IT1244 Project Report: Breast Cancer Dataset

Team 25

Lee Yan Le Ryan - A0253123U

Lowe David Robert - A0289965Y

Mefsut Nicholas - A0290168Y

Nevin Allen Marian - A0269753M

Introduction

Breast cancer is the most common type of cancer affecting women in the world, as it is one of the leading causes of female mortality, especially in developing countries. The chances of having breast cancer increases with age, and the early detection of tumours in breast tissue is crucial for having a higher probability of an individual recovering with successful treatment.

Our project aims to effectively diagnose the type of tumours found in a breast mass, benign (not harmful) and malignant (infectious and indicative of breast cancer) using basic methods of machine learning that were taught in this course, as well as some intermediate methods that we researched outside of the syllabus.

There are several challenges in diagnosing breast cancer in women with current detection methods such as mammograms. It overlooks certain cases of cancer, especially in women with dense tissue mass. With the use of machine learning, breast cancer detection can be improved by analysing the attributes in imaging data to identify suspicious or abnormal changes in the breast tissue and deploying different models to evaluate its performance and choose the best-performing model in distinguishing malignant from benign tumours.

Dataset

The dataset given was based on the attributes of images of a fine needle aspirate of the breast mass. The attributes describe the characteristics of the cell nuclei present in the image. The first column contained the diagnosis based on the cell, benign or malignant, labelled as ‘B’ and ‘M. The other 30 columns can be segregated into 3 sets of statistics: the mean, standard error, and the worst measurements. Each set consists of 10 measurements of the cell, including the radius, texture, smoothness, concavity, etc.

We converted the labels in diagnosis from ‘B’ and ‘M’ to 0 and 1 respectively as binary target variables are easier to work with. Since there are many feature variables, we used a correlation matrix to select the measurements in each set of statistics. We then identified variables that have high correlation with the diagnosis.

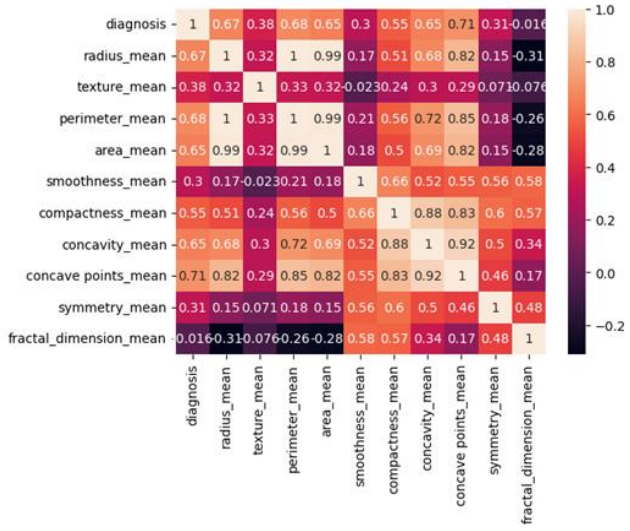

Figure 1: Correlation Matrix of the “mean” variables.

After computing three different correlation matrices for each statistic; mean, standard error and worst, we picked the features that have a correlation of 0.6 and above with the diagnosis, and ignored area and perimeter as they had perfect correlation with radius.

- radius_mean
- concavity_mean
- concave points_mean
- radius_worst
- concavity_worst
- concave_points worst

Methods

Mislabel Detection

Building upon the features selected from the correlation analysis, we aimed to explore the potential separation of data points into two distinct clusters, one for benign, and another for malignant. To achieve this, we visualised all possible combinations of the selected features, plotting each pair against one another and colour-coding the points based on their diagnosis. However, the resulting clusters exhibited considerable overlap and lacked clear discernible patterns, prompting us to consider alternative approaches.

Feature Engineering

Considering these observations, we turned to feature engineering to uncover new insights within the dataset. We learnt from online research the distinct differences between a benign tumour and a malignant tumour. Most notably, a benign tumour is smoother along its borders as compared to a malignant tumour which has variations and spikes along its borders. So, there may be patterns in the difference between worst and mean values.

The features that were selected from correlation analysis are highly related to the smoothness of its border. We performed feature scaling by standardising the selected features, then calculated the cubed and squared differences between worst and mean values to reveal potential patterns. The resulting patterns across the three combinations of features remained consistent, with points of malignant diagnosis following an upward curve and points of benign diagnosis following a downward curve.

