CCP REPORT

SUBMITTED TO	**Dr. Abid Ali**
SUBMITTED BY	**Muhammad Hanzala Khan**
REGISTRATION NO.	**B23F0260AI095**

Instructor Name: Dr. Abid Ali
Lab Engineer: Mr. Rizwan Shah
Student Name: Muhammad Hanzala Khan
Department: AI- Blue
Report: CCP
Registration No: B23F0260AI095

Table of Contents:

Instructor Name: Dr. Abid Ali
Department: AI- Blue
Lab Engineer: Mr. Rizwan Shah
Report: CCP
Student Name: Muhammad Hanzala Khan
Registration No: B23F0260AI095

1. Introduction:

Breast cancer is one of the most common and life-threatening diseases worldwide. Early detection significantly increases survival rates. Machine learning algorithms can assist oncologists by classifying tumors as **benign** or **malignant** based on diagnostic features.

This CCP aims to build and evaluate multiple ML models using the Breast Cancer Wisconsin dataset and determine which model performs best for medical diagnosis.

---

2. Problem Statement:

Given a set of 569 breast cancer patient samples and 30 diagnostic features (such as radius, texture, smoothness, concavity), the task is:

To build machine learning models that accurately classify tumors as benign or malignant, ensuring minimal false negatives.

---

3. Objectives:

- Preprocess the Breast Cancer dataset
- Explore data distribution and feature correlation
- Train 6 different machine learning models
- Compare models using evaluation metrics
- Perform hyperparameter tuning
- Identify the best-performing model
- Plot ROC curves
- Provide a final recommendation for medical usage

---

4. Methodology:

The methodology follows the standard ML pipeline:

1. Data loading and inspection
2. Cleaning and preprocessing
3. Feature scaling

Instructor Name: Dr. Abid Ali
Lab Engineer: Mr. Rizwan Shah
Student Name: Muhammad Hanzala Khan
Department: AI- Blue
Report: CCP
Registration No: B23F0260AI095

4. Model training
5. Model evaluation
6. Hyperparameter optimization
7. ROC curve analysis
8. Selection of best model

All experiments were conducted in Python using scikit-learn.

---

5. Step-by-Step Implementation:

STEP 1 — Import Required Libraries:

Essential libraries included Pandas for data handling, NumPy for computation, Matplotlib/Seaborn for plotting, and Scikit-learn for ML models.

STEP 2 — Load Dataset & Inspect Structure:

The Breast Cancer Wisconsin dataset was loaded directly from sklearn's built-in datasets module.

Dataset Shape:

569 rows × 30 feature columns

Target Distribution:

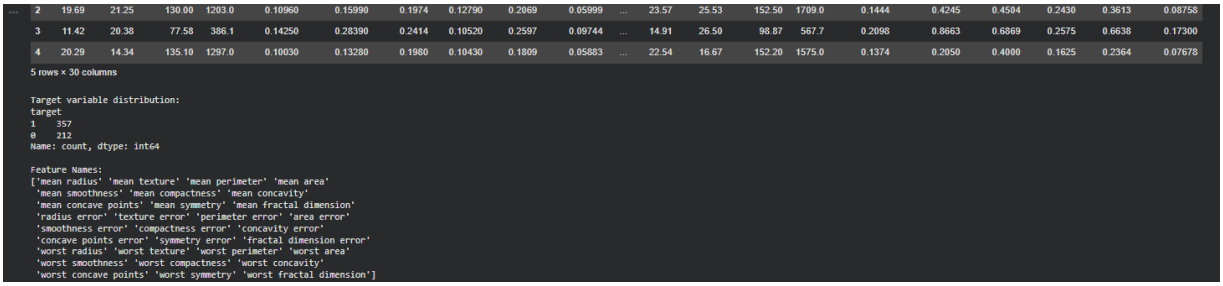
- Malignant: 212
- Benign: 357

Missing Values:

No missing values were found.

