Mohamed Mahmoud Emam 202202236
Mahmoud Tarek 202202051
Seif Aboshanab 202201838
Ameen Gamal 20222219

GitHub Repo: https://github.com/mohamed-7oda/DSAI-305-Project

Mid Progress Report

1) Preprocessing & EDA:

- *Data Overview:*
  - **Columns:** 31
  - **Rows:** 569
  - **Target Column Name:** Diagnosis
  - **Target Column Type:** object
  - **Feature Columns Names:** (radius1', 'texture1', 'perimeter1', 'area1', 'smoothness1', 'compactness1', 'concavity1', 'concave_points1', 'symmetry1', 'fractal_dimension1', 'radius2', 'texture2', 'perimeter2', 'area2', 'smoothness2', 'compactness2', 'concavity2', 'concave_points2', 'symmetry2', 'fractal_dimension2', 'radius3', 'texture3', 'perimeter3', 'area3', 'smoothness3', 'compactness3', 'concavity3', 'concave_points3' 'symmetry3', 'fractal_dimension3')
  - **Feature Columns Type:** float64

- *Preprocessing & Cleaning:*
  - **Nulls:** 0, There were no nulls in the data
  - **Duplicates:** 0, There were no duplicates in the data
  - There were **outliers**, and we handled them using the IQR method

- *Exploratory Data Analysis (EDA):*
  - **Univariate Analysis:** We used the .describe() function to calculate the necessary statistics for each feature, such as the mean, standard deviation, minimum, and maximum. 63% of the patients had benign tumors, while the remaining 37% had malignant tumors. 350 of the patients had benign tumors, while the remaining 219 had malignant tumors. We used the histogram and KDE plots to know the distribution of each feature

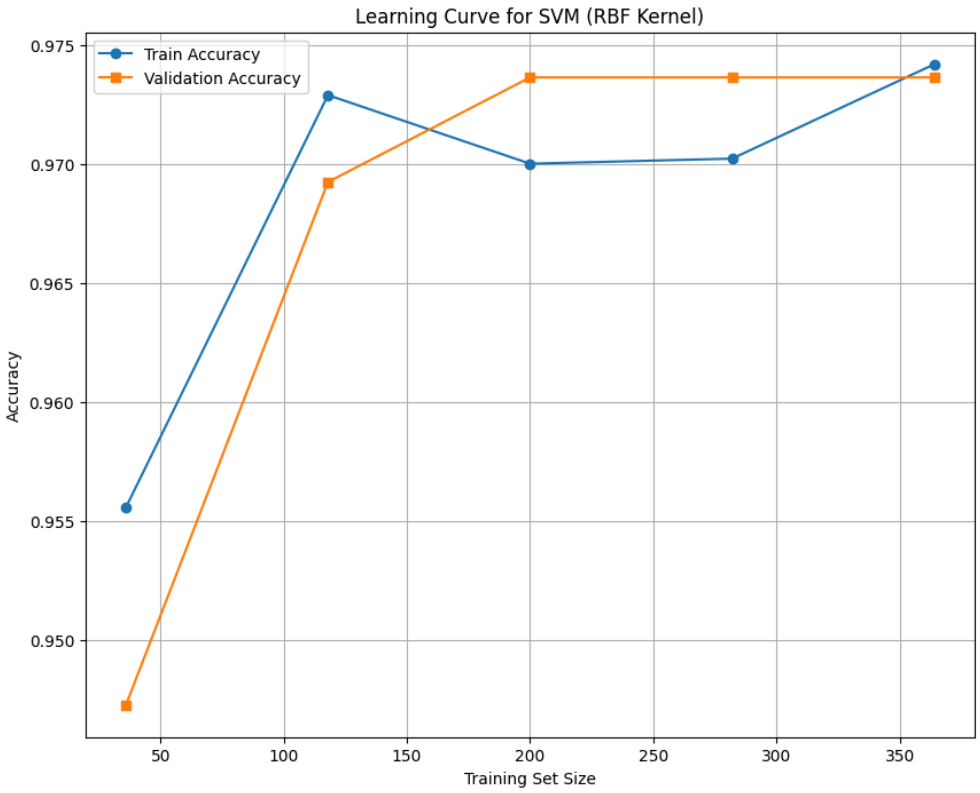
- **Bivariate/Multivariate Analysis:** We plotted the correlation matrix to understand the correlation between each feature and the others, as well as between each feature and the target, We found that there are many features that are correlated with each other. We plotted a boxplot for each feature with the target to understand the relationship between each feature and the target.

- *Feature Engineering and Selection:*
  - **Remove High Correlated Features:** We used the variance inflation factor (VIF) to identify any significant multicollinearity and to remove any features with high VIF. We found that the radius1 column had the highest VIF, followed by perimeter1, so we had to remove them.
  - **Feature Importance Using Mutual Information:** We calculated the feature importance using Mutual Information and found that the top three features were: (perimeter3, area3, and radius3), while the bottom three features were: (texture2, symmetry2, and fractal_dimension1)
  - **Feature Importance Using ANOVA:** Also we calculated the feature importance using ANOVA and found that the top three features were: (concave_points3, perimeter3, and concave_points1), while the bottom three features were: (fractal_dimension1,smoothness2, and texture)
  - **Feature Selection Using Variance Threshold**
  - In the end, we dropped a group of columns.

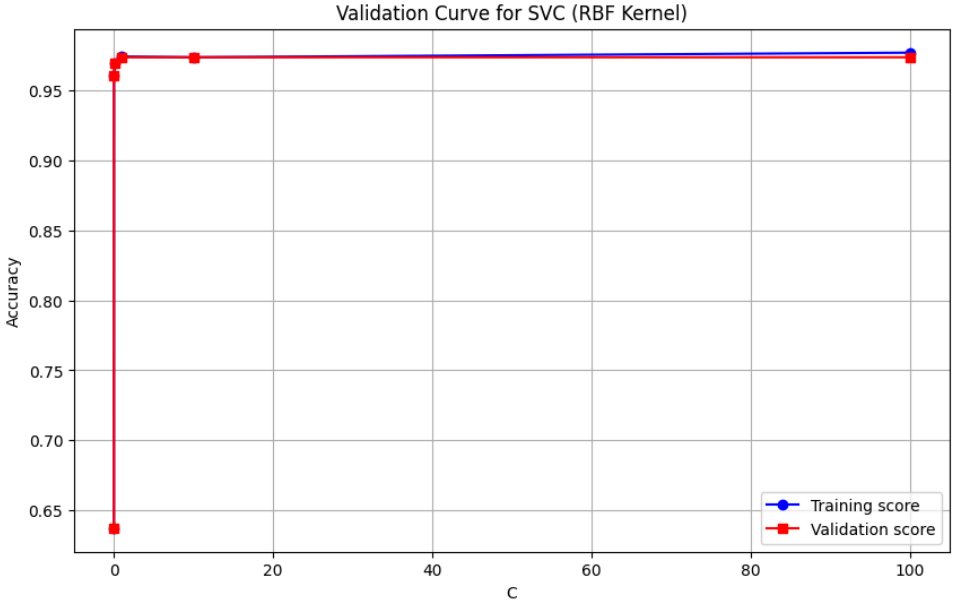
2) Model 1 SVM-LDA (Mohamed Mahmoud Emam):