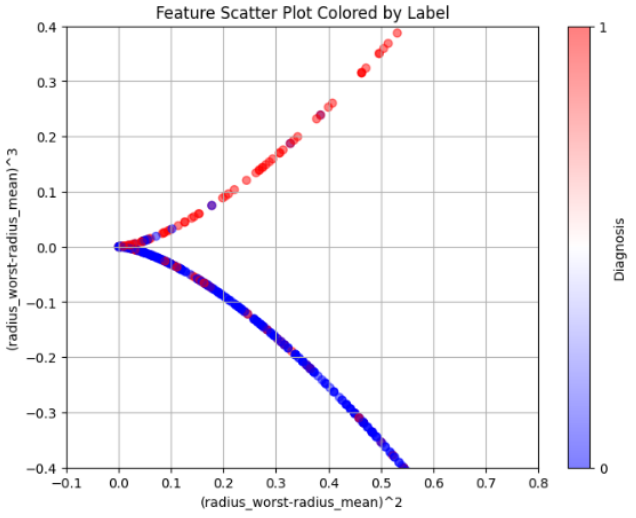

Figure 2: The zoomed in scatter plot of cubed difference against squared difference of radius worst and radius mean.

With these insights in mind, we proceeded to identify outliers within the dataset by pinpointing benign (blue) points on the upward curve and malignant (red) points on the downward curve. We then compared the indices of identified outliers against those provided in the given file.

Indices derived from the feature engineered cubed difference of radius worst and radius mean exhibited the best matching-to-predicted ratio. We then saved the predicted indices and the modified dataset for further investigation and analysis.

K-means Clustering Approach

In our continued exploration of the dataset, we employed K-means clustering to investigate the potential separation of benign and malignant points into two distinct clusters. Recognizing the sensitivity of K-means to initialization, we conducted up to 50 iterations, varying the random state parameter to explore the results obtained from varying centroid initializations.

For each of the 50 iterations, we evaluated the clustering results by identifying indices where false positives and false negatives occurred, as some of them could be the mislabelled indices we were searching for. We retained only the indices that were consistently predicted across all 50 iterations by using dictionaries to keep track of the frequencies, ensuring robustness in our selection process.

We then took the intersection of the two arrays of predicted indices from feature engineering and k-means clustering to obtain a new array of predicted indices, to further refine our predictions.

Logistic Regression Model Application

We employed a logistic regression model on the dataset to refine our predictions further. Segregating the testing data based on the predicted mislabelled indices allowed us to evaluate the model's performance specifically on potentially mislabelled instances. The resultant confusion matrix highlighted a significant proportion of false positives and false negatives, indicative of the presence of mislabelling within the testing data.

Iterating this process enabled us to progressively narrow down the predicted mislabelled indices while maintaining consistency with the actual mislabelled data. By repeating the validation steps, we refined our predictions, ultimately achieving a high degree of correspondence between predicted and actual mislabelled instances.

Classification

Our study's focus is on accurately classifying breast cancer samples into benign (B) and malignant (M) categories. Planning for an early diagnosis and course of therapy depends on this work. Using characteristics extracted from fine needle aspirate (FNA) pictures of breast masses, we employed the use of various machine learning models that are well-known for binary classification.

Logistic Regression

The standard method for binary classification tasks is Logistic Regression (LR), which outputs a probability value between 0 and 1 using the sigmoid function. It was chosen for its efficiency, and it is less prone to overfitting.

k-Nearest Neighbours

The non-parametric technique k-Nearest Neighbours (kNN) was chosen due to its simplicity and intuitiveness. It also does not make any assumptions about the data and is effective with small datasets, indicating its potential use for our breast cancer dataset which has 569 samples.

Random Forest (not covered in IT1244)

One of the intermediate models we utilised was Random Forest (RF). During the training phase, the RF ensemble learning technique generates a forest of decision trees, each one derived from a random subset of the training data (Breiman, 2001).

For tasks involving classification or regression, this approach yields the mean prediction for all trees or the mode of the classes. RF tries to generate a diverse set of trees by adding randomness in the selection of data and feature subsets for each tree's construction. This increases overall accuracy and lowers the possibility of overfitting.

