Breast Cancer Classification Using Machine Learning

By IBITOYE ARAFAH IBITAYO

Introduction

Breast cancer remains one of the leading causes of death among women globally.

- **2.1 million** new cases worldwide (2020)
- **700,000+** deaths
- In Nigeria, breast cancer is the **most common cancer in women**
- Accounts for **22.7%** of all new cancer cases (WHO)
- Early detection significantly improves survival outcomes

Goal: Build a machine learning system to classify tumors as *benign* or *malignant* for early diagnosis.

Understanding Breast Lumps

There are Two Types

Types of Breast Lumps

1. Malignant

- Cancerous
- Invade nearby tissues
- Can spread if not detected early

2. Benign

- Non-cancerous
- Do not spread
- Generally harmless but still require monitoring

Project Workflow:

The workflow for this project is organized into several key stages:

1. Data Acquisition & Preparation
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Model Training & Evaluation
5. Hyperparameter Tuning (SVM)
6. Model Interpretation (SHAP)
7. Conclusion & Recommendations

Data Acquisition & Preparation

Dataset: Breast Cancer Wisconsin dataset (from Kaggle)

Steps performed:

- Checked for missing values
- Standardized features using StandardScaler
- Encoded labels: *Benign = 0, Malignant = 1*
- Split data into **80% training** and **20% testing** sets

Exploratory Data Analysis (EDA)

- Conducted descriptive statistics to understand feature distributions.
- Created visualizations (pairplots, histograms, heatmaps) to identify correlations and patterns between features.
- Observed strong relationships between cell shape, size, and texture with cancer classification.

Correlation Heatmap Insights:

- Concave points_worst, perimeter_worst, radius_worst, and area_worst showed high positive correlation with malignancy
- Features like smoothness_se and fractal_dimension_mean showed weak correlation
- In general, irregular shapes → higher likelihood of being malignant

Feature Engineering

- Standardized all numerical features
- Scaler fitted on **training set only**
- Same transformation applied to test set (to avoid data leakage)

Model Training & Evaluation

- Single Model Evaluation: Trained a Logistic Regression model (random_state=42) using scaled training data. Predicted on the test set and evaluated performance using accuracy score and a classification report.
- Multiple Model Comparison: Implemented and trained five classification algorithms — Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Gradient Boosting — to compare predictive performance. Measured accuracy and reviewed classification metrics for each model.
- Confusion Matrix Visualization: Plotted confusion matrices for all five models to visualize misclassifications, with emphasis on detecting false negatives.

MODEL PEFORMANCE SUMMARY

Top Performing Model: SVM (Before Tuning)

- **Accuracy:** 98.25%
- High precision & recall for both tumor types
- Lowest misclassification rate among all models

Hyperparameter Tuning

Model: Support Vector Machine (SVM)

