Breast Cancer Classification using XGBoost

Data Mining Presentation

Reza Ahmadizadeh

Faculty of Engineering and Technology
University of Mazandaran

November 3, 2025

OUTLINE

1. Introduction
2. Dataset
3. Exploratory Data Analysis (EDA)
4. Paper introduction and basic concepts
5. Feature Selection in action
6. Model Training and Results
7. Comparison and Summary
8. Conclusion

Introduction

Importance of the topic

Why breast cancer is important?

- Second cause of death in women (After lung cancer)
- About 8% of women are diagnosed with breast cancer during their lifetime
- The five-year survival rate is only 14% (in advanced cases)

What role does data mining plays?

- Early and accurate detection using artificial intelligence can be life-saving
- To build a classification model to detect cancerous (Malignant) cells from non-cancerous (Benign) cells

Dataset

Wisconsin Breast Cancer Diagnostic (WDBC)

Specifications

- **Samples:** 569 rows, each representing a digitized image of a breast mass.
- **Features:** After removing `id` and `Unnamed:`, 30 features remain.
- **Target variable:** `diagnosis` (M = Malignant, B = Benign).
- **Class distribution:** 357 Benign (62.7%), 212 Malignant (37.3%).

Feature source:

- Extracted from FNA (Fine Needle Aspirate) cell images

Dataset (cont.)

Feature types (Data Construction concepts)
For each cellular attribute (e.g., radius, texture, perimeter), three values are computed:

- Mean
- SE (Standard Error)
- Worst

Example: `radius_mean, texture_se, area_worst`.

Exploratory Data Analysis (EDA)

Understand data distribution and identify strong predictive features

Which features are most discriminative?

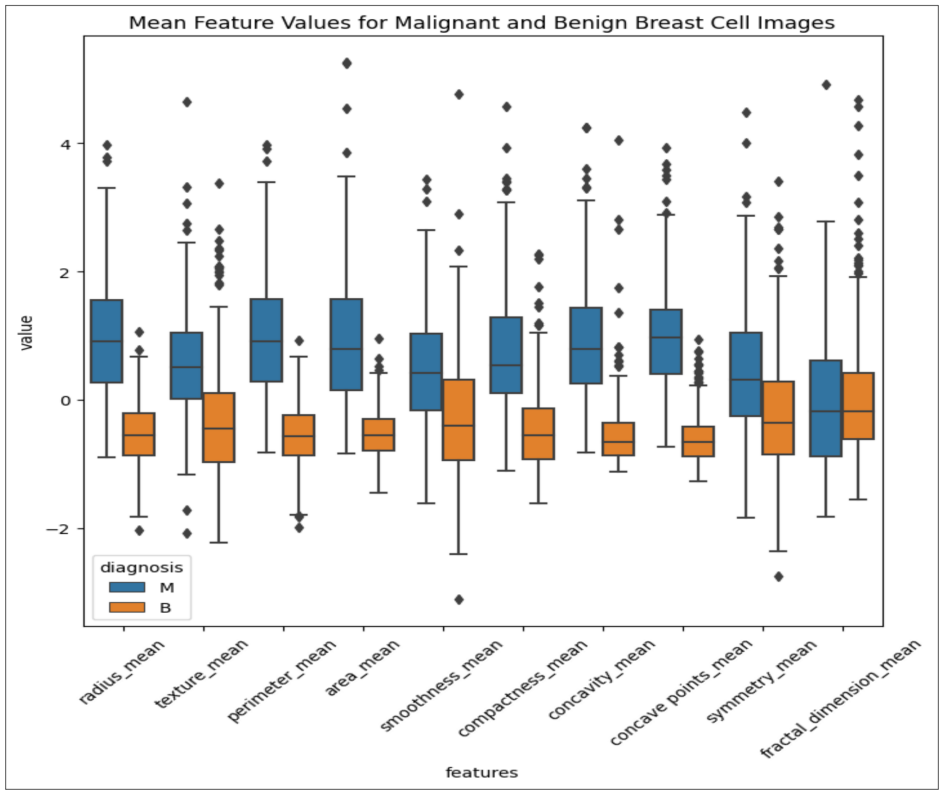

Figure 1: Mean feature values for malignant and benign breast cells.

Exploratory Data Analysis (cont.)