As it aggregates a variety of trees, RF is a reliable and adaptable algorithm that may be applied to a broad range of machine learning problems (Breiman, 2001). The selection of RF was based on its high-dimensional data handling capabilities and performance.

Support Vector Machine (not covered in IT1244)

The final method used was Support Vector Machine (SVM). Strongly suited for both classification and regression applications, SVM is a supervised learning technique. According to Cortes and Vapnik (1995), it finds the best hyperplane to divide the classes in the feature space by maximising the gap between the classes' closest points, or support vectors.

Due to its effectiveness in high-dimensional spaces and its ability to handle non-linear relationships using the kernel technique, SVM can convert the input space into higher dimensions where a linear separator can be implemented. According to (Cortes and Vapnik, 1995), SVM's versatility renders it appropriate for intricate classification applications where classic linear classifiers are inadequate.

Cross-Fold Validation

To guarantee the validity of performance indicators and their indication of the models' capacity to generalise to new data, a 10-fold cross-validation approach was employed for each model's evaluation. Due to the limited size of our data, performing 10-fold cross validation makes efficient use of the given data we have, while mitigating overfitting.

Our method includes advancements meant to get over drawbacks found in earlier studies, like handling possible class imbalance and guaranteeing reliable model evaluation. This was achieved with the implementation of the 10-fold cross-validation, as well as making use of multi-metric evaluation.

Evaluation Metrics

We used three metrics to evaluate the models.

- Accuracy ($\text{ACC} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$)
- Receiver Operating Characteristic Area Under the Curve (ROC = Area Under Curve of TPR against FPR)
- F1 score ($\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$)

Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied, used to assess the trade-off between the true positive rate and false positive rate (Fawcett, 2006). The F1 score is the harmonic mean of precision and recall. Precision is the ratio of true positives to all positive predictions, and recall is the ratio of true positives to all actual positives. These metrics were all used to find the most effective model in successfully predicting the diagnosis for breast cancer.

Results and Discussion

Mislabel Detection

Results

After every step, we compared the indices that we predicted using our method to the actual swapped indices to find the number of indices that were matching.

	Matching	Predicted
Feature Engineering	26	76
k-Means Clustering	24	70
Intersection	24	37
Logistic Regression	24	28

Table 1: Our methods and their matching-to-predicted ratio

Discussion

We believe radius to be the most influential feature. In the feature engineering method, the feature that provided the best matching-to-predicted ratio was $(\text{radius}_{\text{worst}} - \text{radius}_{\text{mean}})^3$. This likely was the case as after standardizing the differences, the smaller differences would become negative, and the larger differences would become positive. Since benign tumours are smoother along its border, it would have a smaller difference. Since malignant tumours have lots of spikes around its border, it would have a larger difference. Thus, when cubing the difference, there is a clear distinction between the two diagnoses. In hindsight, there was no need to cube the difference, as the two diagnoses would already have distinguished itself by splitting into the negative and positive, but cubing may have also made the pattern clearer.

Instead of performing k-Means clustering and taking the intersection of the two predicted indices obtained, we could probably also have tried instead to directly employ the use of logistic regression, using the 76 predicted indices as testing data and all other indices as training data. This may have simplified and sped up the process of detection, and

likely preserve the initial 26 matching indices obtained from feature engineering.

A recurring issue with the methods we employed was the significantly high number of extra predicted indices. This is likely indicative of a general issue with the data quality, or that the models we utilised are too simple to predict something as complex as breast cancer. Thus, this led to us researching and utilising more complex models such as Random Forest and Support Vector Machine for the classification task.

Modelling

Results

After running all models previously mentioned, they all performed well, achieving values of greater than 0.93 in all metrics.

	ACC	F1	ROC
LR	0.974	0.963	0.968
kNN	0.956	0.935	0.942
RF	0.965	0.952	0.960
SVM	0.979	0.971	0.975

Table 2: Metrics used to evaluate the models.

SVM clearly outperforms other methods across all metrics and was thus chosen as the optimal method to predict breast cancer in patients.

