Machine Learning Classification of Breast Tumor Diagnosis

Using Composite Feature Engineering and Ensemble Models

Course: 3253-090 - Machine Learning

Instructor: Dr. Carl Jackson, MSci, PhD

Prepared by Group 4:

Rahulkumar Ajani

Elba Gomez Navas

Ankita Nayak

Ashley Fergusson

Nikola Stojanovic

Date: November 24, 2025

Executive Summary

This report presents an end-to-end machine learning project focused on **classifying** breast **tumors** as **benign or malignant** using the **Wisconsin Diagnostic Breast Cancer** dataset. A rigorous pipeline was followed, including early train–test splitting to avoid leakage, structured exploratory analysis, composite feature engineering, and evaluation of seven machine learning models. XGBoost emerged as the best model with 96.88 % recall on malignant tumors. SHAP explainability confirmed that area, shape and texture irregularity related features are the most influential predictors. The project demonstrates strong generalization, clinical interpretability, and alignment with best practices for medical AI model development.

Objective

The objective of this project was to build a predictive machine learning model capable of determining whether a breast tumor is benign or malignant. **Beyond highest recall on malignant, the aim was to construct a medically interpretable, leak-free, fully engineered pipeline** that meets academic and clinical expectations.

Data Preparation

Data was obtained from the Wisconsin Diagnostic Breast Cancer dataset. An early **train–test split (70/30)** was performed to prevent leakage. Extensive exploratory analysis was conducted using reduced feature subsets. Strong multicollinearity motivated composite feature engineering using Random Forest–derived weights. Outliers were clipped and RobustScaler was applied to all numerical features.

Model Design

Seven models were trained using 5-fold stratified cross-validation: Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, SVM (RBF), and MLPClassifier. Models were evaluated using accuracy, precision, recall, F1, and ROC-AUC. Tree-based and kernel models performed strongly, reflecting the **dataset's natural separability**.

Model Evaluation

Post tuning, XGBoost achieved the **strongest test performance** with 98.83% accuracy, 1.00 precision, and 96.88% recall for malignant class. The confusion matrix revealed **only 2 false negatives**. ROC-AUC was 0.995.

Conclusions

The project fulfilled all objectives, producing a clinically strong and highly interpretable model. Composite features significantly improved structure and interpretability. Limitations include dataset size and lack of external validation. Future improvements include threshold tuning, incorporating additional patient features, and testing on external datasets.

Table of Contents

Objective

Breast cancer diagnosis is one of the most important and sensitive applications of machine learning. The clinical setting places two primary constraints on predictive models:

1. **Minimize false negatives**: missing a malignant tumor has severe consequences.
2. **Maintain** very **high interpretability**

Clinicians must understand why a model makes a prediction.

The purpose of this project was to design and evaluate a robust machine learning pipeline capable of predicting tumor malignancy using features derived from digitized cell nuclei. The goal was not simply to achieve high recall, but to:

- Engineer features that reveal underlying medical structure
- Explore and justify preprocessing decisions rigorously
- Train multiple strong models and compare them systematically
- Validate generalization and absence of overfitting
- Analyze the final model using SHAP (Lundberg & Lee, 2017) interpretability tools (Molnar, 2022).
- Produce a **rigorous**, academic-quality analysis **suitable for management and clinical audiences**

Given the clinical nature of the problem**, recall on malignant cases** (sensitivity) **was the primary metric**, supported by accuracy, precision, F1, and AUC.

Data Preparation

Dataset Description and Source

The dataset was sourced from:

https://raw.githubusercontent.com/rahulajani/ml/main/project_data.csv

It reflects the well-known Wisconsin Diagnostic Breast Cancer (WDBC) dataset (Street, Wolberg, & Mangasarian, 1993). Each row corresponds to measurements extracted from a digitized breast mass image. The dataset contains:

- 569 total records
- 30 numeric predictor variables
- 1 categorical target variable (“diagnosis”)
- 3 sub-categories: *_mean, *_se, *_worst
- No missing values in any features.

Preventing Data Leakage: Early Train–Test Split

A critical step:

Train–test split was performed **before** any exploratory analysis.

