Hybrid Models for Classification of Healthy and Diseased Crops Using Machine Learning

Riyan Gupta(22116079)
Piridi Bhanuja(22116066)
Puramsetti Sushma(22116076)
Vislavath Himabindu(22116107)
Department of Electronics and Communication Engineering
Indian Institute of Technology Roorkee

***Abstract*—In this report we are using the hybrid models in order to classfy the plant and crop images as either healthy or diseased. We will analyze the performance of these models and compare them with respect to how accurately are they able to classify with the use of confusion matrix, classification reports and so on. Our concern is to find the best model so that it can be applicable for real time detection of diseased crops.**

I. Introduction

In order to make sure that the plants and crops can be analyzed and hence treated at a large scale, we cannot just simply rely on manual methods. Instead we need to use sophisticated models that have high computaional power and are also accurate. Hybrid models are an important class of such models in which we are using various models with different functionalities in order to pool all these functionalities in a single model which can be operated on the data.

II. Dataset and Preprocessing

The dataset is taken from Kaggle and is named as CCMT Final Dataset. Although the dataset has more than 10000 images, we have taken the total no of images as 1590 across all the four classes which include cashew, cassava, maize and tomato.This was done in order to avoid the problem of memory allocation.These images are resized to 256 × 256 pixels and normalized to improve model training.

III. Feature Extraction

In feature extraction, we are extracting the histogram data and analyzing the GLCM(Gray Level Co Occurence Matrix) features which can include the contrast, correlation, energy and homogenity.Fig. 2 shows the feature vector which clubs all of these properties.Feature extraction is very crucial so that these features can be subsequently used in the training of the model.

IV. Hybrid Models

Four hybrid models are implemented:

1) **Voting Classifier:** The classifiers involved are Random Forest, Gradient Boosting and Extra Trees classifiers.
2) **PCA + RFE:** It is a combination of Principal Component Analysis and Recursive Feature Elimination. PCA is used to reduce the dimensions and the RFE is used for feature selection.
3) **Augmented Features + LightGBM:** Incorporates additional statistical features like mean, variance, skewness and kurtosis and employs LightGBM or Light Gradient Boosting Machine.

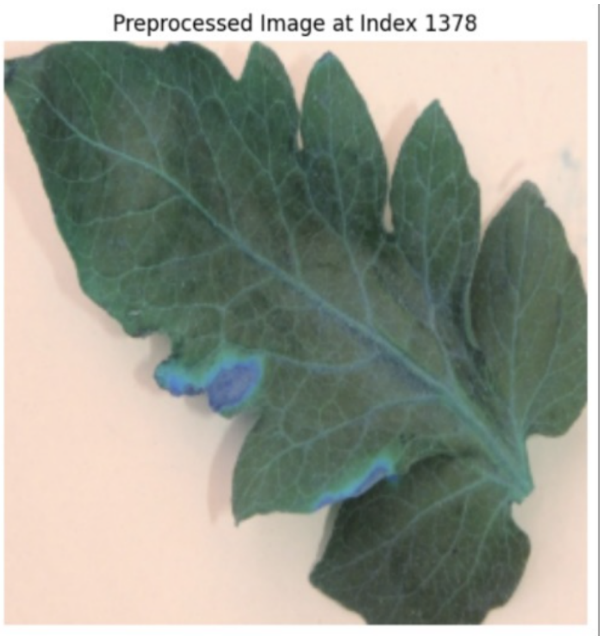
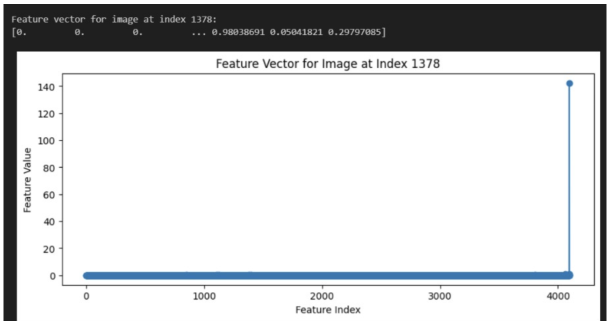


Fig. 1: Preprocessed Image Example



```
Feature vector for image at index 1378:
[0.         0.         0.         ... 0.98038691 0.05041821 0.29797085]
```

Fig. 2: Feature Vector for a Sample Image

4) **Clustering + Classification:** The clustering is done by K means and classification is done with the help of the Random Forest.

V. Results and Discussion

A. *Model Evaluation*

The model is trained with an 80 percent of training data and 20 percent of test data. The models were performing in such a manner that the test accuracies were coming out to be very very high, so it reduces any possibility of overfitting or underfitting because these problems happen when the test data has a very less accuracy. This is a major benefit of hybrid models.The metrics of model performance are going to include the confusion matrix, which will help to analyze the accuracy of the model in greater details and the classification report which includes the values of precision, recall, f1-score and so on.The best part of all these models is that all of these models are classifying the plants at a very high accuracy which makes these models very very useful in real life applications where accuracy of classification is the key.