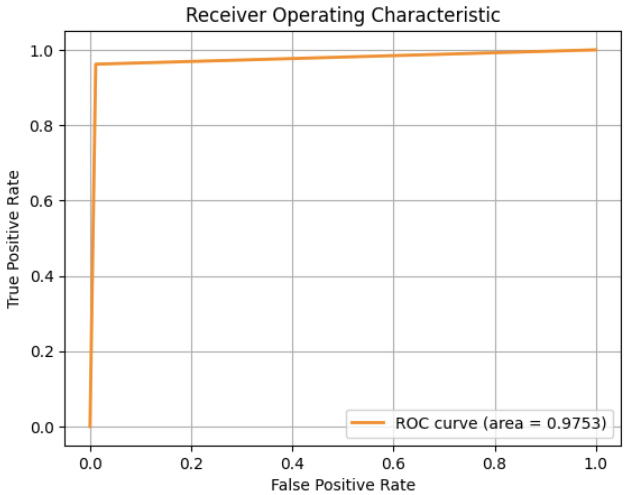

Figure 3: ROC for SVM model.

Discussion

The high accuracy indicates that SVM is overall the most effective in classifying instances correctly across both classes. The high F1 score indicates that SVM is most effective in class imbalance and is robust. The high AUC indicates that SVM is most optimal at distinguishing between the positive and negative classes. There could be a few reasons for the effectiveness of the SVM model.

- It is effective in high-dimensional spaces, which is applicable to the breast cancer dataset as it has a substantial number of features. Breast cancer is also a complex topic, and a linear model would likely have subpar performance.
- SVM can handle non-linear input spaces effectively by using the kernel trick, creating a non-linear boundary between benign and malignant tumours.
- SVM can find extremely difficult points to classify, which is of immense use in a complex classification.
- SVM's formulation also allows for cost-sensitive learning, which can be crucial for the case of breast cancer diagnosis where the cost incurred from a false negative is much greater than the cost incurred from a false positive.

Interestingly, despite being suited for linear data, LR performed exceptionally well, coming in at second place across all metrics. This may suggest that the decision boundary between benign and malignant may not be as complex as initially thought, only slightly edging towards non-linear.

Unsurprisingly, kNN performed the worst out of the four models. It suffers from the 'curse of dimensionality', where it becomes less effective in datasets with a substantial number of features. In hindsight, we should have attempted to reduce the dimensionality of the data, either through feature selection by correlation matrix as done earlier, or feature extraction through Principal Component Analysis to perhaps bring out the potential in kNN.

There are some features highlighted in the correlation matrix that do not have a high correlation with the diagnosis, and some with absolutely zero correlation. This perhaps contributed to the reduced effectiveness of RF, as it likely suffered from overfitting due to the inclusion of irrelevant features. It can also be greatly affected by noise in the data, leading to complex trees that do not generalize well.

Nevertheless, we should take the effectiveness of SVM with a pinch of salt. In fact, changing the seed for the random state would sometimes lead to LR outperforming every model. Additionally, for much larger datasets, LR is more practical due to its computational efficiency as compared to SVM. Training the model with LR is generally faster as the Gradient Descent technique can converge quickly with large datasets. This could mean the difference between life and death as there may be dire cases where a prediction is urgently required. In hindsight, we also should have performed hyperparameter tuning to fine-tune SVM.

Overall, this work overcomes the limitations of earlier research by offering a strong framework for comparing models that can be used for comparable binary classification tasks and by utilising an extensive range of evaluation metrics to make sure the selected model is performing well in more areas than just overall accuracy.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.

Appendix

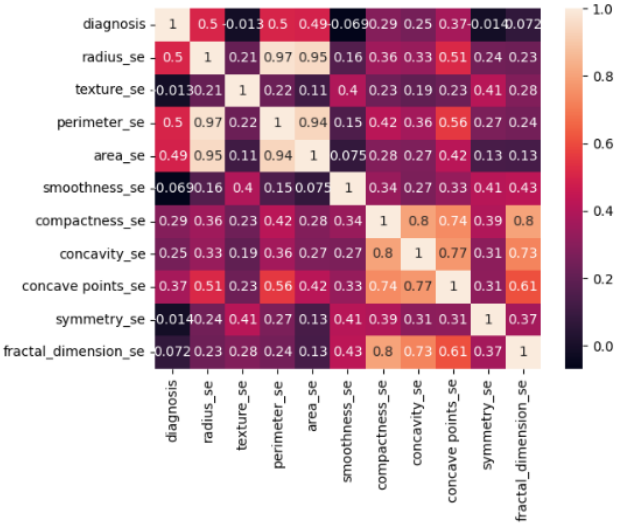

Figure 4: Correlation Matrix of the "se" variables.