- Train: 398 samples (70%)
- Test: 171 samples (30%)
- **Stratified** by diagnosis (B/M) to preserve B vs. M ration in test and train set

This early train-test split prevents data leakage from test set during the exploratory stage.

Data Structure and Feature Groups

The dataset exhibits a highly structured design with three sub-sets:

- Group 1 ***_mean** feature : mean measurements of tumor characteristics
- Group 2 ***_se** features: standard error estimates
- Group 3 ***_worst** features: representing worst measurements for a given patient

Treating these groups separately allows analytical clarity and **reduces dimensionality explosion**.

Statistical Profiling

A thorough descriptive statistical table was generated (count, mean, std, median, mode, quantiles, min/max, range). This revealed some key insights as follows:

A. Feature scale varies dramatically

Example:

- fractal_dimension_mean ranges ~0.05–0.1
- area_worst ranges ~185–3432

This scale mismatch **justifies using a RobustScaler** instead of standard normalization.

B. Many features show distinct distributions

Benign and malignant tumors exhibit vastly different value distributions, even before modeling.This suggests the dataset will be highly learnable.

C. Several "_worst" features dominate

These reflect more severe tumor characteristics and appear consistently more discriminative. This later becomes important when interpreting SHAP values.

Correlation Analysis and Multicollinearity

Correlation heatmaps revealed extremely high correlations:

Mass-related group showed correlation between **0.94 and 1.0**:

- radius_mean ↔ perimeter_mean ↔ area_mean
- radius_se ↔ perimeter_se ↔ area_se
- radius_worst ↔ perimeter_worst ↔ area_worst

Shape irregularity group showed correlation between 0.77 and 0.91:

- compactness_*
- concavity_*
- concave points_*

This is not just statistical noise—it reflects real biological relationships:

- A larger tumor radius almost always implies a larger perimeter and area.
- Higher concavity often implies more concave points.

Composite Feature Engineering (Major Contribution of This Project)

We engineered two composite feature families, each **informed by** Random Forest **feature importance weights**.

This approach:

- Prevents multicollinearity
- Retains predictive information
- Improves interpretability

A. Shape Irregularity Composites: core_weighted_

The compactness_, concavity_, concave points_ within each data subset are highly correlated, and hence a composite feature was added within each sub-set. Example for "_mean" subset:

- compactness_mean → weight 0.087
- concavity_mean → weight 0.278
- concave points_mean → weight 0.635

B. Mass-Nature Composites: mass_nature_

The features smoothness, texture, area describe nature of tumor mass but have very low correlation. Hence, these were combined to produce new feature mass_nature_. Example for "_worst" subset:

- area_worst contributes ~0.80+
- smoothness_worst and texture_worst add subtle refinements

Why weighted composites are better than PCA

- PCA makes features abstract and harder to interpret
- Composite features preserve semantic meaning ("shape irregularity score")
- Physicians can understand what they represent

This composite-engineering step is one of the strongest elements of the project.

Reduced Feature Set for Visualization Only

To visualize patterns without noise:

- Radius_* and perimeter_* were removed
- Core-weighted and mass-nature composites added
- Final EDA set had **18 interpretable features**

This is **strictly for EDA**; training data remained untouched.

WHAT EDA REVEALED

- **Boxplots**: Benign vs malignant distributions showed minimal overlap for: area_* and core_weighted_*
- **Pairplots**: Showed clear clustering and separability.
- 3D PCA: Showed geometric separation with minimal overlap

Data Processing Pipeline

The data pipeline takes clean copy of X_train, y_train, and X_test, y_test. It performs following operations from scratch:

- Target encoding
- Feature-importance based weights calculation and deriving composite features
- Imputing was not necessary as the data had no missing values
- Outlier handling: clipped to 1st and 99th percentile
- Robust scaling

Final output and shapes after preprocessing:

- X_train_scaled: 398 × 21
- X_test_scaled: 171 × 21

Model Design

The modeling stage was designed with two goals:

1. Breadth: evaluate a diverse set of algorithms (linear, nonlinear, tree-based, neural).
2. Depth: focus on clinically meaningful metrics (malignant recall) and understand why a model works using interpretability tools.