- **Accuracy:** 0.9824
- **Precision:** 0.9829
- **Recall:** 0.9824
- **F1-Score:** 0.9823

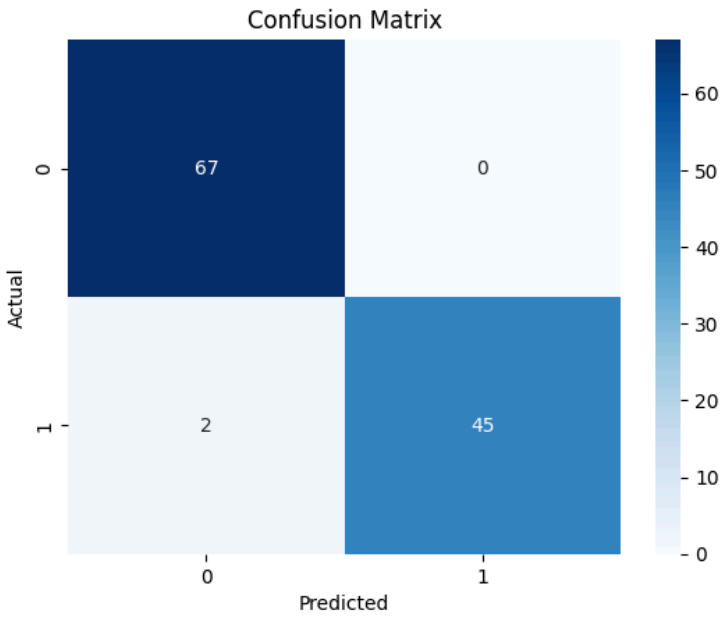
- **Learning Curve:**



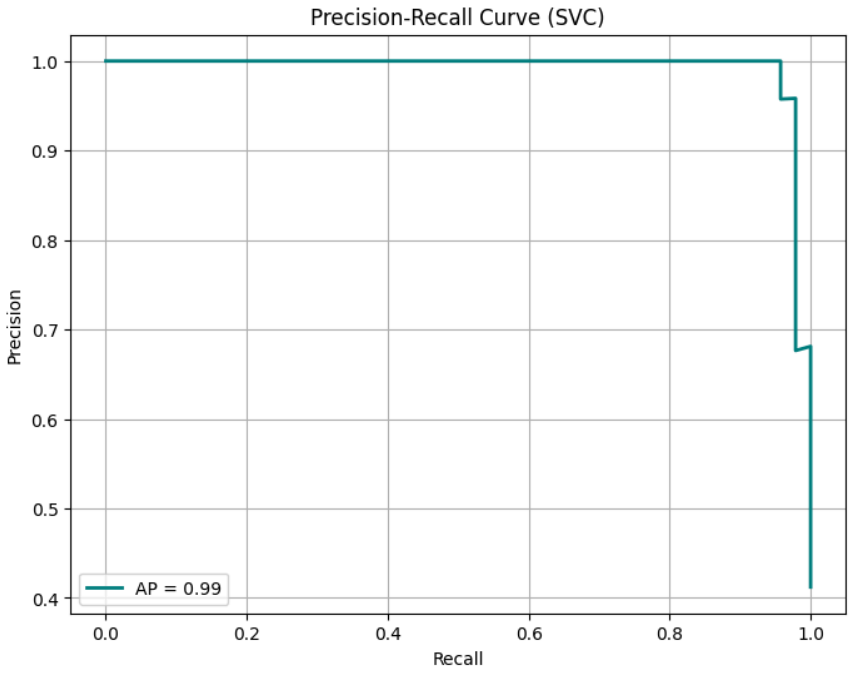
- **Validation Curve:**



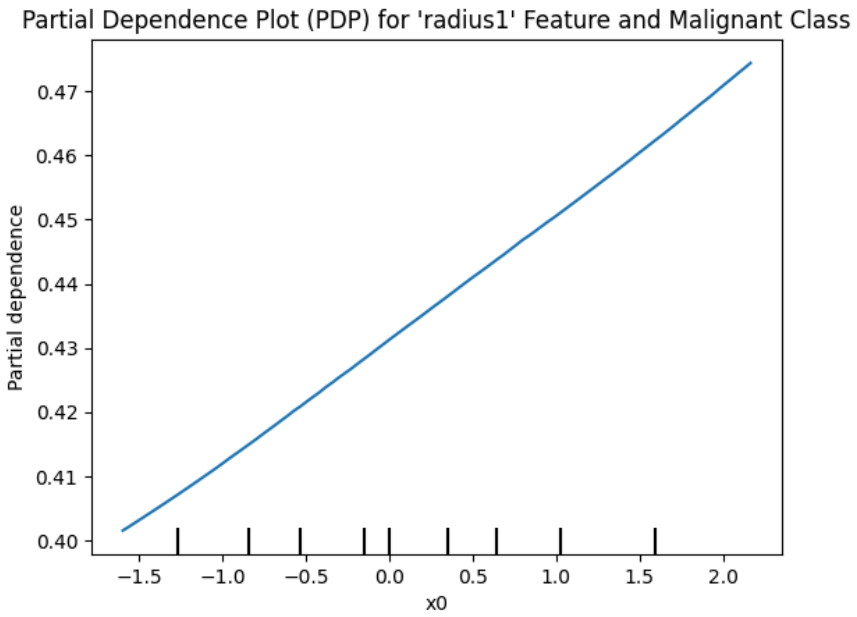
- **Confusion Matrix:**



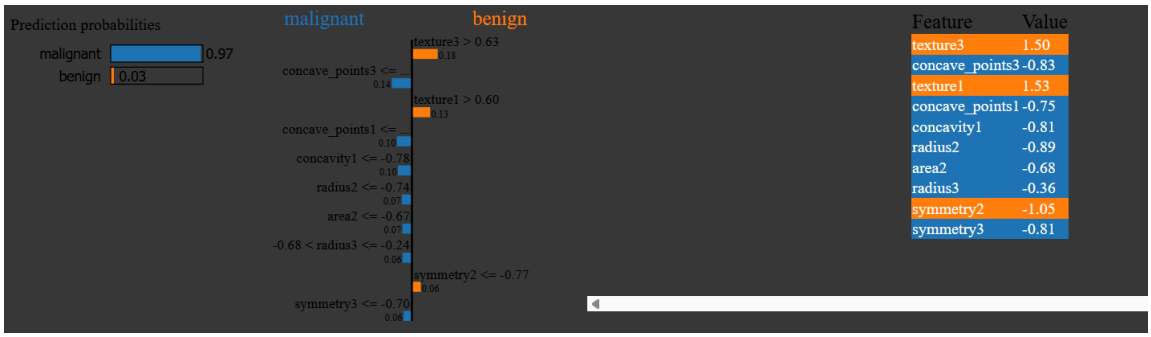
- **Precision-Recall Curve:**

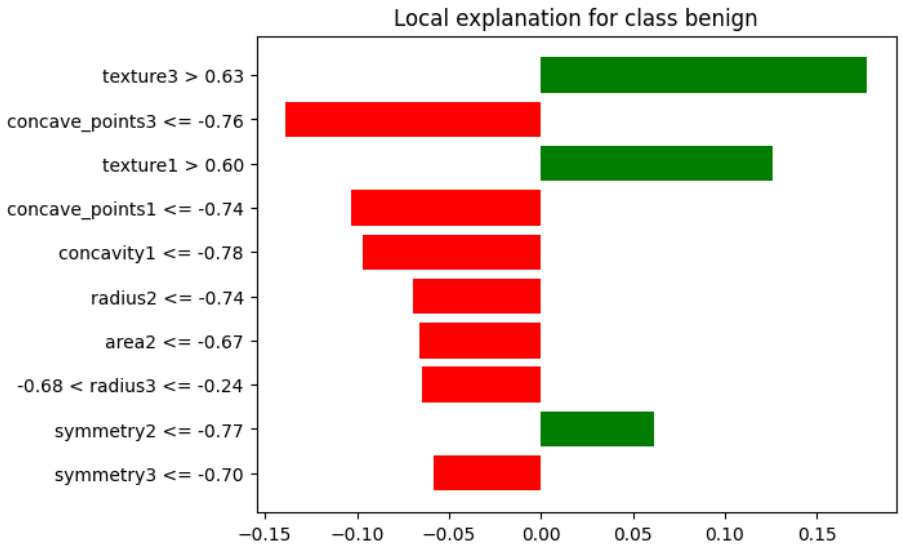