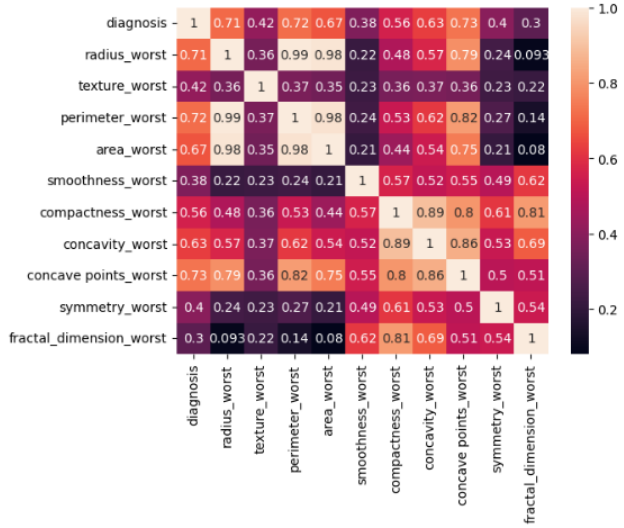

Figure 5: Correlation Matrix of the "worst" variables.

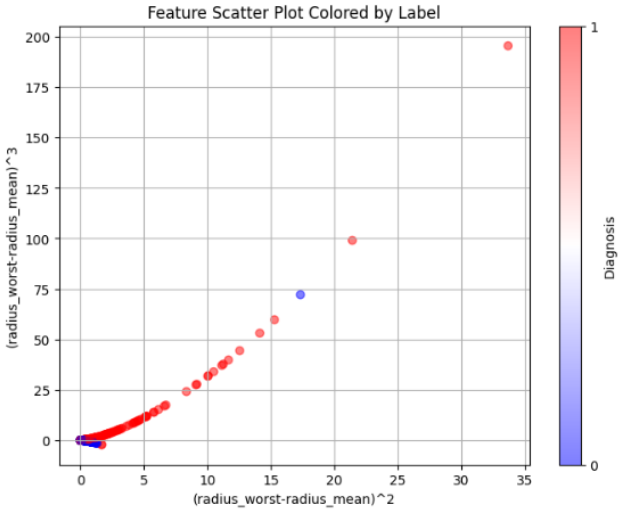

Figure 6: $(\text{radius}_{\text{worst}} - \text{radius}_{\text{mean}})^3$ against $(\text{radius}_{\text{worst}} - \text{radius}_{\text{mean}})^2$.

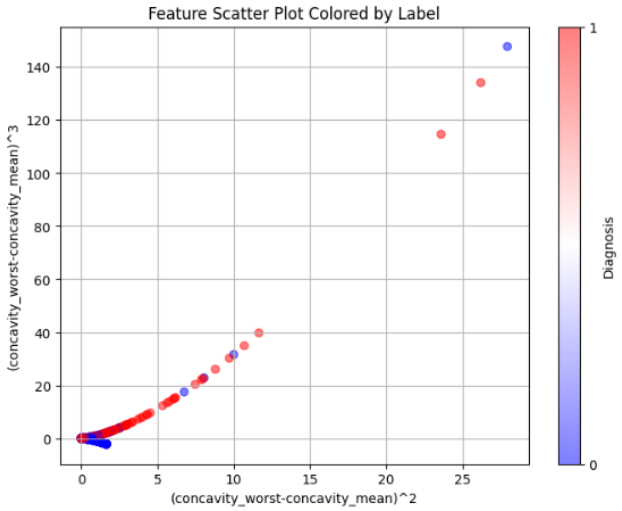

Figure 7: $(\text{concavity}_{\text{worst}} - \text{concavity}_{\text{mean}})^3$ against $(\text{concavity}_{\text{worst}} - \text{concavity}_{\text{mean}})^2$.