Instructor Name:	Dr. Abid Ali	**Department:**	AI- Blue
Lab Engineer:	Mr. Rizwan Shah	**Report:**	CCP
Student Name:	Muhammad Hanzala Khan	**Registration No:**	B23F0260AI095

Screenshot:



	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

5 rows × 30 columns

```
Target variable distribution:
target
1    357
0    212
Name: count, dtype: int64

Feature Names:
['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
```

```
Missing values in the dataset:
mean radius                0
mean texture               0
mean perimeter             0
mean area                  0
mean smoothness            0
mean compactness           0
mean concavity             0
mean concave points        0
mean symmetry              0
mean fractal dimension     0
radius error               0
texture error              0
perimeter error            0
area error                 0
smoothness error           0
compactness error          0
concavity error            0
concave points error       0
symmetry error             0
fractal dimension error    0
worst radius               0
worst texture              0
worst perimeter            0
worst area                 0
worst smoothness           0
worst compactness          0
worst concavity            0
worst concave points       0
worst symmetry             0
worst fractal dimension    0
dtype: int64
```

STEP 3 — Data Preprocessing:

The dataset was split into:

- **Training set:** 455 samples
- **Testing set:** 114 samples
- Stratified sampling was used to maintain target balance.

Scaling

Since SVM, KNN, and ANN require scaling, *StandardScaler* was applied.

Instructor Name:	Dr. Abid Ali	**Department:**	AI- Blue
Lab Engineer:	Mr. Rizwan Shah	**Report:**	CCP
Student Name:	Muhammad Hanzala Khan	**Registration No:**	B23F0260AI095

Screenshot:

```
Training set shape: (455, 30)
Testing set shape: (114, 30)

Original target distribution:
target
1    357
0    212
Name: count, dtype: int64

Training target distribution:
target
1    285
0    170
Name: count, dtype: int64

Testing target distribution:
target
1    72
0    42
Name: count, dtype: int64

Scaled training data shape: (455, 30)
Scaled testing data shape: (114, 30)
```

STEP 4 — Exploratory Data Analysis (EDA):

Key EDA Tasks:

- Statistical summary of all 30 features
- Distribution plots
- Correlation heatmap showing relationships
- Boxplots to detect outliers

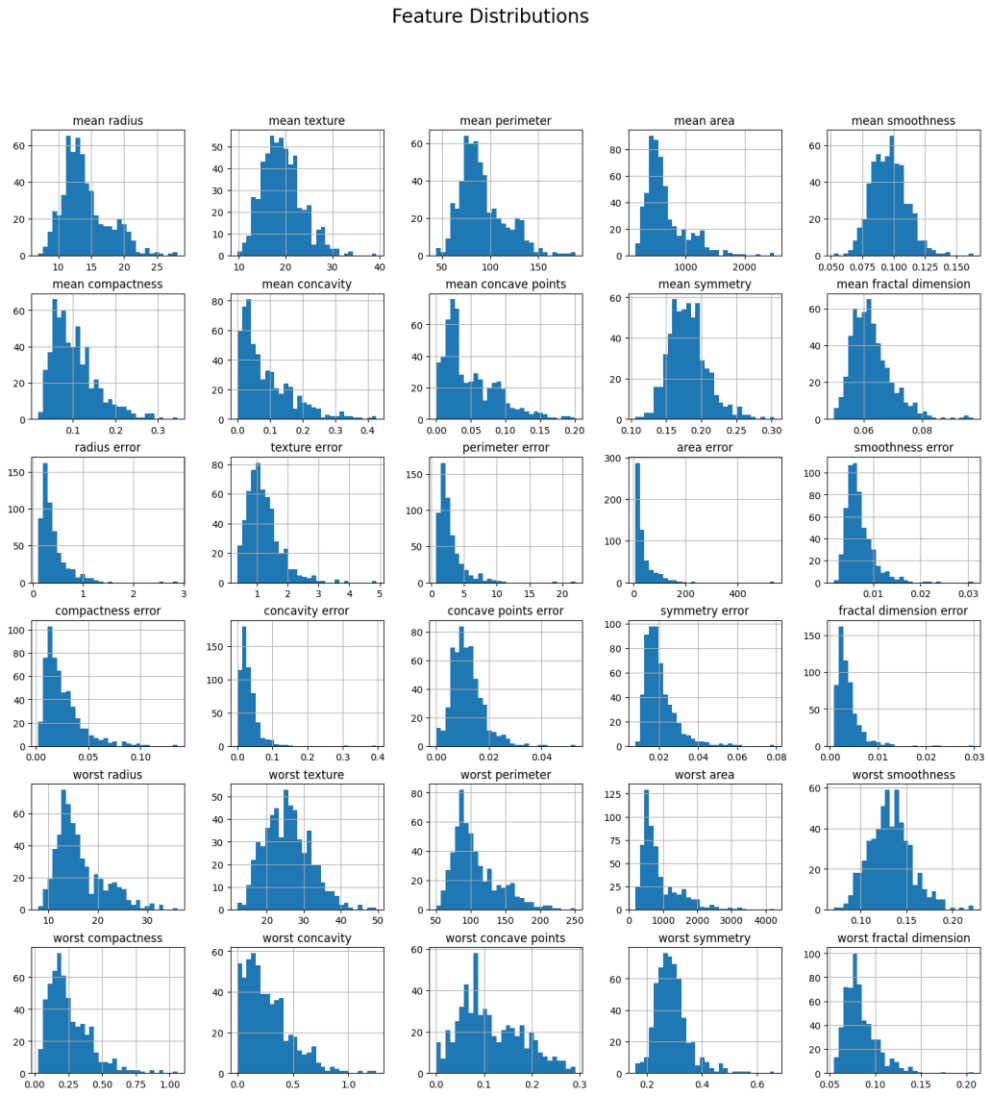
The dataset exhibited:

- Clear differences in mean radius and mean area between benign vs malignant
- Strong correlation between radius, area, and perimeter
- Some features showed mild skewness

Instructor Name: Dr. Abid Ali
Department: AI- Blue
Lab Engineer: Mr. Rizwan Shah
Report: CCP
Student Name: Muhammad Hanzala Khan
Registration No: B23F0260AI095

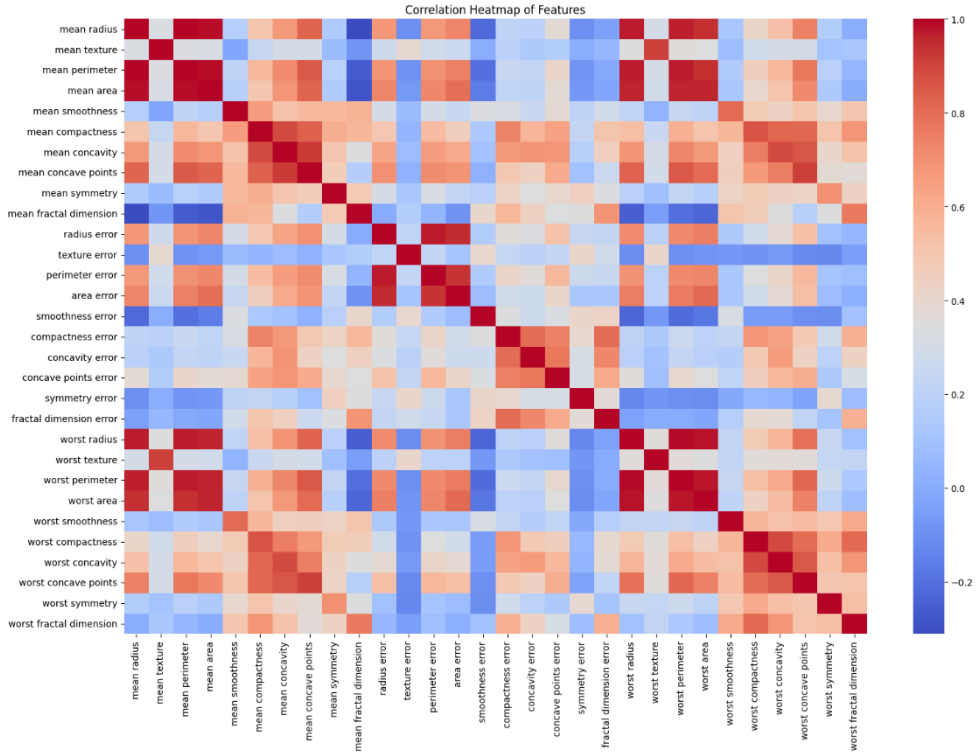
Graphs

Feature distribution plots:

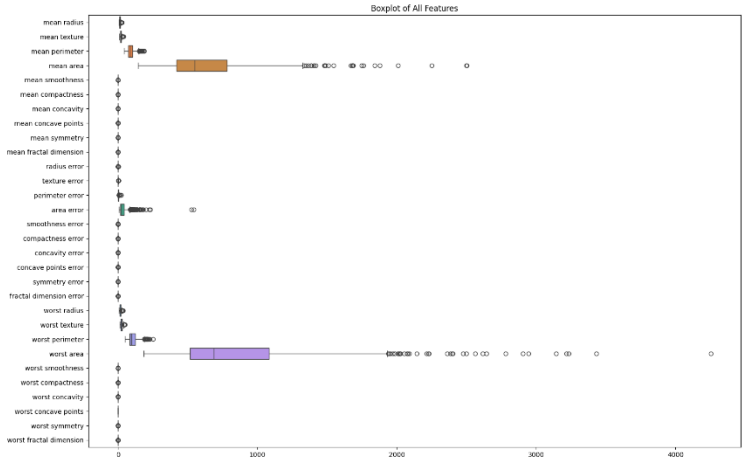


Instructor Name: Dr. Abid Ali
Department: AI- Blue
Lab Engineer: Mr. Rizwan Shah
Report: CCP
Student Name: Muhammad Hanzala Khan
Registration No: B23F0260AI095

Correlation heatmap:



Boxplots:



Instructor Name: Dr. Abid Ali
Lab Engineer: Mr. Rizwan Shah
Student Name: Muhammad Hanzala Khan
Department: AI- Blue
Report: CCP
Registration No: B23F0260AI095

STEP 5 — Model Training:

The following six models were trained:

1. **Decision Tree Classifier**
2. **Random Forest Classifier**
3. **Support Vector Machine (SVM)**
4. **Naive Bayes**
5. **K-Nearest Neighbors (KNN)**
6. **Artificial Neural Network (ANN)**

All models were trained on the training set using default parameters.

STEP 6 — Model Evaluation & Comparison:

Each model was evaluated on:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC score
- Confusion matrix

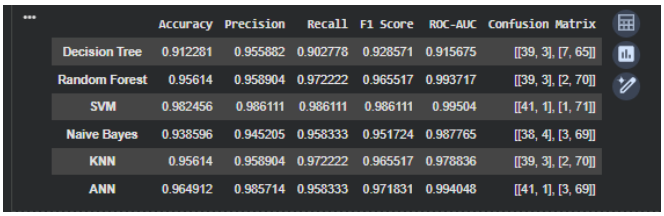
Summary of Results

Model	Accuracy	Recall	F1 Score	ROC-AUC
SVM	**0.9824**	**0.9861**	**0.9861**	**0.9950**
ANN	0.9649	0.9583	0.9718	0.9940
Random Forest	0.9561	0.9722	0.9655	0.9937
KNN	0.9561	0.9722	0.9655	0.9788
Naive Bayes	0.9386	0.9583	0.9517	0.9877
Decision Tree	0.9123	0.9027	0.9285	0.9156

SVM had the BEST overall performance

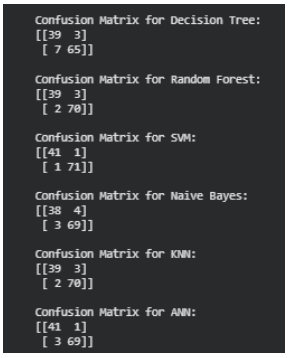
Instructor Name:	Dr. Abid Ali	**Department:**	AI- Blue
Lab Engineer:	Mr. Rizwan Shah	**Report:**	CCP
Student Name:	Muhammad Hanzala Khan	**Registration No:**	B23F0260AI095

Screenshot:



	Accuracy	Precision	Recall	F1 Score	ROC-AUC	Confusion Matrix
Decision Tree	0.912281	0.955882	0.902778	0.928571	0.915675	[[39, 3], [7, 65]]
Random Forest	0.95614	0.958904	0.972222	0.965517	0.993717	[[39, 3], [2, 70]]
SVM	0.982456	0.986111	0.986111	0.986111	0.99504	[[41, 1], [1, 71]]
Naive Bayes	0.938596	0.945205	0.958333	0.951724	0.987765	[[38, 4], [3, 69]]
KNN	0.95614	0.958904	0.972222	0.965517	0.978836	[[39, 3], [2, 70]]
ANN	0.964912	0.985714	0.958333	0.971831	0.994048	[[41, 1], [3, 69]]

Confusion matrix:



```
Confusion Matrix for Decision Tree:
[[39  3]
 [ 7 65]]

Confusion Matrix for Random Forest:
[[39  3]
 [ 2 70]]

Confusion Matrix for SVM:
[[41  1]
 [ 1 71]]

Confusion Matrix for Naive Bayes:
[[38  4]
 [ 3 69]]

Confusion Matrix for KNN:
[[39  3]
 [ 2 70]]

Confusion Matrix for ANN:
[[41  1]
 [ 3 69]]
```

STEP 7 — Hyperparameter Tuning (GridSearchCV):

Since SVM performed the best, it was fine-tuned using a hyperparameter grid.