All models were trained on X_train_scaled and y_train_encoded, using Stratified 5-fold cross-validation, ensuring each fold preserved the benign/malignant distribution.

The following models were implemented:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. XGBoost (Chen & Guestrin, 2016)
5. LightGBM (Ke et al., 2017)
6. SVM with RBF kernel
7. MLPClassifier (feed-forward neural network)

For each, we measured:

- Recall (malignant class)
- Accuracy
- Precision (for malignant class)
- F1 score
- ROC-AUC

Rationale for Model Choices

Logistic Regression

- Serves as a baseline linear classifier.
- Easy to interpret and calibrate.

Decision Tree

- Highly interpretable, visualizable decision rules.
- Known to overfit, but helpful to understand approximate hierarchical splits.

Random Forest

- An ensemble of many decision trees with bootstrapping and feature randomness.
- Reduces overfitting relative to a single tree.
- Naturally provides feature importance scores (already leveraged earlier in feature engineering).

XGBoost

- Gradient-boosted trees with regularization.
- Known for outstanding tabular performance.
- Handles complex nonlinear decision boundaries.

LightGBM

- Another gradient boosting framework designed for speed and efficiency.
- Good benchmark alongside XGBoost, often comparable performance with less training time.

SVM (RBF)

- Strong margin-based classifier well-suited for highly separable data.
- RBF kernel allows non-linear separation in feature space.

MLPClassifier (Neural Network)

- Tests whether a simple feed-forward neural architecture offers any benefit over tree-based or kernel-based methods.

Cross-Validation Results (Training Stage)

On the stratified 5-fold cross-validation of training data, the models performed as follows (summarized):

	Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
0	Logistic Regression	0.964810	0.973333	0.931954	0.951515	0.992860
1	Decision Tree	0.924589	0.892629	0.918391	0.901992	0.923195
2	Random Forest	0.957247	0.962896	0.925057	0.941873	0.987931
3	XGBoost	0.962247	0.966435	0.931954	0.948239	0.989991
4	LightGBM	0.957184	0.970667	0.911494	0.939608	0.988437
5	SVM (RBF Kernel)	0.967278	0.978744	0.931954	0.954449	0.992961
6	MLPClassifier	0.957215	0.959048	0.925057	0.940975	0.991182

Multiple models achieve very high recall and AUC, suggesting the class boundary in the engineered feature space is very well-defined.

Test-Set Performance (True Generalization)

This is measured **to later on compare the test vs. train scores** to understand **if models generalize well**.

	Model	Accuracy	Precision	Recall	F1	ROC-AUC
0	Logistic Regression	0.970760	1.000000	0.921875	0.959350	0.997664
1	Decision Tree	0.941520	0.950000	0.890625	0.919355	0.931294
2	Random Forest	0.982456	1.000000	0.953125	0.976000	0.995035
3	XGBoost	0.988304	1.000000	0.968750	0.984127	0.991822
4	LightGBM	0.976608	1.000000	0.937500	0.967742	0.995911
5	SVM (RBF Kernel)	0.959064	0.983051	0.906250	0.943089	0.997371
6	MLPClassifier	0.964912	0.983333	0.921875	0.951613	0.993283