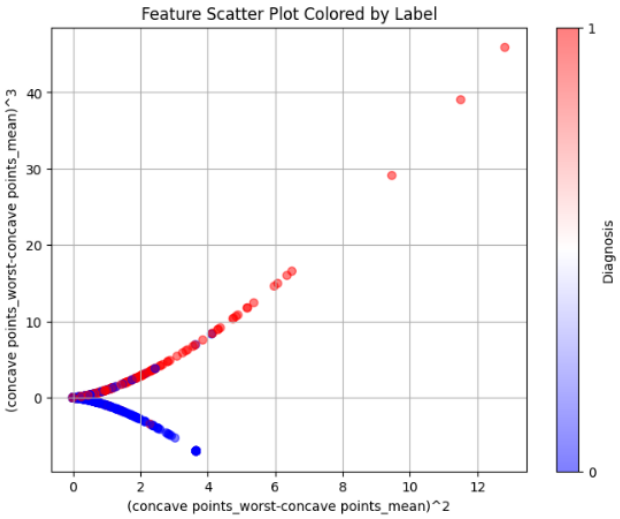

Figure 8: $(\text{concave points}_{\text{worst}} - \text{concave points}_{\text{mean}})^3$ against $(\text{concave points}_{\text{worst}} - \text{concave points}_{\text{mean}})^2$.

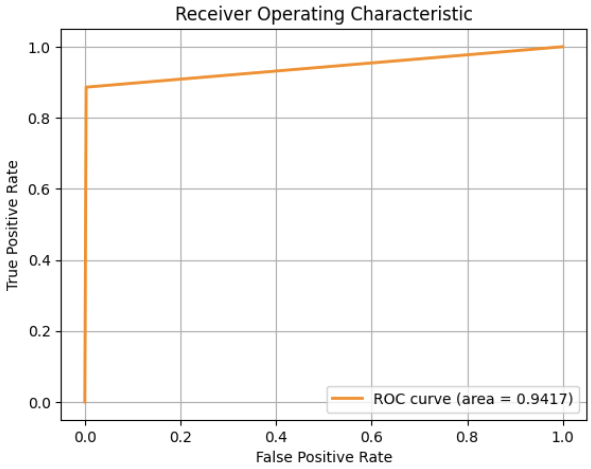

Figure 9: ROC for kNN.

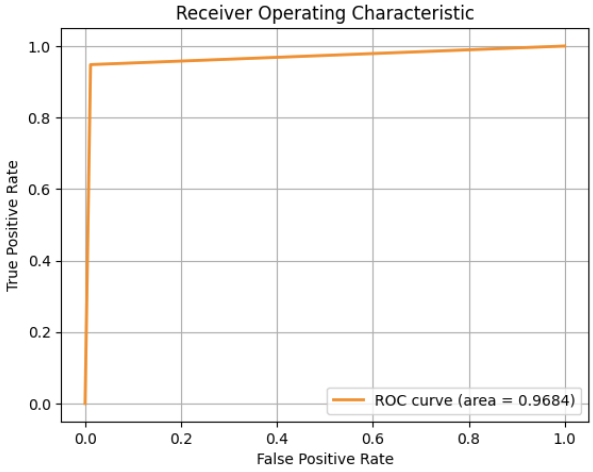

Figure 10: ROC for Logistic Regression.

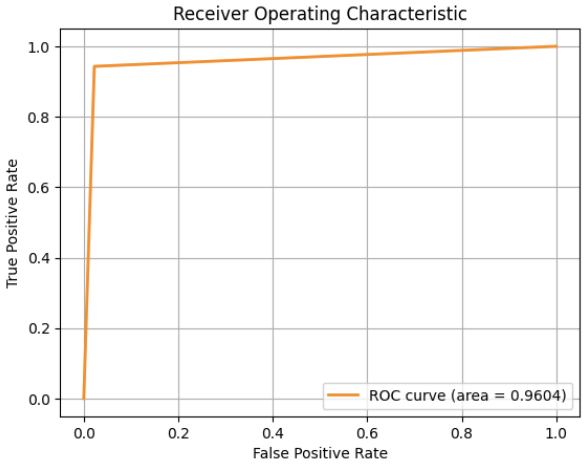

Figure 11: ROC for Random Forest.