- **PDP Curve:**

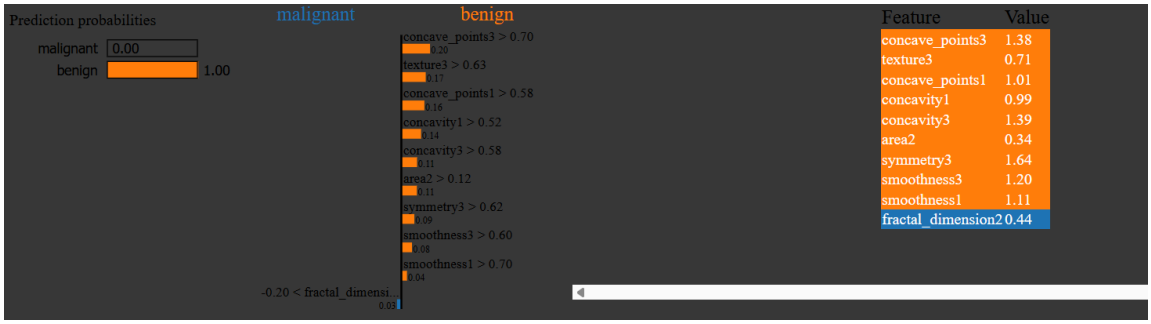


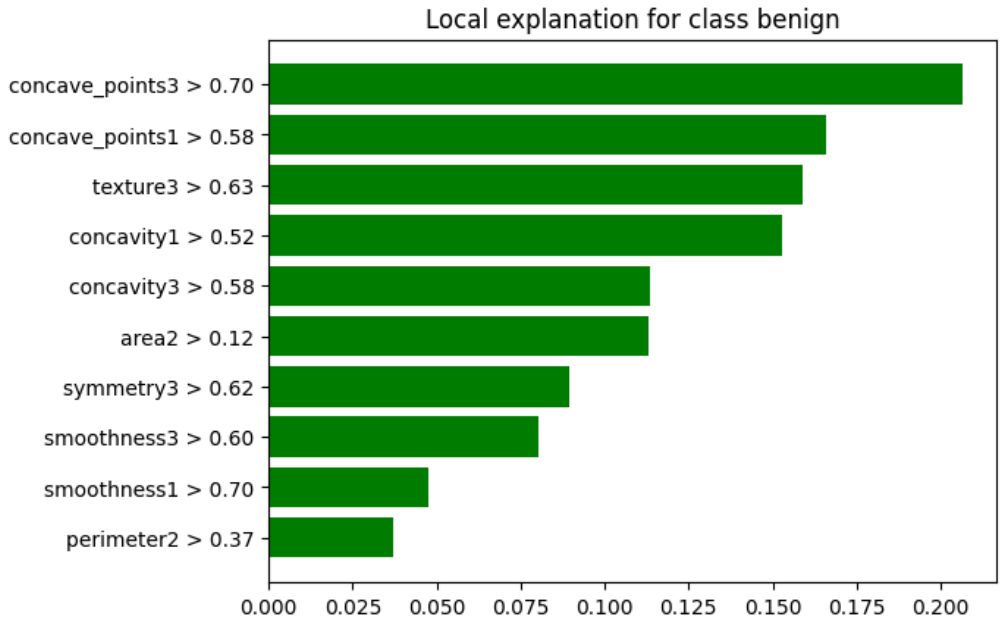
- **LIME for Sample 1:**





- **LIME for Sample 2:**





3) **Model 2 ANN (Saif Abushanab):**

- **Structure:**
  - **Input Layer**: 27 features (from the dataset)
  - **Hidden Layer 1**: 64 neurons, ReLU activation
    regularization
    Batch Normalization
    Dropout (rate = 0.2)