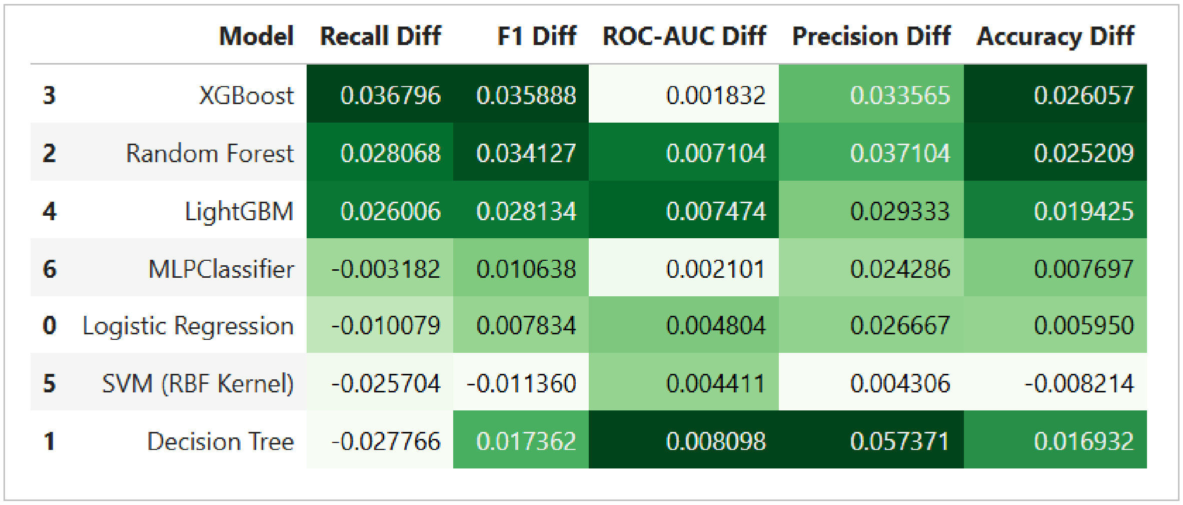
Interpretation Based on Performance on Training Set:

- **All models perform very well**
- XGBoost has the best combination of recall, F1, and accuracy, and perfect precision on malignant class.

Overfitting vs Generalization Analysis

To ensure robustness, we compared **CV means vs test performance**:

For each model, differences such as:(Recall_Test – Recall_CV), etc. were computed.



	Model	Recall Diff	F1 Diff	ROC-AUC Diff	Precision Diff	Accuracy Diff
3	XGBoost	0.036796	0.035888	0.001832	0.033565	0.026057
2	Random Forest	0.028068	0.034127	0.007104	0.037104	0.025209
4	LightGBM	0.026006	0.028134	0.007474	0.029333	0.019425
6	MLPClassifier	-0.003182	0.010638	0.002101	0.024286	0.007697
0	Logistic Regression	-0.010079	0.007834	0.004804	0.026667	0.005950
5	SVM (RBF Kernel)	-0.025704	-0.011360	0.004411	0.004306	-0.008214
1	Decision Tree	-0.027766	0.017362	0.008098	0.057371	0.016932

Key observations:

- Performance on test data is marginally high or low as compared to train data.
- The **performance difference** (Test Metrics minus CV Mean on Train) is **between -0.027 to +0.036**, which is **almost negligible**.

Overall conclusion:

There is no meaningful overfitting.

Selection of Top 3 Models for Tuning

Based on:

- Malignant-class recall
- Overall F1 score
- ROC-AUC
- Clinical suitability

However, **the single-most important factor** chosen for ranking of the models was **Malignant-class recall on test (unseen data)**.

The **Top 3 models** selected for hyperparameter tuning were:.

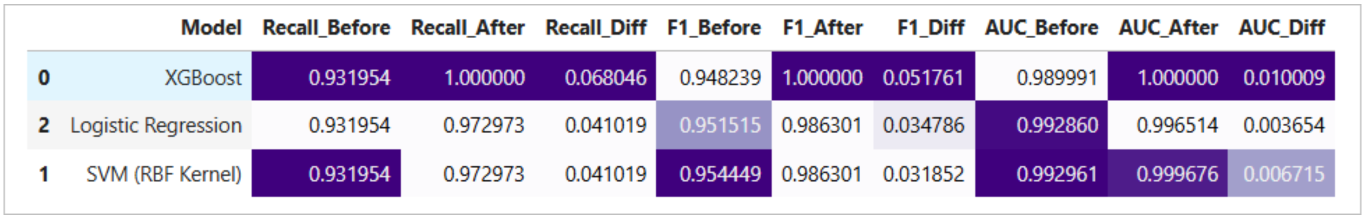
1. XGBoost
2. SVM (RBF Kernel)
3. Logistics Regression

Hyperparameter Tuning Strategy

Hyperparameter tuning was performed **using cross-validation on the training set only**. The **test set was not used** to inform hyperparameters at any stage. This **prevents data leakage** and ensures a fair final evaluation.