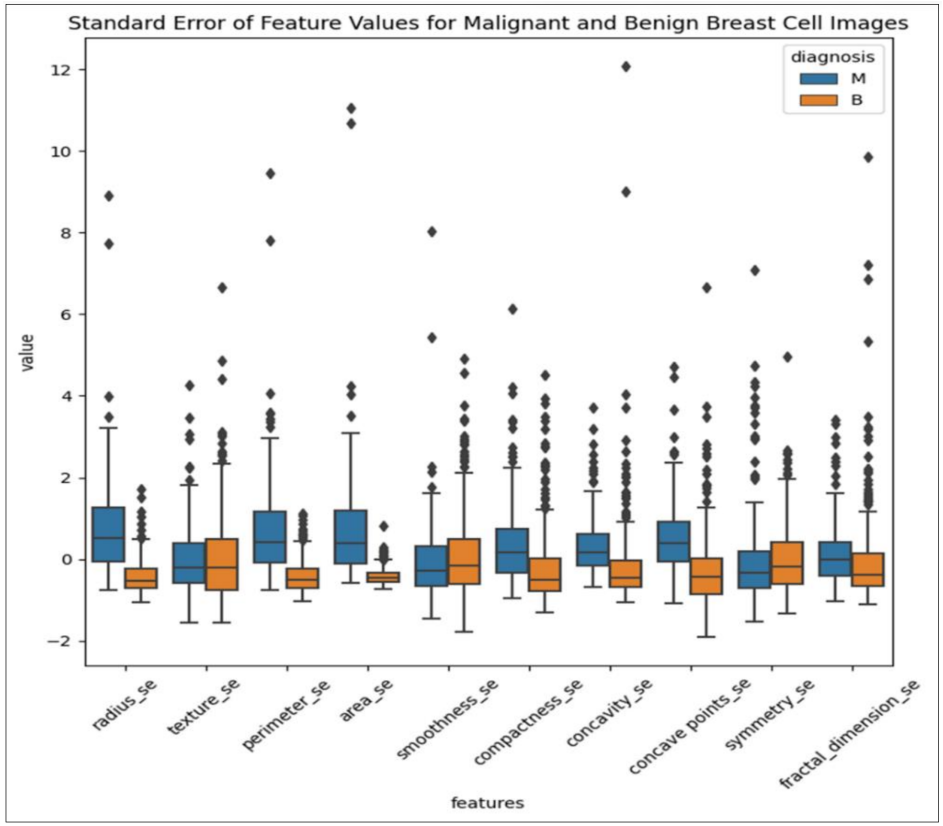

Figure 2: Standard error of feature values.

Exploratory Data Analysis (cont.)