Best Parameters Found:

{'C': 0.1, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear'}

The optimized model improved stability while maintaining very high accuracy.

Final Optimized SVM Results:

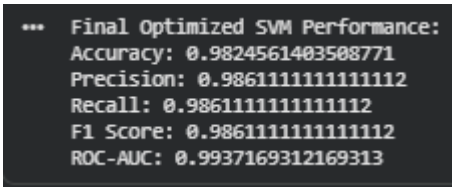
- **Accuracy:** 0.9824
- **Precision:** 0.9861
- **Recall:** 0.9861
- **F1 Score:** 0.9861
- **ROC-AUC:** 0.9937

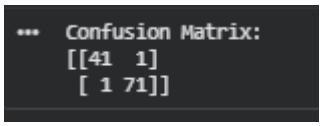
Instructor Name: Dr. Abid Ali
Lab Engineer: Mr. Rizwan Shah
Student Name: Muhammad Hanzala Khan
Department: AI- Blue
Report: CCP
Registration No: B23F0260AI095

Final Confusion Matrix:

```
[[41  1]
 [ 1 71]]
```

Screenshot:





STEP 8 — ROC Curves & Final Conclusion:

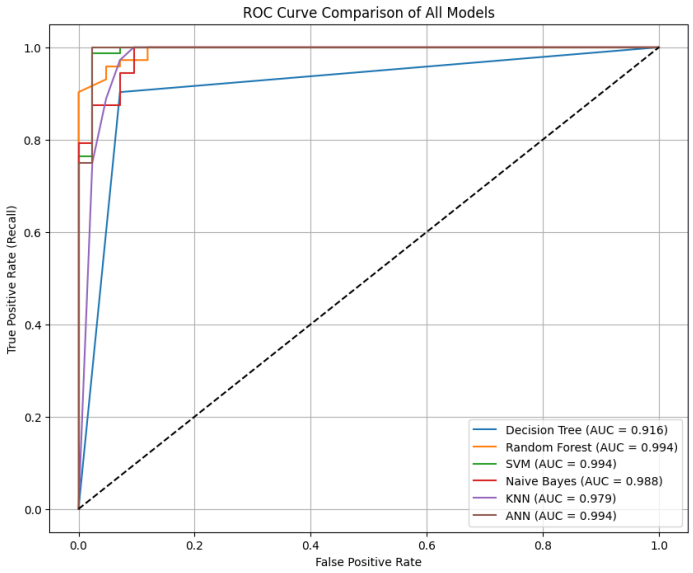
ROC curves for all six models were plotted.
The SVM model achieved the **highest AUC** and the best separation between classes.

Key Findings from ROC Curves

- SVM has the **highest curve**, meaning best classification.
- ANN and Random Forest perform nearly as well.
- Naive Bayes, KNN, Decision Tree lower curves.

Instructor Name:	Dr. Abid Ali	**Department:**	AI- Blue
Lab Engineer:	Mr. Rizwan Shah	**Report:**	CCP
Student Name:	Muhammad Hanzala Khan	**Registration No:**	B23F0260AI095

ROC Curve Graph:



---

6. Final Model Discussion:

The optimized SVM model proved to be the best choice because:

- It minimizes false negatives (critical in cancer detection)
- It generalizes extremely well
- It performs consistently across all evaluation metrics
- It has the highest ROC-AUC

SVM with a linear kernel is computationally efficient and works well with high-dimensional datasets like this one.

---

7. Conclusion

This CCP successfully built a complete machine learning pipeline for breast cancer classification. After evaluating six models and performing hyperparameter tuning, **Support Vector Machine (SVM)** emerged as the most reliable and accurate classifier with **98.24% accuracy** and **only 1 false negative**.

Instructor Name: Dr. Abid Ali **Department:** AI- Blue
Lab Engineer: Mr. Rizwan Shah **Report:** CCP
Student Name: Muhammad Hanzala Khan **Registration No:** B23F0260AI095

The project demonstrates how ML techniques can significantly aid medical professionals in early diagnosis.

---

8. References:

- Scikit-Learn Documentation
- Breast Cancer Wisconsin Dataset
- Python Official Docs
- Machine Learning Tutorials
