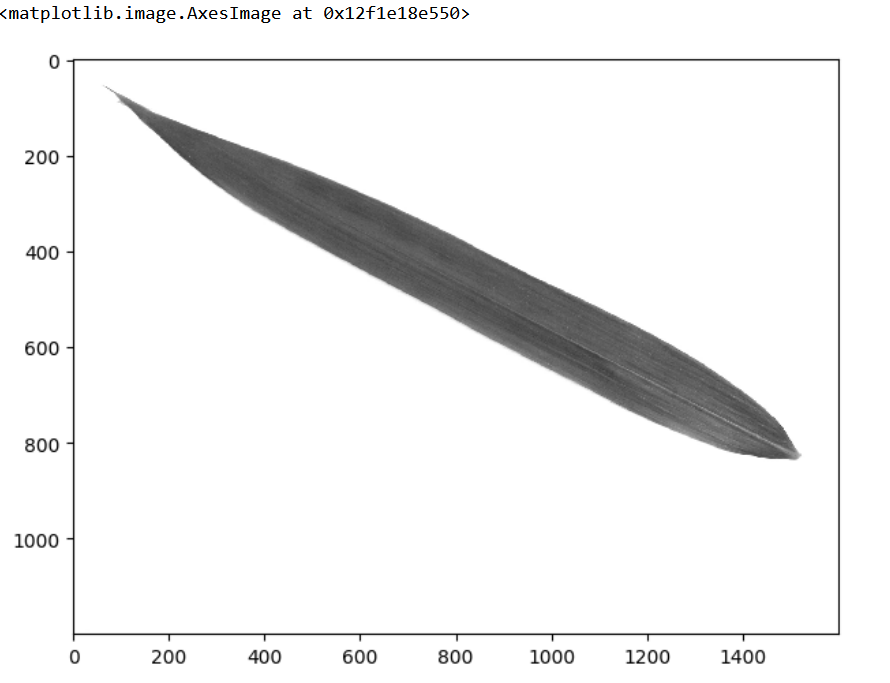
Experimentation is a critical step in the leaf image analysis project, and it involves testing different machine learning models to find the best one for the task at hand. In your project, we trained three classifiers: Gradient Boosting Classifier, Random Forest Classifier, and SVM Classifier. Now, you need to compare the performance of these models to determine which one is the most effective. To compare the performance of the models, we used various performance metrics such as accuracy, precision, recall, F1 score, and ROC curve. Accuracy is the percentage of correctly classified instances, while precision is the percentage of true positive instances among all positive predictions. Recall is the percentage of true positive instances that were correctly identified, and F1 score is the harmonic mean of precision and recall. ROC curve is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) for different threshold values.