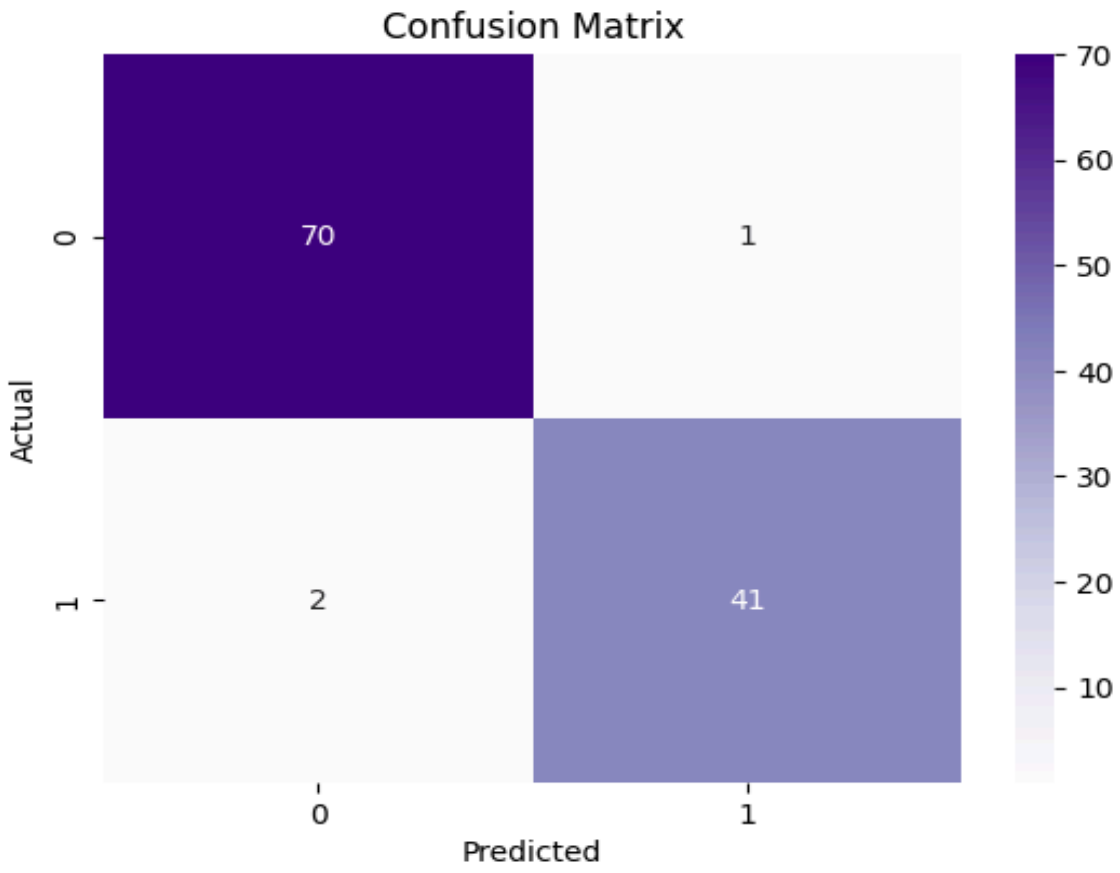
- **Hidden Layer 2**: 32 neurons, ReLU activation
  regularization
  Batch Normalization
  Dropout (rate = 0.2)

- **Output Layer**: 1 neuron, Sigmoid activation (for binary classification)

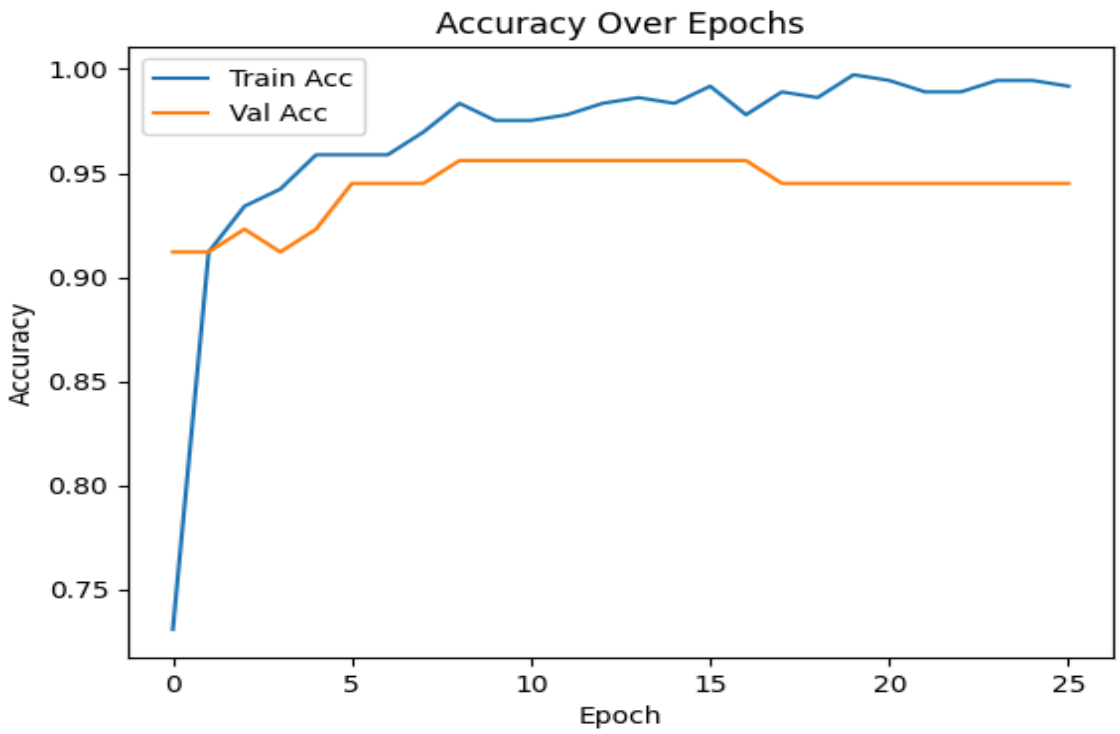
- **Output:**

```
Accuracy: 0.9736842105263158

              precision    recall  f1-score   support


           0       0.97      0.99      0.98        71

           1       0.98      0.95      0.96        43


    accuracy                           0.97       114

   macro avg       0.97      0.97      0.97       114

weighted avg       0.97      0.97      0.97       114


Final Epoch Accuracy :

Training Accuracy: 0.9918

Validation Accuracy: 0.9451
```