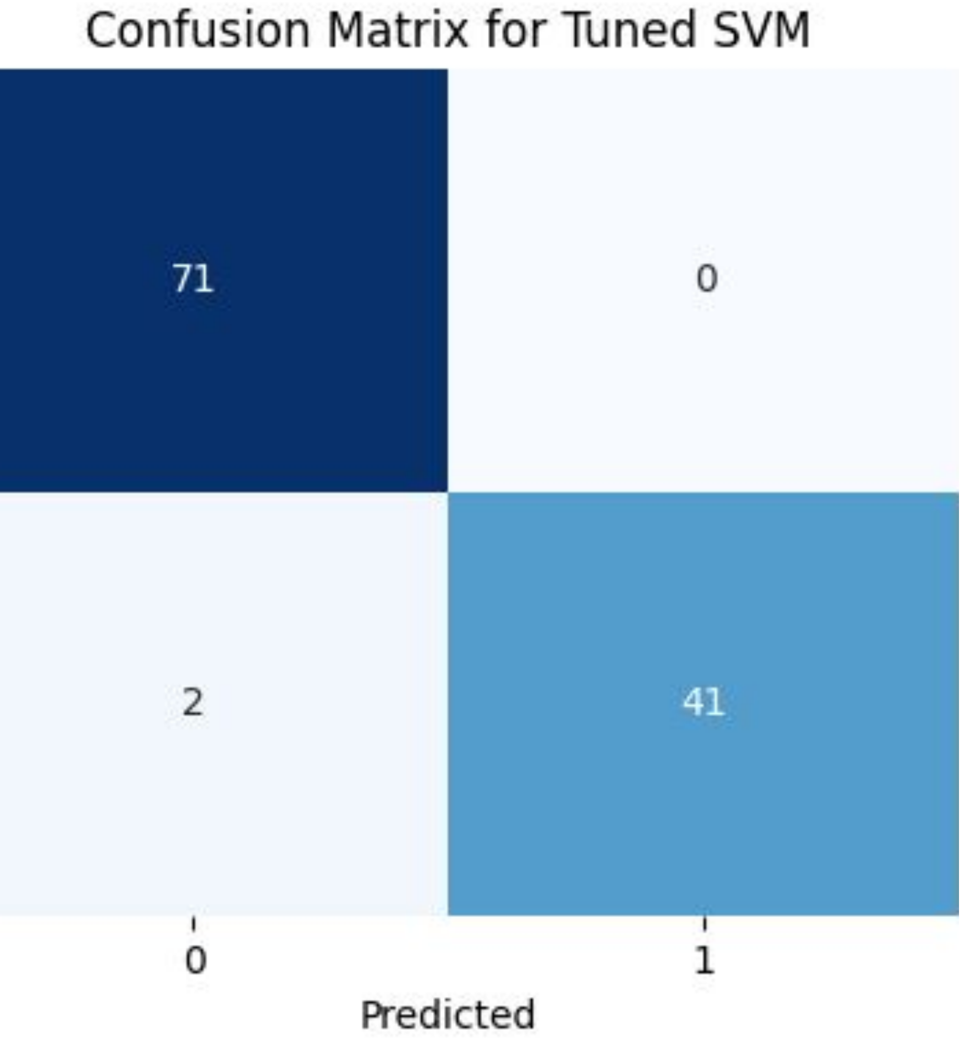
Method: **GridSearchCV (5-fold CV)**

Parameters tuned:

- C (regularization strength)
- Kernel (linear, rbf)
- Gamma (for rbf kernel)

Tuned Model Performance:

- **Accuracy:** 98%
- Benign Recall: **100%**
- Malignant Recall: **95%**
- Strong balance between sensitivity & specificity



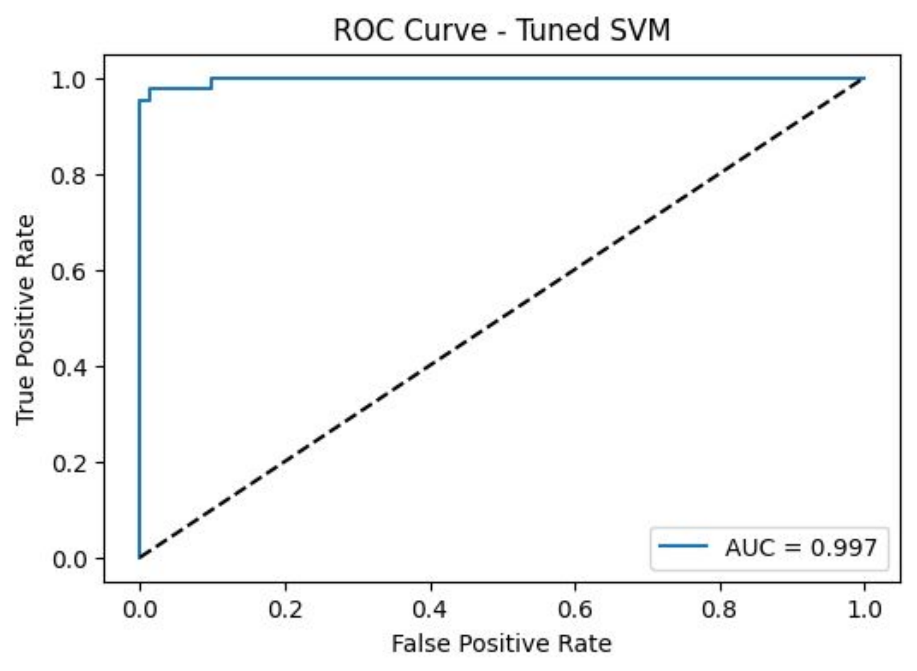
Confusion Matrix Insights

The confusion matrix shows:

- 71 benign cases were all correctly classified (no false negatives for benign).
- 43 malignant cases had 2 false negatives, meaning they were misclassified as benign.