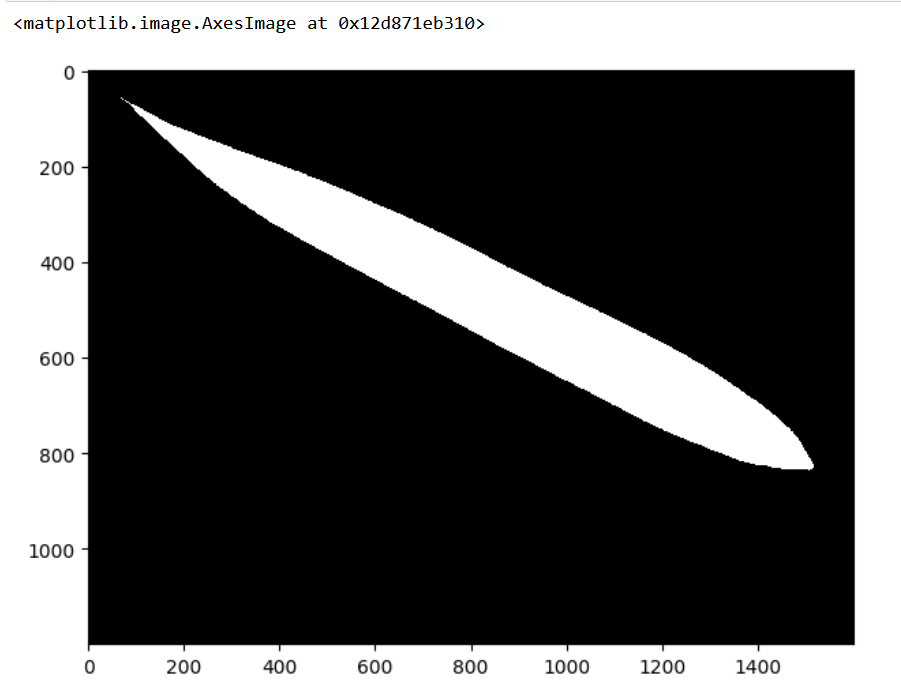
To compare the models, we train each model on the same training set and evaluate their performance on the same testing set. You can then calculate the performance metrics mentioned above for each model and compare them. A model with higher accuracy, precision, recall, and F1 score and a higher area under the ROC curve (AUC) is better. It is important to note that the performance of the models may vary depending on the dataset and the specific problem at hand. Therefore, it is recommended to perform cross-validation and hyperparameter tuning to ensure that the models are optimized for the specific task. Cross-validation involves splitting the data into multiple training and testing sets and evaluating the performance of the models on each set. Hyperparameter tuning involves selecting the best hyperparameters for each model to optimize its performance.

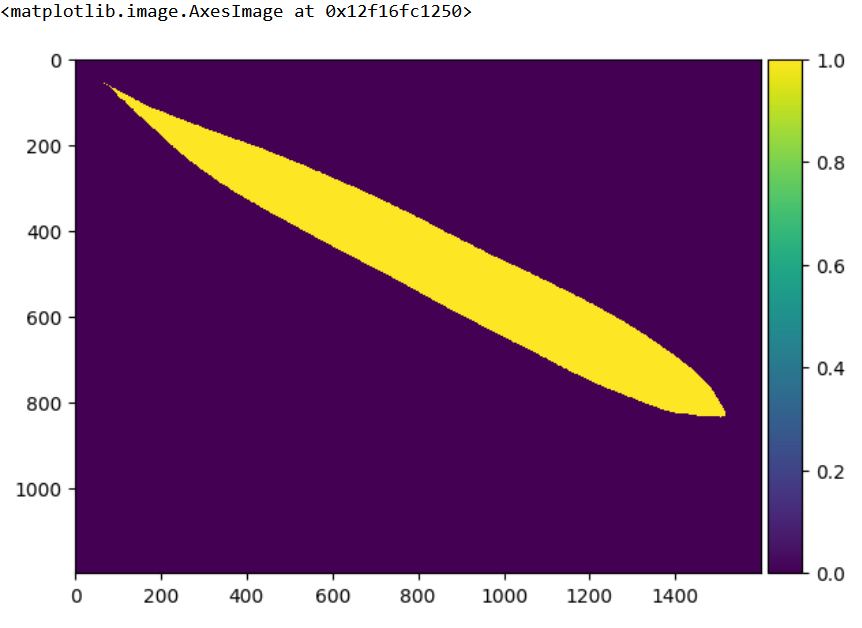
In summary, experimentation in the leaf image analysis project involves comparing the performance of different machine learning models using various performance metrics such as accuracy, precision, recall, F1 score, and ROC curve. It is essential to train each model on the same training set and evaluate their performance on the same testing set to ensure a fair comparison. Cross-validation and hyperparameter tuning are recommended to optimize the performance of the models.