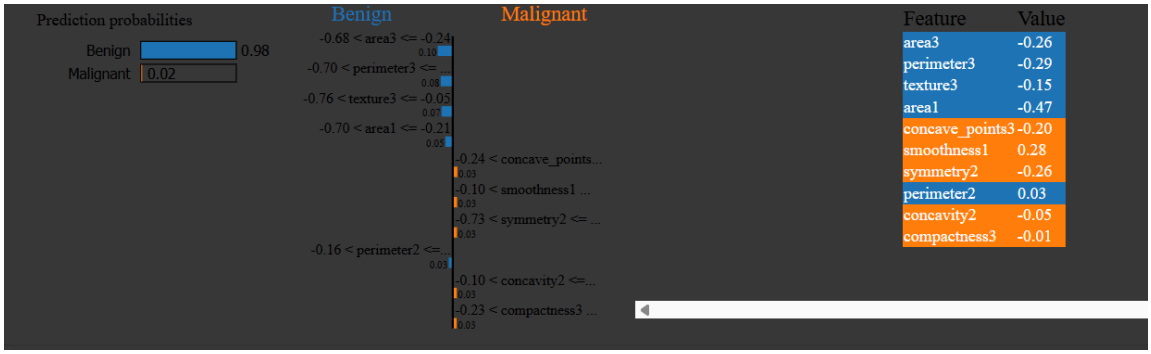
- **Confusion Matrix**



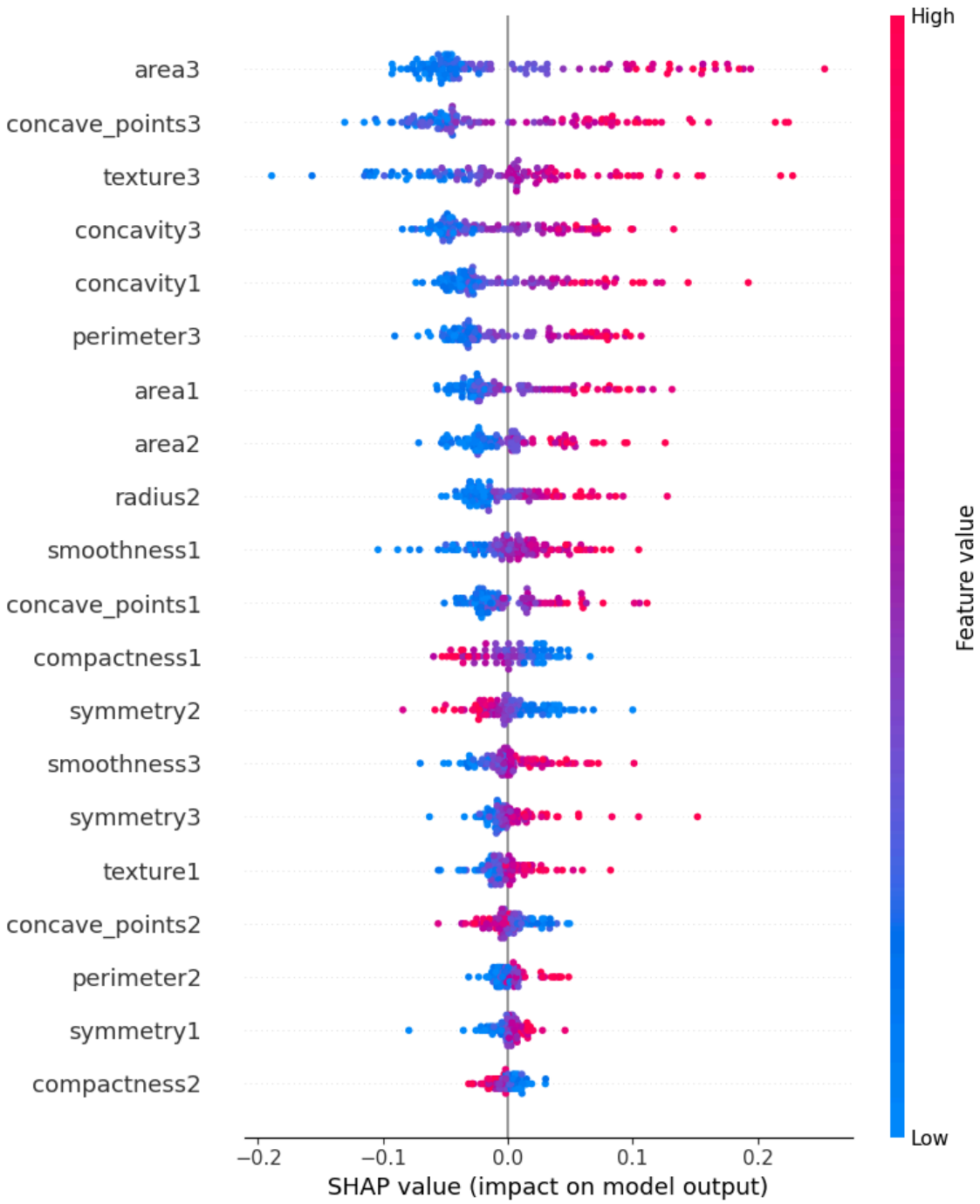
- **Training Performance**


Accuracy Over Epochs
Train Acc
Val Acc
Accuracy
Epoch
1.00
0.95
0.90
0.85
0.80
0.75
0
5
10
15
20
25