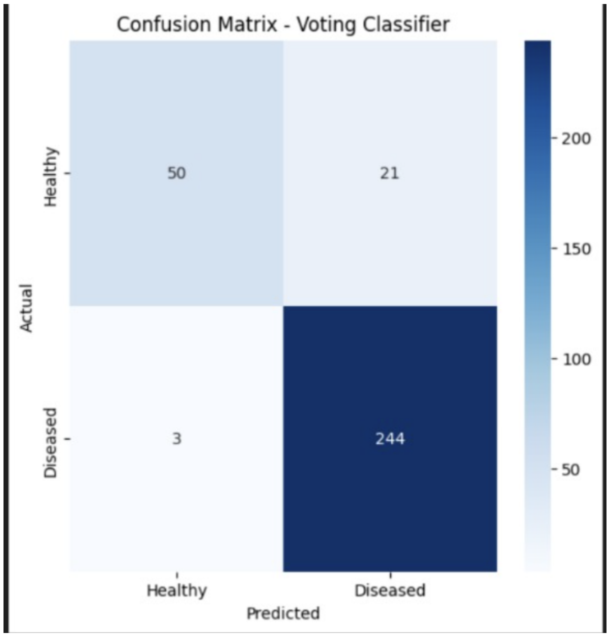
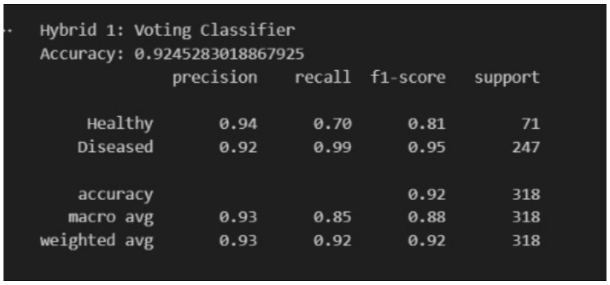


Fig. 3: Confusion Matrix - Voting Classifier



```
Hybrid 1: Voting Classifier
Accuracy: 0.9245283018867925
              precision    recall  f1-score   support

     Healthy       0.94      0.70      0.81        71
    Diseased       0.92      0.99      0.95       247

    accuracy                           0.92       318
   macro avg       0.93      0.85      0.88       318
weighted avg       0.93      0.92      0.92       318
```

Fig. 4: Classification Report - Voting Classifier

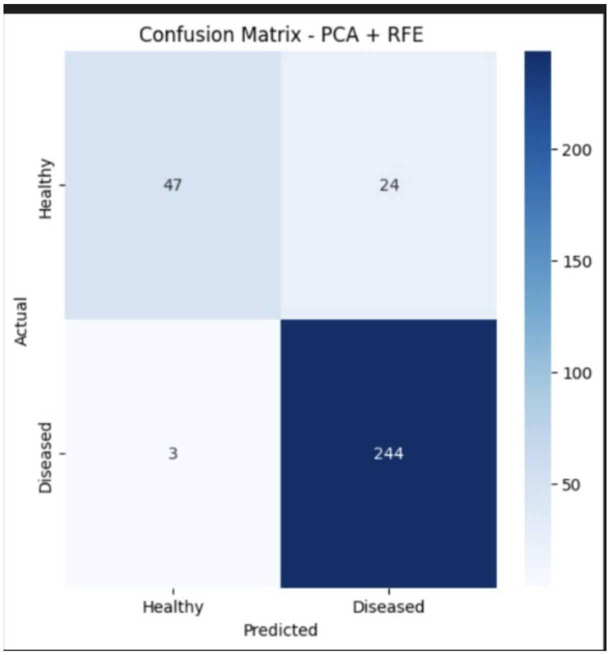
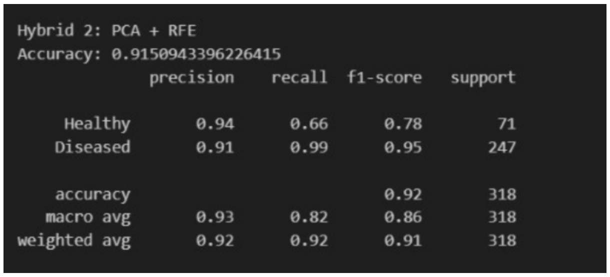


Fig. 5: Confusion Matrix - PCA + RFE



```
Hybrid 2: PCA + RFE
Accuracy: 0.9150943396226415
              precision    recall  f1-score   support

     Healthy       0.94      0.66      0.78        71
    Diseased       0.91      0.99      0.95       247

    accuracy                           0.92       318
   macro avg       0.93      0.82      0.86       318
weighted avg       0.92      0.92      0.91       318
```

Fig. 6: Classification Report - PCA + RFE

VI. Conclusion

The use of hybrid models as a concept and as a useful tool to improve the classification of data will surely be a great area of research because these hybrid models have a great accuracy of classification and have a wide range of applications apart from just classification. They have performed very well with an astonishing accuracy, especially the Augmented Features and LightGBM one which has an accuracy of almost 97 percent. Every model has achieved classification with an accuracy of greater than 90 percent. So in this report we have analyzed the performance of four hybrid models in classification of plants as diseased or healthy and concluded that they performed really well, even better than singular models like SVM or LR.