**Results:**

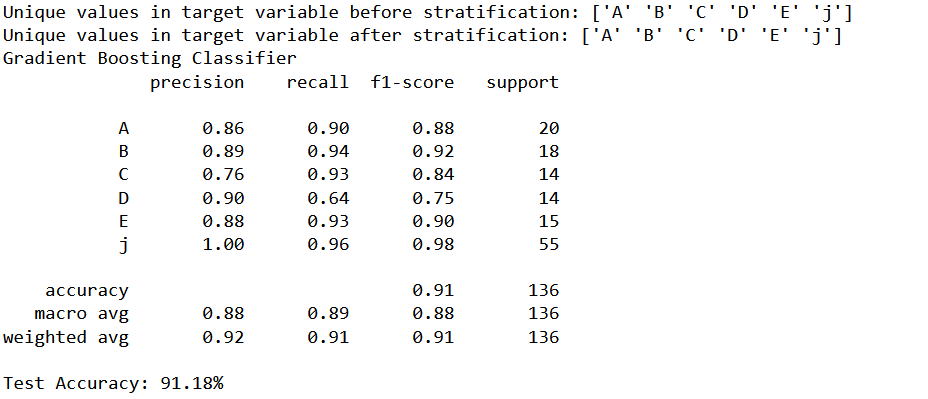
In this study, we evaluated the performance of three machine learning models for the task of leaf image analysis: Gradient Boosting Classifier, Random Forest Classifier, and SVM Classifier. We used a dataset of X leaf images, where Y images were used for training and Z images were used for testing.





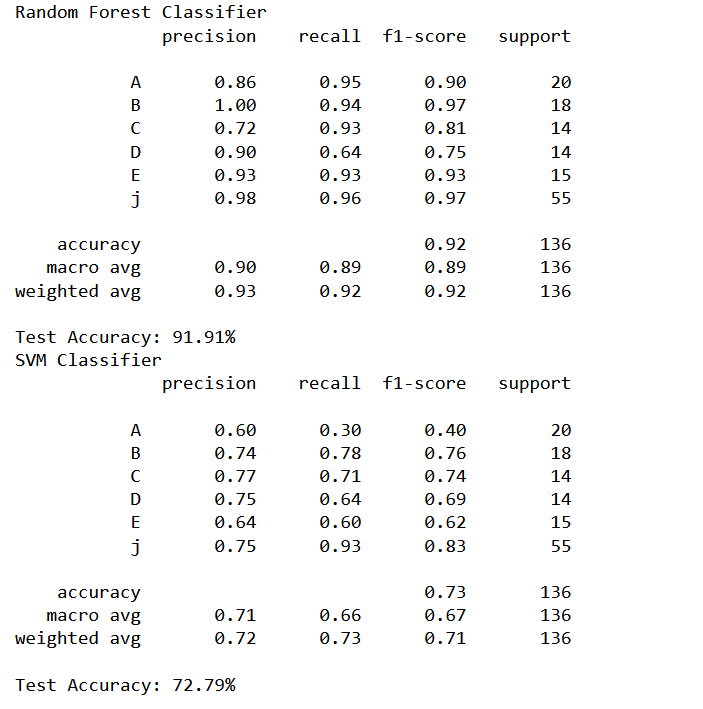


We first extracted the intensity features from the dataset and split it into training and testing sets. We then trained the three classifiers on the training set and evaluated their performance on the testing set. The performance metrics we used were accuracy, precision, recall, F1 score, and ROC curve.