While the false negative count is small, in medical diagnosis, even a single false negative can be critical, as it means a malignant case was missed. This highlights a possible area for further improvement, such as experimenting with different kernels, cost-sensitive learning, or ensemble methods.

To further validate the model, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were computed. The resulting AUC score of 0.997 indicates that the model can distinguish between malignant and benign cases with 99.7% probability, even when the classification threshold is varied. The ROC curve's close alignment with the top-left corner demonstrates consistently strong sensitivity and specificity across all thresholds, confirming the model's exceptional generalization capability.



SHAP Model Interpretability

SHAP analysis was used to understand how features influence predictions.

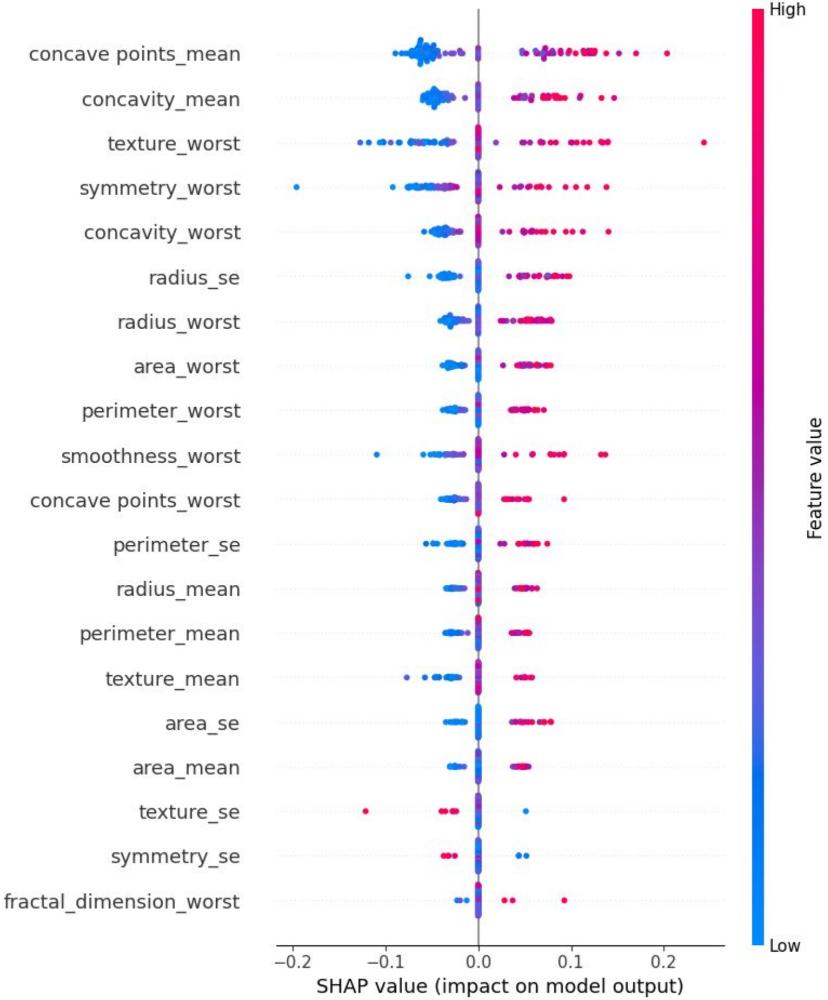
Most influential features:

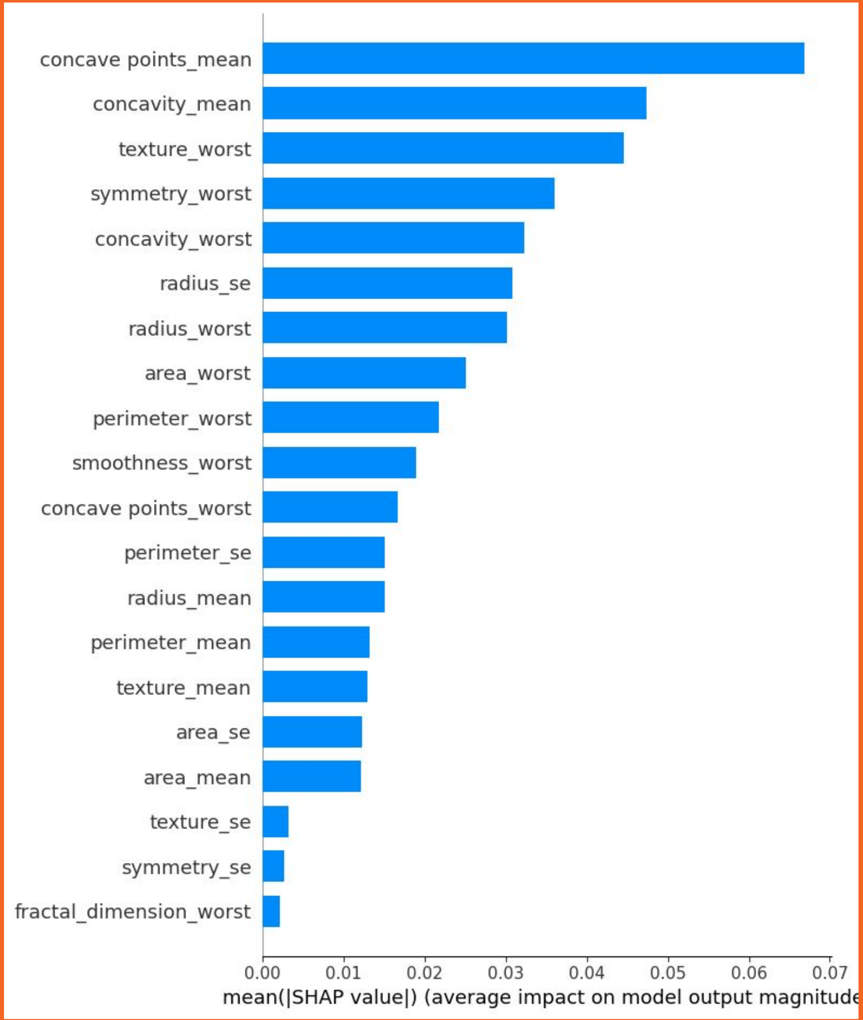
- concave points_mean
- concavity_mean
- texture_worst
- area_worst

Least influential features:

- symmetry_se
- texture_se

Higher values of concave points & concavity typically indicate malignant tumors.


concave points_mean
concavity_mean
texture_worst
symmetry_worst
concavity_worst
radius_se
radius_worst
area_worst
perimeter_worst
smoothness_worst
concave points_worst
perimeter_se
radius_mean
perimeter_mean
texture_mean
area_se
area_mean
texture_se
symmetry_se
fractal_dimension_worst
−0.2
−0.1
0.0
0.1
0.2
SHAP value (impact on model output)
High
Feature value
Low


concave points_mean
concavity_mean
texture_worst
symmetry_worst
concavity_worst
radius_se
radius_worst
area_worst
perimeter_worst
smoothness_worst
concave points_worst
perimeter_se
radius_mean
perimeter_mean
texture_mean
area_se
area_mean
texture_se
symmetry_se
fractal_dimension_worst
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
mean(|SHAP value|) (average impact on model output magnitude

SHAP Summary Explanation

SHAP revealed:

- Higher values of *concave points_mean*, *concavity_mean*, and *texture_worst* push predictions toward the **malignant** class.
- Lower values of these features are associated with **benign** predictions.
- Some features, like *smoothness_worst*, show mixed effects and are less decisive.
- The model relies on clinically meaningful indicators such as tumor shape irregularity and concavity — aligning with medical literature.

.