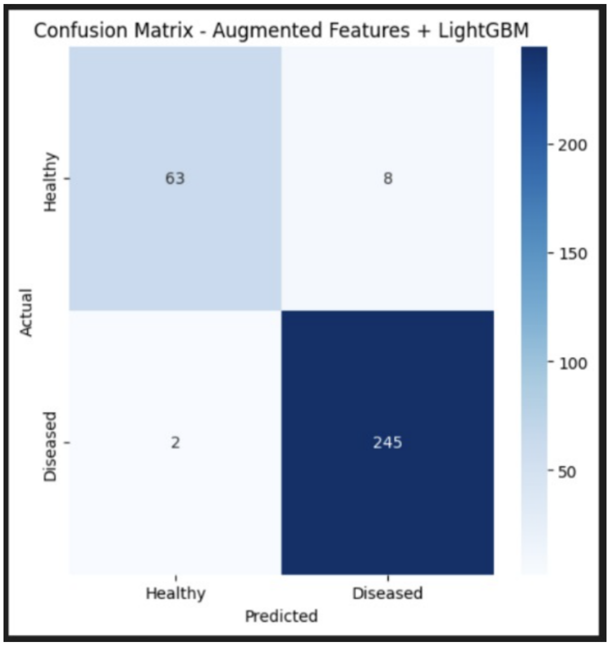
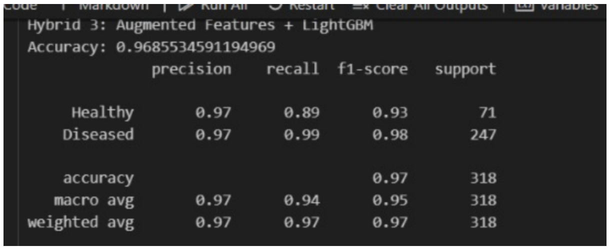


Fig. 7: Confusion Matrix - Augmented Features + LightGBM



```
Hybrid 3: Augmented Features + LightGBM
Accuracy: 0.9685534591194969
              precision    recall  f1-score   support

     Healthy       0.97      0.89      0.93        71
    Diseased       0.97      0.99      0.98       247

    accuracy                           0.97       318
   macro avg       0.97      0.94      0.95       318
weighted avg       0.97      0.97      0.97       318
```

Fig. 8: Classification Report - Augmented Features + LightGBM

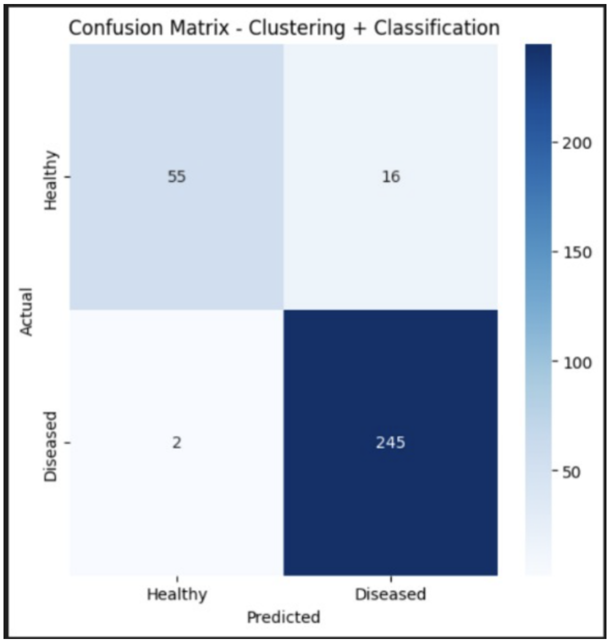
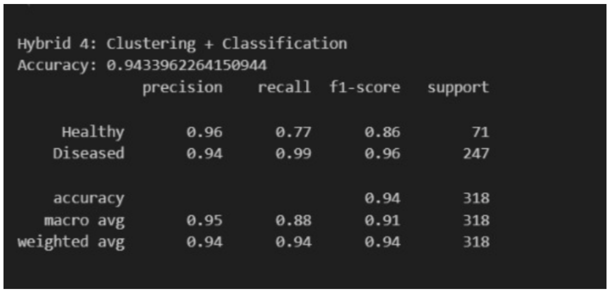


Fig. 9: Confusion Matrix - Clustering + Classification



```
Hybrid 4: Clustering + Classification
Accuracy: 0.9433962264150944
              precision    recall  f1-score   support

     Healthy       0.96      0.77      0.86        71
    Diseased       0.94      0.99      0.96       247

    accuracy                           0.94       318
   macro avg       0.95      0.88      0.91       318
weighted avg       0.94      0.94      0.94       318
```

Fig. 10: Classification Report - Clustering + Classification