- **LIME**

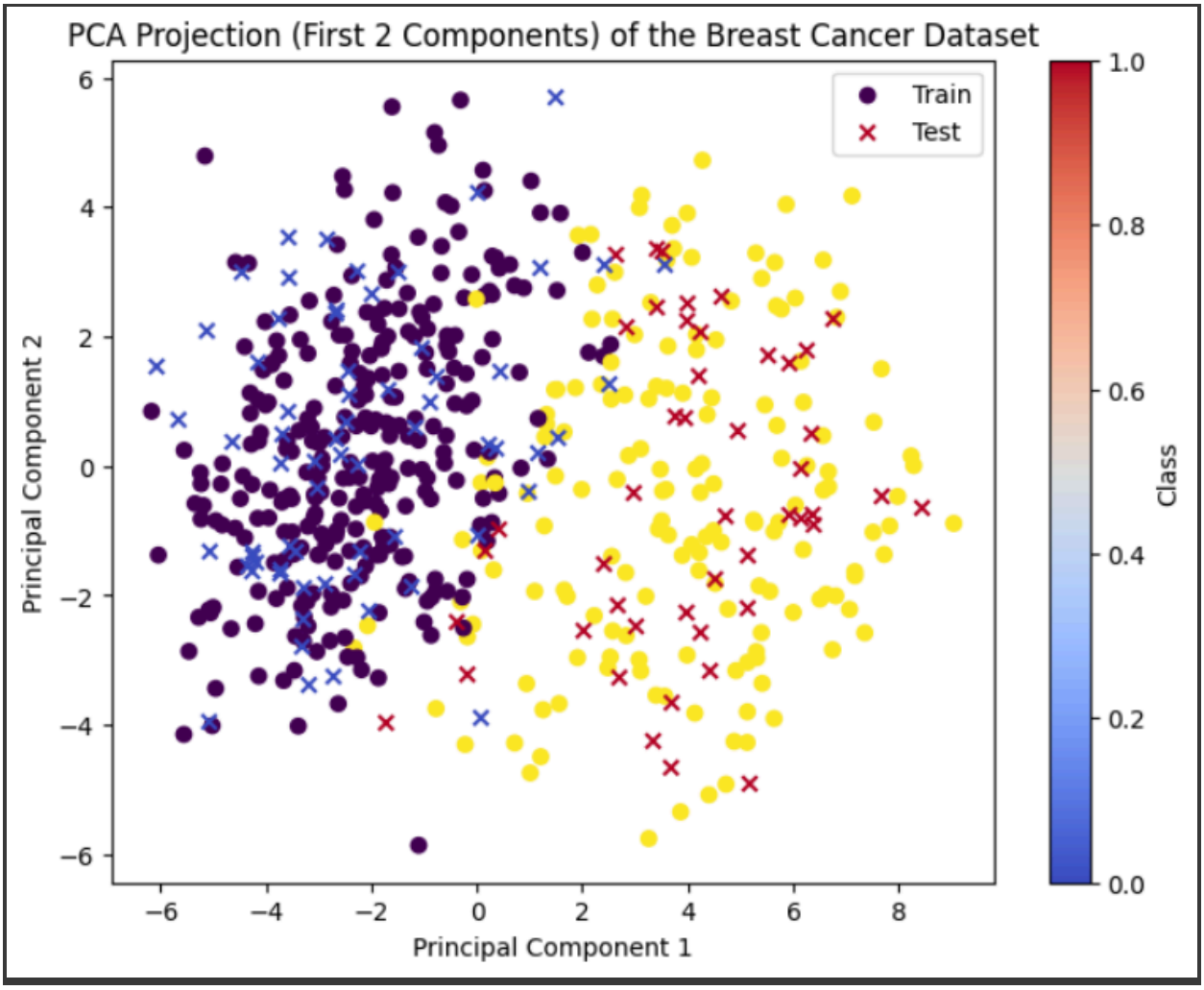

Prediction probabilities
Benign
Malignant
0.98
0.02
Benign
Malignant
-0.68 < area3 <= -0.24
-0.70 < perimeter3 <= ...
-0.76 < texture3 <= -0.05
-0.70 < area1 <= -0.21
-0.24 < concave_points...
-0.10 < smoothness1 ...
-0.73 < symmetry2 <= ...
-0.16 < perimeter2 <= ...
-0.10 < concavity2 <=...
-0.23 < compactness3 ...
Feature
Value
area3 -0.26
perimeter3 -0.29
texture3 -0.15
area1 -0.47
concave_points3 -0.20
smoothness1 0.28
symmetry2 -0.26
perimeter2 0.03
concavity2 -0.05
compactness3 -0.01