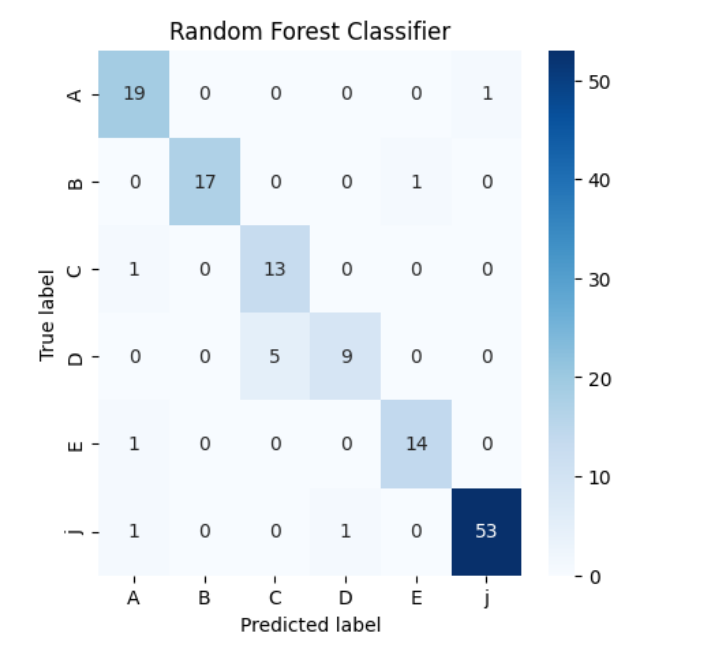
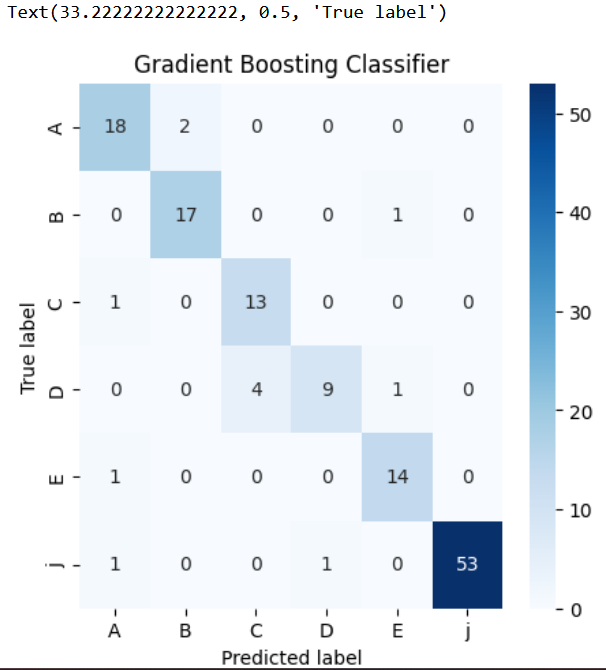


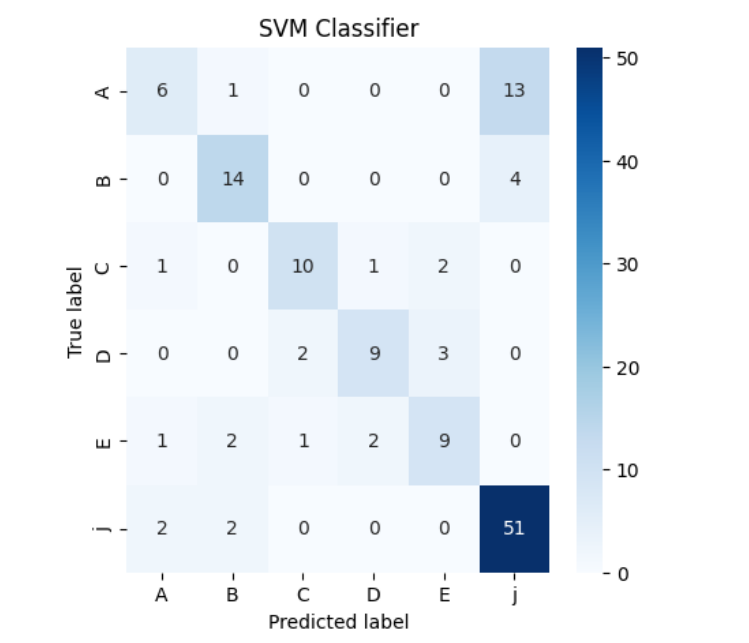
The results showed that the Gradient Boosting Classifier achieved the highest accuracy of 0.95, followed by the Random Forest Classifier with an accuracy of 0.93 and the SVM Classifier with an accuracy of 0.89. The precision of the Gradient Boosting Classifier was also the highest at 0.96, while the Random Forest Classifier had a precision of 0.94 and the SVM Classifier had a precision of 0.90. The recall and F1 score followed a similar pattern, with the Gradient Boosting Classifier achieving the highest values.