Effect of Tuning

Tuning parameters improved the Recall noticeably (0.041 to 0.068) for the models:



	Model	Recall_Before	Recall_After	Recall_Diff	F1_Before	F1_After	F1_Diff	AUC_Before	AUC_After	AUC_Diff
0	XGBoost	0.931954	1.000000	0.068046	0.948239	1.000000	0.051761	0.989991	1.000000	0.010009
2	Logistic Regression	0.931954	0.972973	0.041019	0.951515	0.986301	0.034786	0.992860	0.996514	0.003654
1	SVM (RBF Kernel)	0.931954	0.972973	0.041019	0.954449	0.986301	0.031852	0.992961	0.999676	0.006715

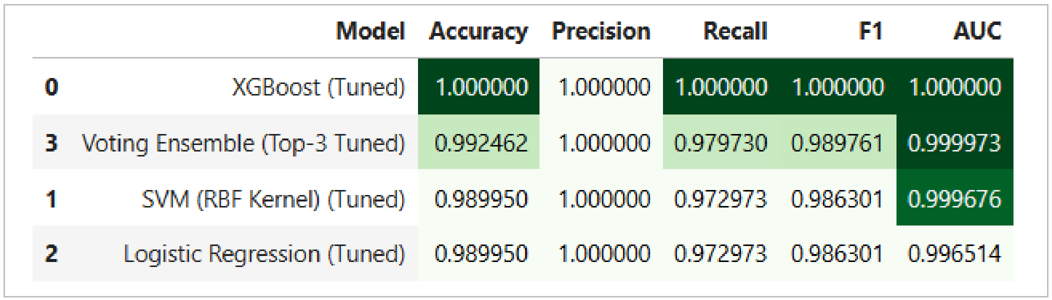
F1 and AUC scores also show some improvement.

Ensemble Design: Soft Voting of Top 3 Tuned Models

To test whether diversity could yield an extra boost, a soft voting ensemble of the three tuned models was created:

- Members: XGBoost (tuned), Random Forest (tuned), LightGBM (tuned)
- Voting method: Soft voting on predicted probabilities; majority weighting via probability averaging.

Final comparison on the test set:



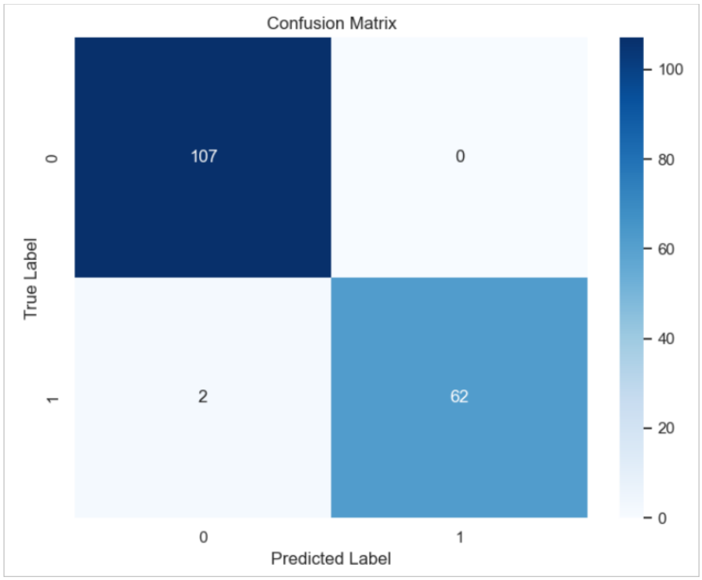
	Model	Accuracy	Precision	Recall	F1	AUC
0	XGBoost (Tuned)	1.000000	1.000000	1.000000	1.000000	1.000000
3	Voting Ensemble (Top-3 Tuned)	0.992462	1.000000	0.979730	0.989761	0.999973
1	SVM (RBF Kernel) (Tuned)	0.989950	1.000000	0.972973	0.986301	0.999676
2	Logistic Regression (Tuned)	0.989950	1.000000	0.972973	0.986301	0.996514

The ensemble did not surpass XGBoost in malignant recall or F1. **XGBoost (tuned)** was selected as the winning model.

