COEN 432 Assignment 2 Report

Kirollos Georgy - 40190279

Abdelaziz Mekkaoui - 40192247

Model and Data Handling

In this project, we used the random forest classifier model for detecting the patients who have cancer. Random forest is basically a collection of decision trees, so it will be able to generalize as it picks the majority vote.Since this is a complex problem with high-dimensional data, we chose this model because of its robustness and ability to handle non-linear data.

To handle missing data, we used the KNN imputer. This imputer replaces missing values with the mean value of the k-nearest neighbors. This approach helps to maintain the integrity of the dataset and ensures that the model has complete data for training and testing.

We also used a grid search to tune the hyperparameters of the random forest classifier to get the best possible parameters for the model. The grid search was performed using 10-fold cross-validation to evaluate the performance of the model on different subsets of the data. The evaluation metrics used in the grid search were accuracy, precision, recall, F1 score, and ROC AUC score. The best parameters found by the grid search were used to train the subsequent models with different training data sizes.

Five different training set sizes were used to evaluate the performance of the model on different amounts of data. The model was trained on 40, 140, 240, 340, and 440 samples, and tested on 10, 35, 60, 85, and 110 respectively, keeping a ratio of 4 : 1 between the training and testing sets. This allowed us to assess how the model performs with varying amounts of training data.

Different Training Set Sizes

For train size = 40, test size = 10:

Prediction Time: 4.9877 milliseconds

Accuracy: 100.0%

Precision: 100.0%

Recall: 100.0%

F1 Score: 100.0%

ROC-AUC: 100.0%

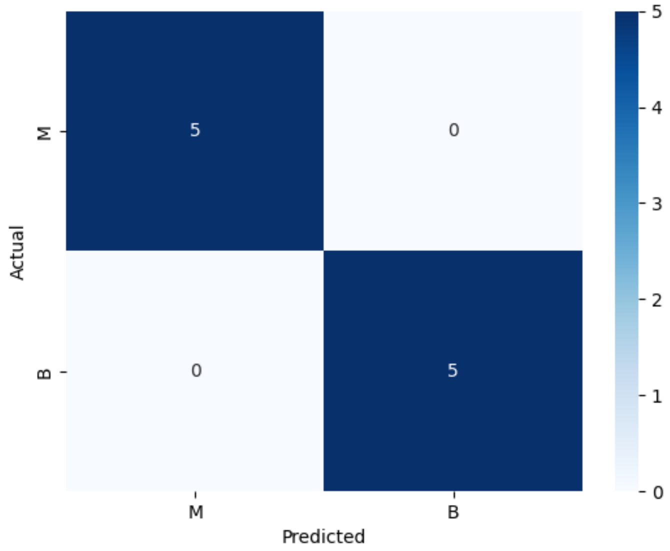

Figure 1: Confusion Matrix

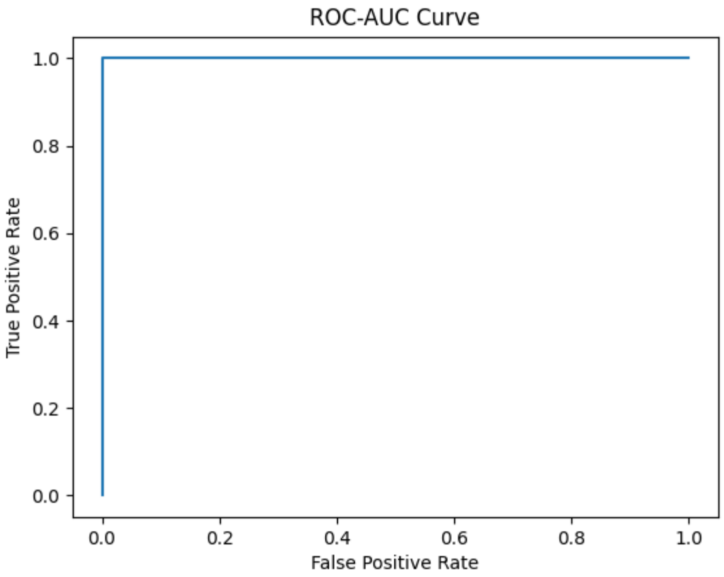

Figure 2: ROC-AUC Curve

For train size = 140, test size = 35:

Prediction Time: 4.9837 milliseconds

Accuracy: 97.14%

Precision: 100.0%

Recall: 91.67%

F1 Score: 95.65%

ROC-AUC: 99.64%

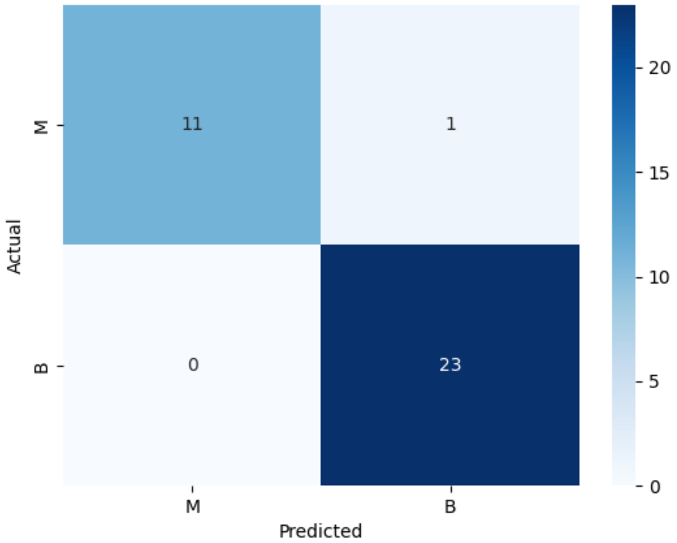

Figure 3: Confusion Matrix

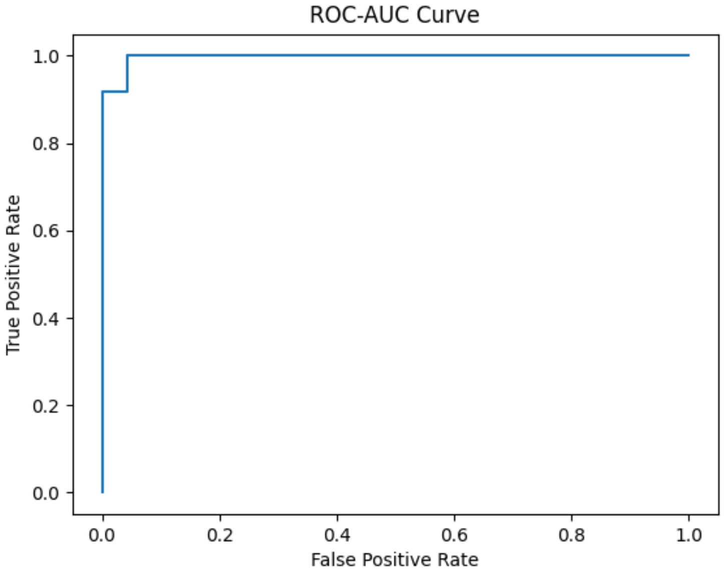

Figure 4: ROC-AUC Curve

For train size = 240, test size = 60:

Prediction Time: 3.9918 milliseconds

Accuracy: 96.67%

Precision: 94.74%

Recall: 94.74%

F1 Score: 94.74%

ROC-AUC: 99.74%

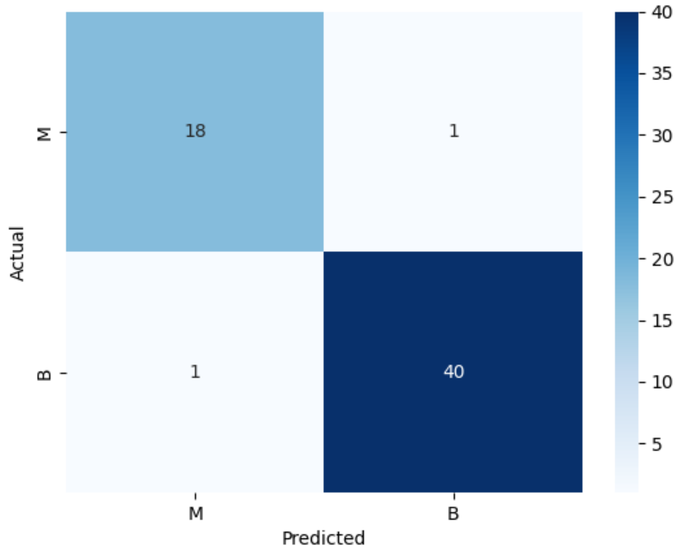

Figure 5: Confusion Matrix

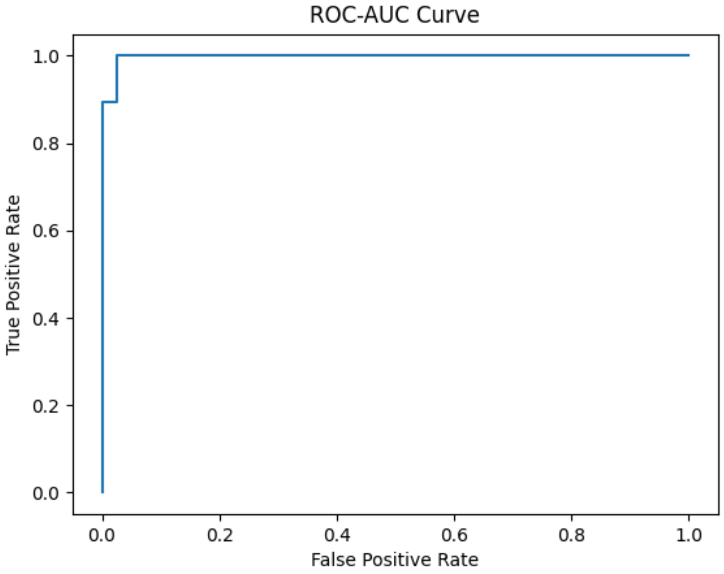

Figure 6: ROC-AUC Curve

For train size = 340, test size = 85:

Prediction Time: 4.9891 milliseconds

Accuracy: 95.29%

Precision: 96.55%

Recall: 90.32%

F1 Score: 93.33%

ROC-AUC: 99.16%

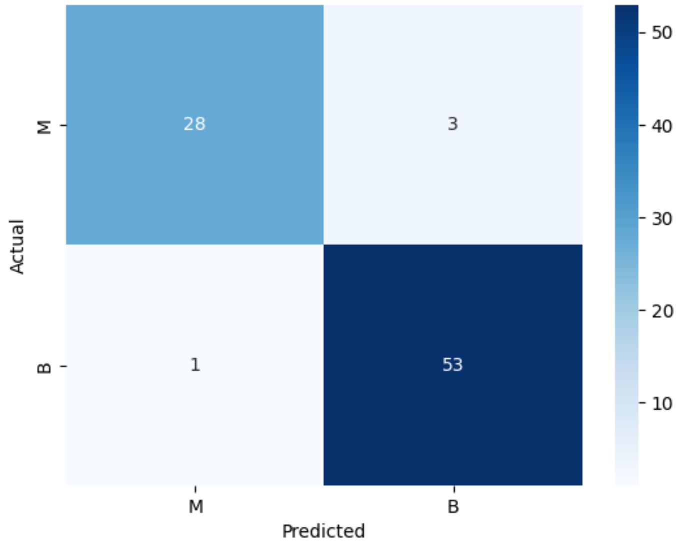

Figure 7: Confusion Matrix

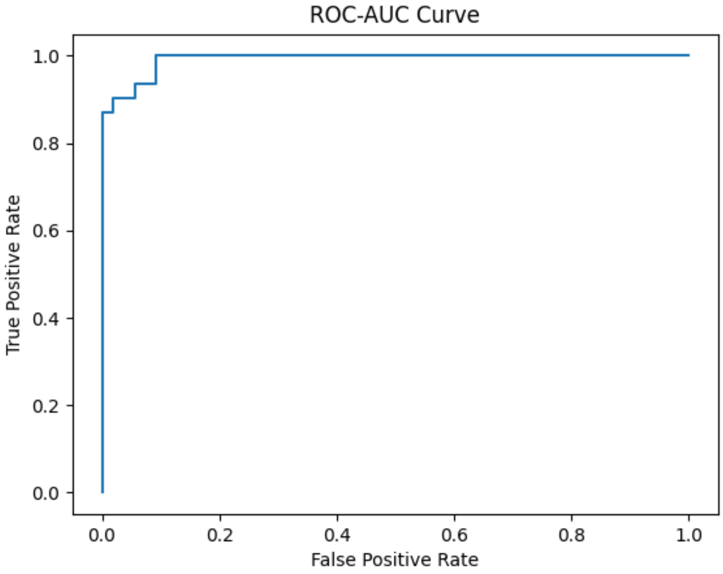

Figure 8: ROC-AUC Curve

For train size = 440, test size = 110:

Prediction Time: 3.9892 milliseconds

Accuracy: 96.36%

Precision: 97.44%

Recall: 92.68%

F1 Score: 95.0%

ROC-AUC: 99.51%

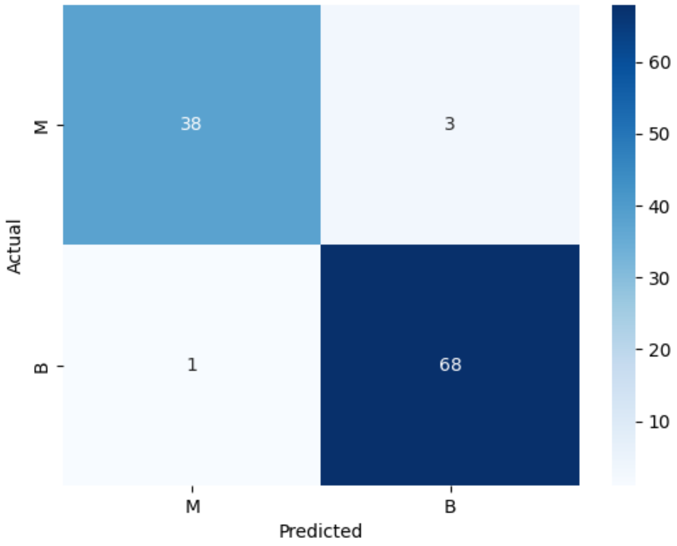

Figure 9: Confusion Matrix

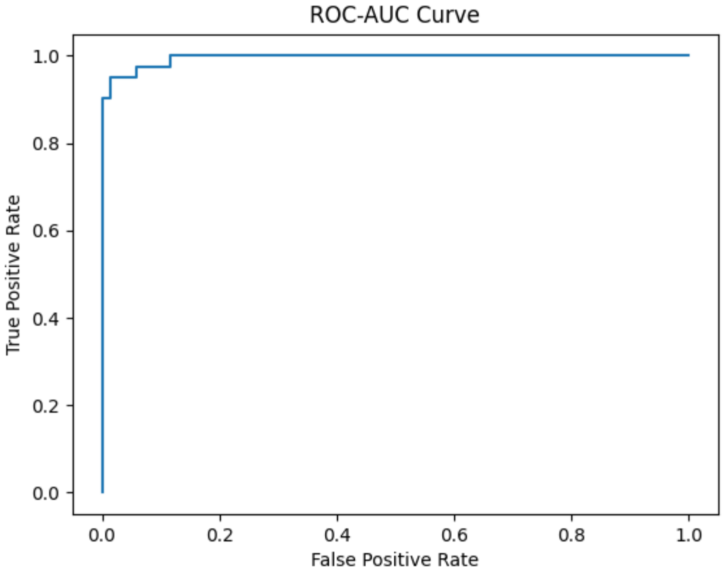

Figure 10: ROC-AUC Curve

Brief Analysis

The model shows good performance with high accuracy, precision and ROC-AUC, indicating it is effective at identifying both positive and negative cases. Also, the recall and F1 score are also high (above 90%), suggesting that the model can detect most of the positive cases while maintaining a good balance between precision and recall. The confusion matrix and ROC-AUC curve provide visual representations of the model's performance, showing that it can distinguish between the two classes effectively with very few false positives and false negatives. Overall, the model is well-suited for cancer detection and can be used to help diagnose patients with high accuracy and reliability.