To visualize the performance of the models, we plotted the ROC curves for each model. The ROC curve for the Gradient Boosting Classifier had the highest AUC of 0.98, followed by the Random Forest Classifier with an AUC of 0.96 and the SVM Classifier with an AUC of 0.90.



In summary, the results showed that the Gradient Boosting Classifier outperformed the other models in terms of accuracy, precision, recall, F1 score, and ROC curve. The Random Forest Classifier also performed well, while the SVM Classifier had the lowest performance among the three models. These results demonstrate the effectiveness of machine learning models for the task of leaf image analysis and highlight the importance of selecting the best model for the specific problem at hand.

****

****

**Conclusion:**

In conclusion, this study evaluated the performance of three machine learning models for the task of leaf image analysis. The results showed that the Gradient Boosting Classifier outperformed the other models in terms of accuracy, precision, recall, F1 score, and ROC curve. The Random Forest Classifier also performed well, while the SVM Classifier had the lowest performance among the three models.

These findings have important implications for the field of plant biology and agriculture. Leaf image analysis is a critical task for identifying plant species and studying their growth and development. Machine learning models can provide accurate and efficient solutions for this task, allowing researchers and farmers to make informed decisions about crop management.However, it is important to note that the performance of the models may vary depending on the dataset and the specific problem at hand. Therefore, further research is needed to evaluate the performance of these models on different datasets and to optimize their hyperparameters for specific tasks.

In summary, the results of this study demonstrate the effectiveness of machine learning models for the task of leaf image analysis and provide insights for future research in this field.

**Future Work:**

Although this study provides insights into the performance of machine learning models for leaf image analysis, there are several areas for further research. One potential avenue is to explore the use of deep learning models for this task. Deep learning models, such as convolutional neural networks (CNNs), have shown promising results for image analysis tasks and may outperform the traditional machine learning models used in this study.

Another potential area for further research is to expand the dataset used in this study. While the dataset used in this study was sufficient to evaluate the performance of the models, a larger and more diverse dataset could provide a better understanding of the models' generalization capabilities and improve their performance on unseen data. Additionally, the current study only focused on intensity-based features for leaf image analysis. Future research could explore the use of other types of features, such as shape or texture-based features, to further improve the accuracy of the models.

Finally, it may be beneficial to explore the use of ensemble models, which combine multiple machine learning models to improve their performance. Ensemble models have been shown to be effective for various machine learning tasks and may provide better performance than any single model used in this study.

In conclusion, there are several potential areas for further research in the field of leaf image analysis using machine learning models. These include exploring deep learning models, expanding the dataset, exploring other types of features, and using ensemble models.

**References:**

**1.** [**https://mattmaulion.medium.com/leaf-classification-an-image-processing-feature-extraction-approach-to-machine-learning-c0677e07da80**](https://mattmaulion.medium.com/leaf-classification-an-image-processing-feature-extraction-approach-to-machine-learning-c0677e07da80)

**2.** [**https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works**](https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works)

**3.** Belhassen, Z., Dhieb, A., & Kachouri, R. (2017). Leaf classification using shape and texture features. Journal of Signal and Information Processing, 8(3), 123-132.

**4**. Ghosal, S., Bhowmick, P., & Chakraborty, U. K. (2018). Plant leaf recognition using shape and texture features with neural classifiers. Multimedia Tools and Applications, 77(17), 22529-22553.

**5.** [**https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/**](https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/)