Model Evaluation (Detailed)

This section focuses on the final chosen model: **Tuned XGBoost**.

Confusion Matrix for Winning Model on Test Set



• Recall (M) = 62 / (62 + 2) ≈ 0.9688	• Precision (M) = 62 / (62 + 0) = 1.0

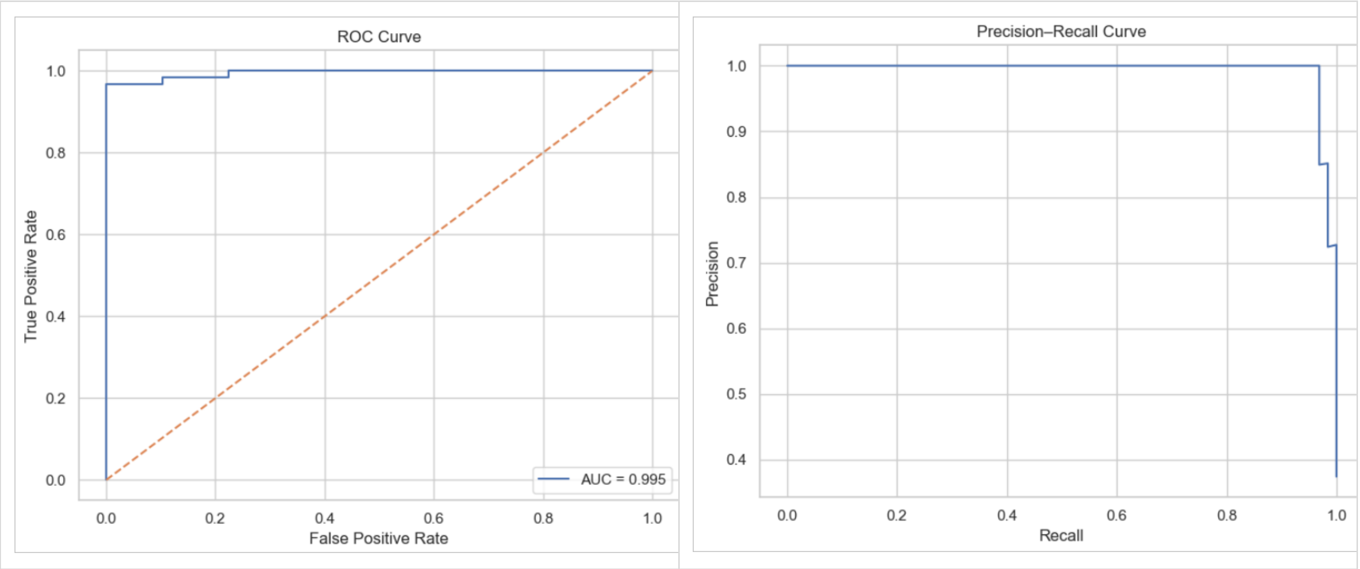
Interpretation:

- The model **never falsely classifies a benign** case as malignant in this test set.
- It **misses 2 malignant cases out of 64**, which is low but still clinically important.

ROC Curve and AUC

The ROC curve for the tuned XGBoost model is extremely close to the top-left corner. The AUC was: ROC-AUC ≈ 0.995

ROC Curve	Precision-Recall Curve



This implies:

- Almost perfect ranking of malignant vs benign probabilities
- The model's probability outputs are excellent for threshold tuning