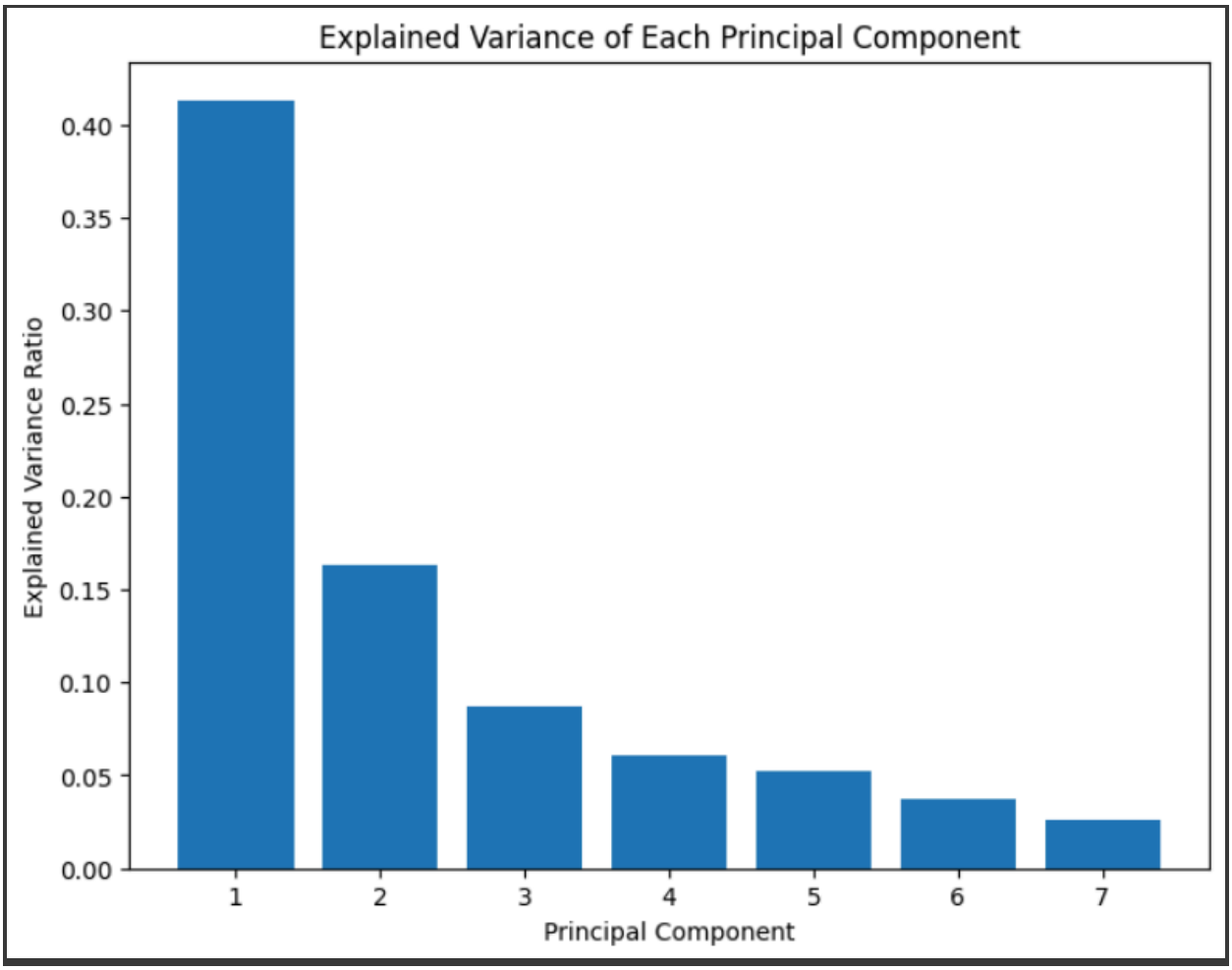
- **SHAPE**

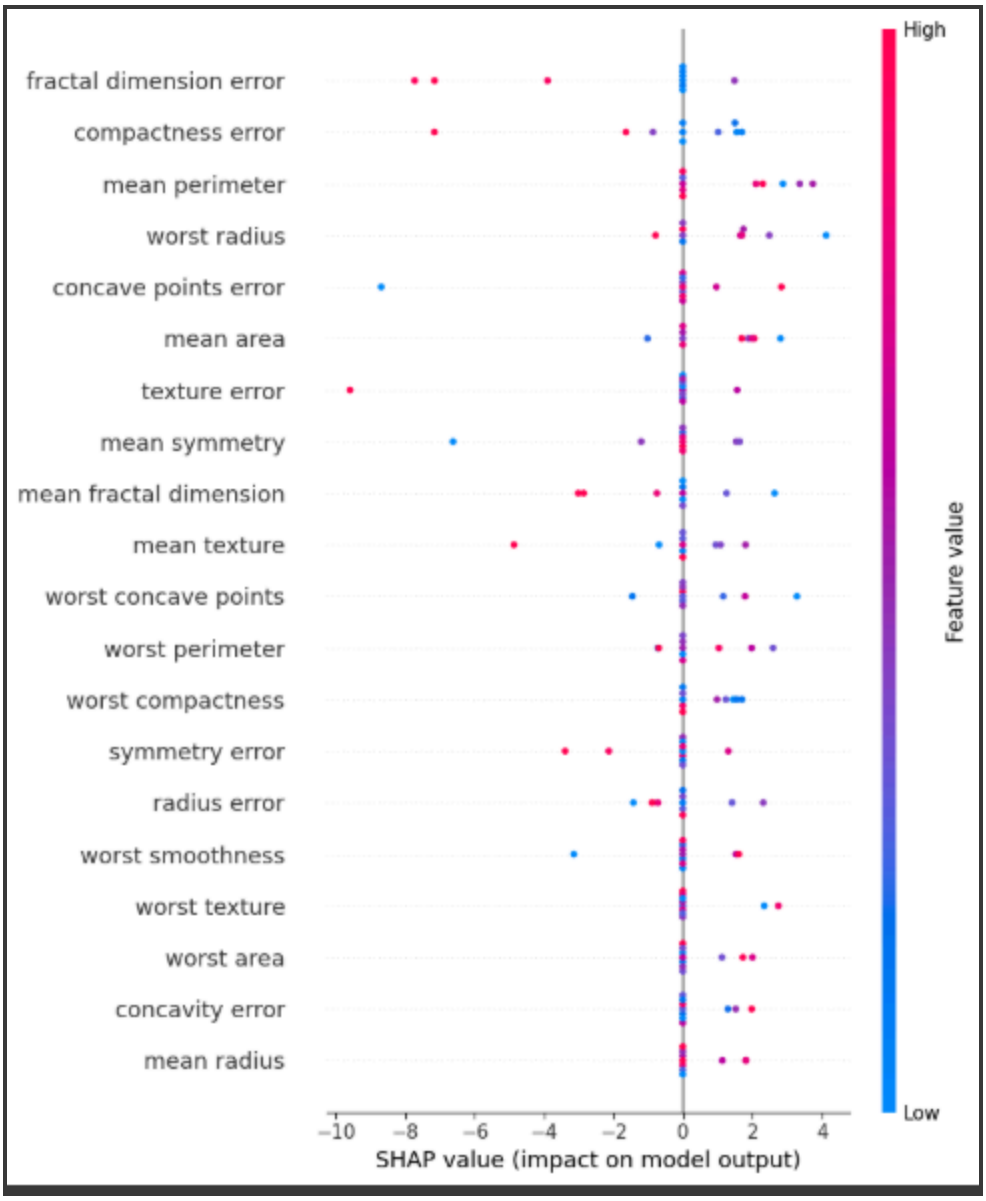


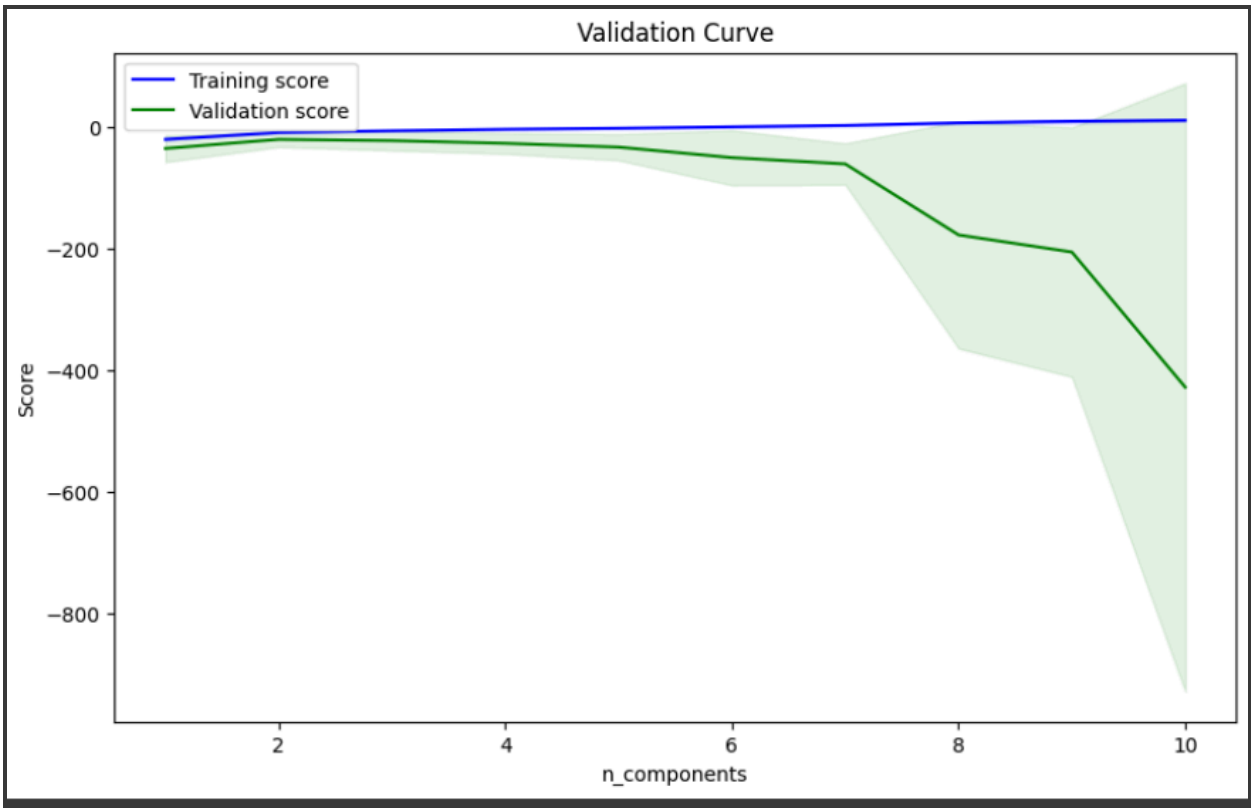
4) Model 3 MLIT (Amin Gamal - 202202219):

- **Accuracy:** `0.9386`
- **Precision:** `0.95`
- **Recall:** `0.94`
- **F1-Score:** `0.95`