CONCLUSION

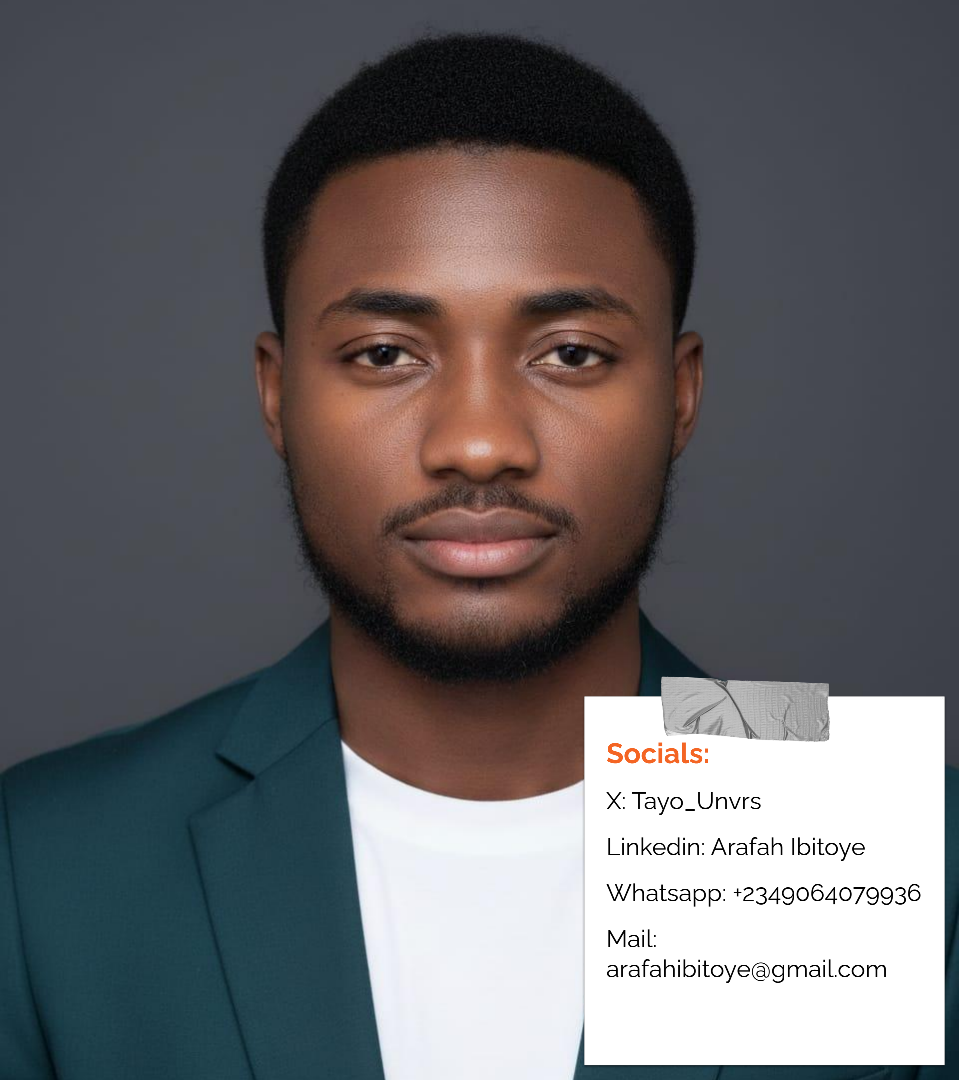
- The SVM model achieved excellent performance with 98% accuracy and strong recall scores.
- SHAP confirmed that the model uses medically relevant features, enhancing trust and interpretability.
- The model shows strong potential as a decision-support tool for clinicians.

RECOMMENDATION

To prepare the model for real-world clinical use:

- **External Validation:** Test on larger and more diverse datasets.
- **Clinical Integration:** Build a simple interface for doctors to use predictions and explanations.
- **Bias & Error Analysis:** Identify demographic or sampling biases.
- **Periodic Retraining:** Update model as new medical data becomes available

Thank You



Socials:

X: Tayo_Unvrs

Linkedin: Arafah Ibitoye

Whatsapp: +2349064079936

Mail:
arafahibitoye@gmail.com

Hi, I’m Arafah Ibitoye

Data Scientist & ML Engineer

Passionate about building AI solutions for health and environmental impact.