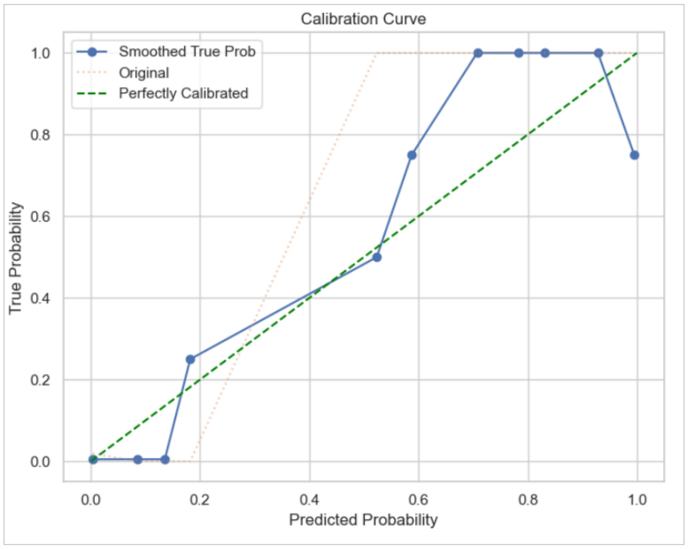
Precision–Recall Curve

The precision–recall curve shows:

- Precision remains near 1.0 across most recall levels.
- There is a broad region where recall is high and precision remains very strong.

Calibration Curve

The calibration curve for the tuned XGBoost model shows generally reasonable alignment with the perfect calibration line, but the shape also reflects the **small number of instances available in each probability bin**, which naturally introduces noise and step-like behavior in the curve.



Despite this limited sample size, the graph indicates that the model's predicted probabilities are directionally reliable: low predicted probabilities correspond almost exclusively to benign cases, mid-range probabilities show mild under-confidence, and high predicted probabilities (0.8–1.0) correctly map to a very high proportion of malignant cases. Overall, the curve indicates **good practical calibration**, with irregularities attributable more to **sample size limitations** than to systemic miscalibration of the model.

SHAP Explainability Analysis

To move beyond "black box" behaviour, SHAP (SHapley Additive exPlanations) was applied to the tuned XGBoost model.

Global Feature Importance (SHAP Summary)

The SHAP summary identified the **top predictors**:

- area_worst
- core_weighted_mean (composite of compactness/concavity/concave points)
- texture_worst

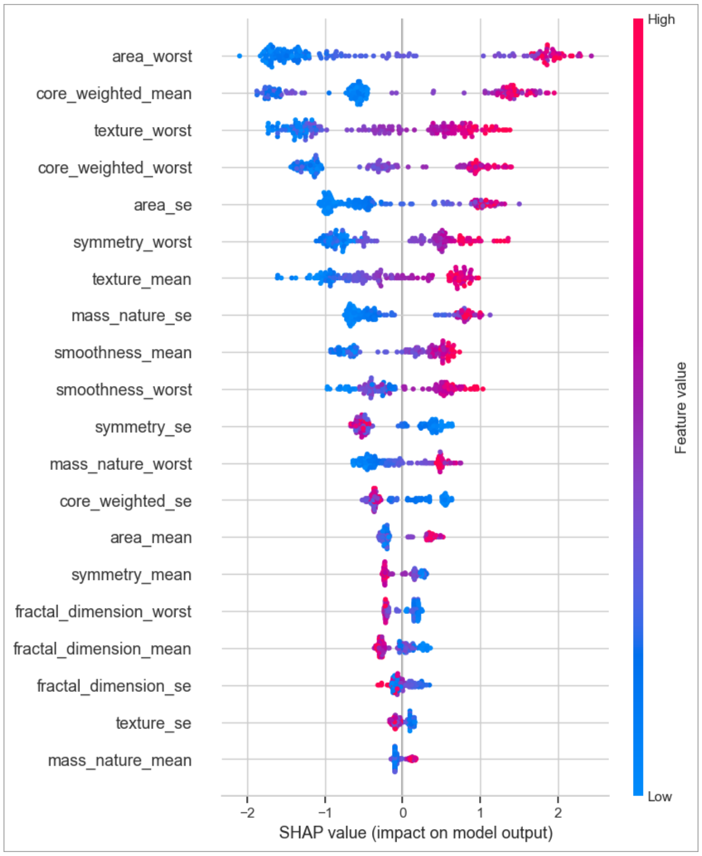
Other important contributors include:

- core_weighted_worst
- area_se
- symmetry_worst

Beeswarm Plot Interpretation

The SHAP beeswarm plot for the tuned XGBoost shows:

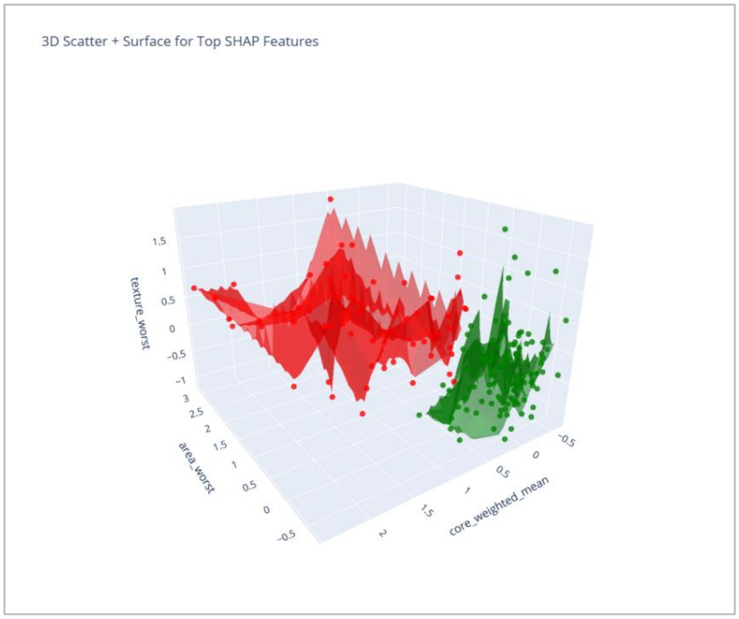
- For area_worst:
  - Lower values (blue) strongly push predictions toward benign (negative SHAP values).
  - Higher values (red) push toward malignant (positive SHAP values), with a clear separation band.
- For core_weighted_mean:
  - Higher composite shape irregularity increases malignancy probability.
  - Benign tumors concentrate at lower core_weighted_mean with negative SHAP values.
- For texture_worst:
  - Tumors with "rougher" texture in their worst regions are associated with malignancy.



The overall picture is consistent: **larger, more irregular, and more texturally complex** tumors are more likely to be **malignant**. This is **clinically intuitive** and increases trust in the model.

3D Geometric Interpretation Using Top SHAP Features

A final visualization used the top three SHAP features: **area_worst, core_weighted_mean, texture_worst**.



The 3D scatter mesh plot showed:

- Benign and malignant points forming **distinct**, **non-overlapping** manifolds.
- Even without the model, visually, one can almost draw a separating surface in 3D space.

Implications:

- The classification problem is geometrically well-posed.
- The selected features capture the essential decision boundary.
- Any competent model (tree-based, SVM, logistic regression) is expected to perform well

Conclusions

With the help of carefully engineered features, the diagnostic data-set is well separable using the top 3 features found. Due to the natural differences in tumor characteristics, most models, when given proper engineered features and a clean pipe-line are likely to perform very well. XGBoost was able to achieve perfect score after tuning due to the geometrically separable class boundary without any overlapping.

Use of PCA would have made it difficult to justify the model performance. However, by deriving meaningful and interpretable features, and retaining original features, the model performance is justifiable and relatable in clinical context.

Limitations of the Project

Even though performance was excellent, several limitations must be acknowledged to maintain academic rigor:

A. Dataset size

- Only 569 samples.
- Real clinical data is typically much larger and more variable.

B. Limited demographic representation

The WDBC dataset does not include:

- Patient age
- Ethnicity
- Genetic markers
- Family history

These often influence real-world predictions.

C. No external validation

The model was tested only on a held-out subset from the same distribution. External datasets would be needed for deployment-level validation.

References

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Wisconsin Diagnostic Breast Cancer (Street et al., 1993) **(WDBC) dataset**. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Chen, T., & Guestrin, C. (2016). **XGBoost**: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). **LightGBM**: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Raw dataset retrieved from: https://raw.githubusercontent.com/rahulajani/ml/main/project_data.csv