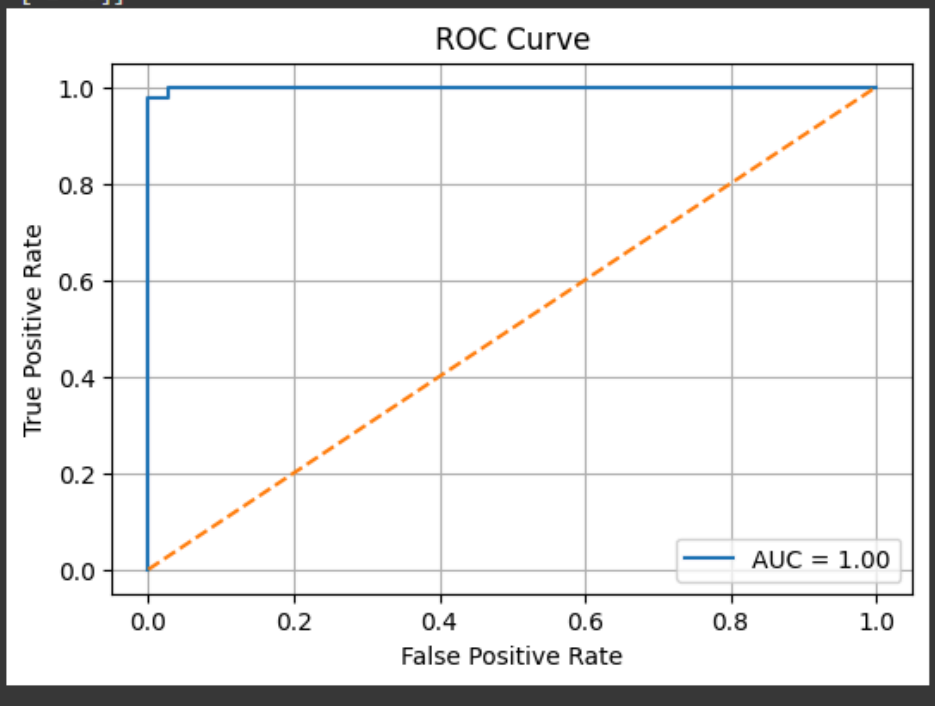

Explained Variance of Each Principal Component
Explained Variance Ratio
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
1
2
3
4
5
6
7
Principal Component


fractal dimension error
compactness error
mean perimeter
worst radius
concave points error
mean area
texture error
mean symmetry
mean fractal dimension
mean texture
worst concave points
worst perimeter
worst compactness
symmetry error
radius error
worst smoothness
worst texture
worst area
concavity error
mean radius
−10
−8
−6
−4
−2
0
2
4
SHAP value (impact on model output)
High
Low
Feature value



4) Model 4 Random Forest Classifier (Mahmoud Tarek):

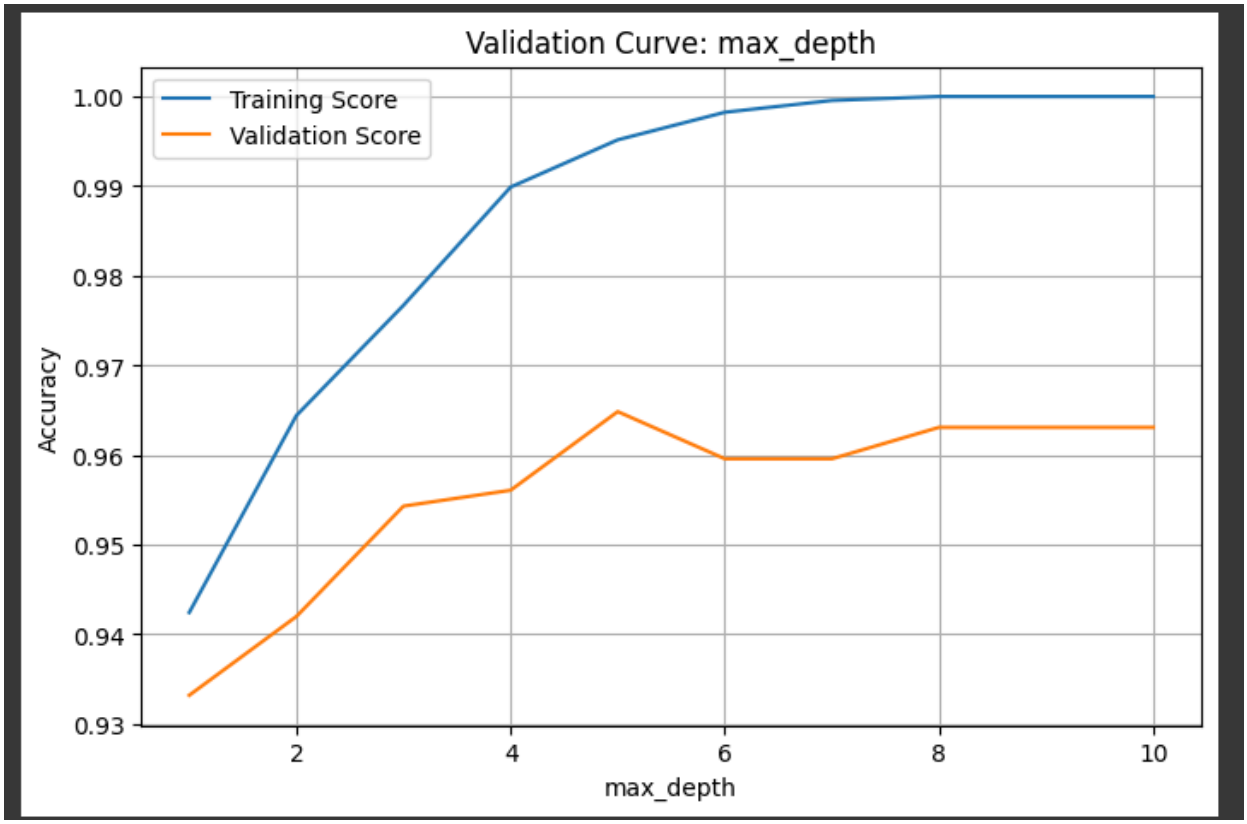
```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        72
           1       1.00      0.95      0.98        42

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114

Confusion Matrix:
 [[72  0]
 [ 2 40]]
```



Classification Report and ROC Curve:

- **The classification report shows high precision, recall, and F1-score, especially for the class 0. The accuracy is 0.98, which is a good indicator of the model's performance.**

- **The ROC curve with AUC = 1.00 suggests perfect classification without any false positives or negatives.**



Validation Curve for max_depth:

- **This plot shows how the model's performance varies with the `max_depth` hyperparameter. The training score continues to improve with depth, but the validation score starts to plateau, indicating potential overfitting if the depth increases beyond a certain point.**

```python
from sklearn.metrics import ConfusionMatrixDisplay

ConfusionMatrixDisplay.from_estimator(rf_model, X_test, y_test, cmap='Blues')
plt.title("Confusion Matrix")
plt.grid(False)
plt.show()
```