Size-related features show clear separation

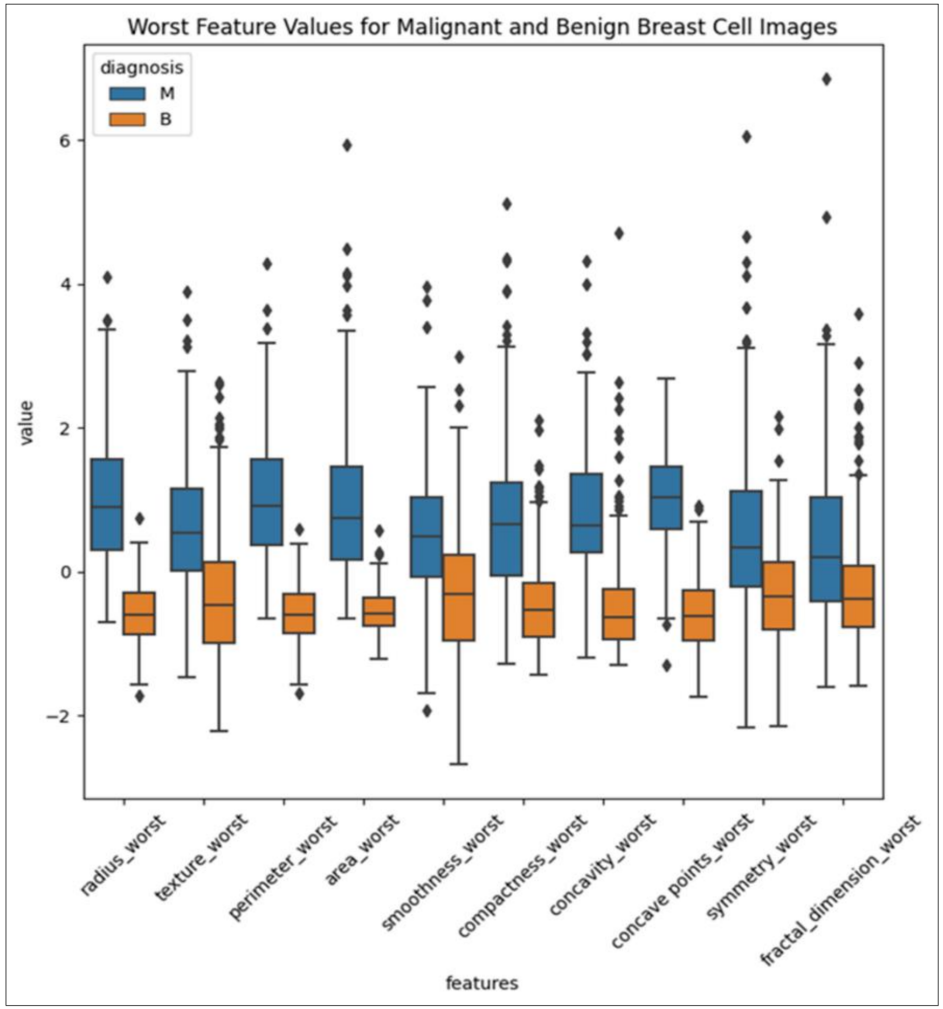

Figure 3: Worst feature values for malignant and benign breast cell images.

Exploratory Data Analysis (cont.)

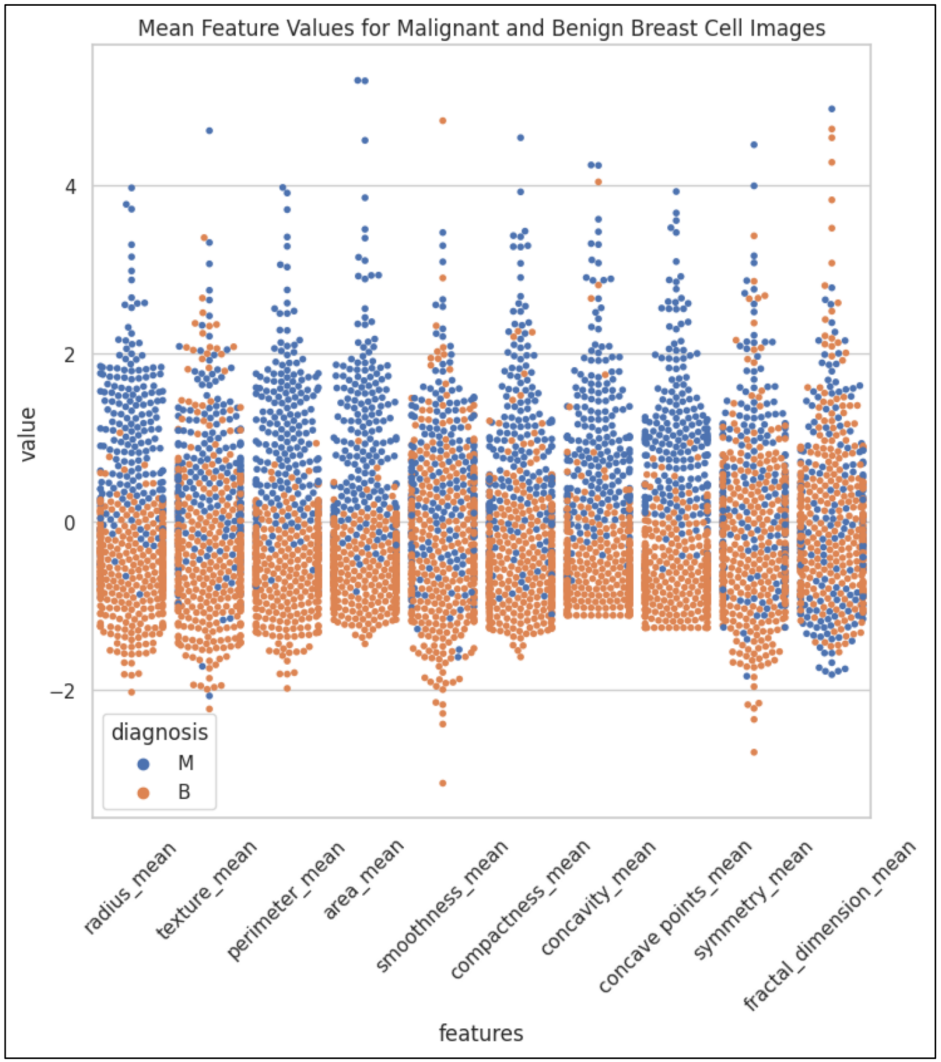

Figure 4: Swarm plots of mean feature values.

Exploratory Data Analysis (cont.)

How do features relate to each other

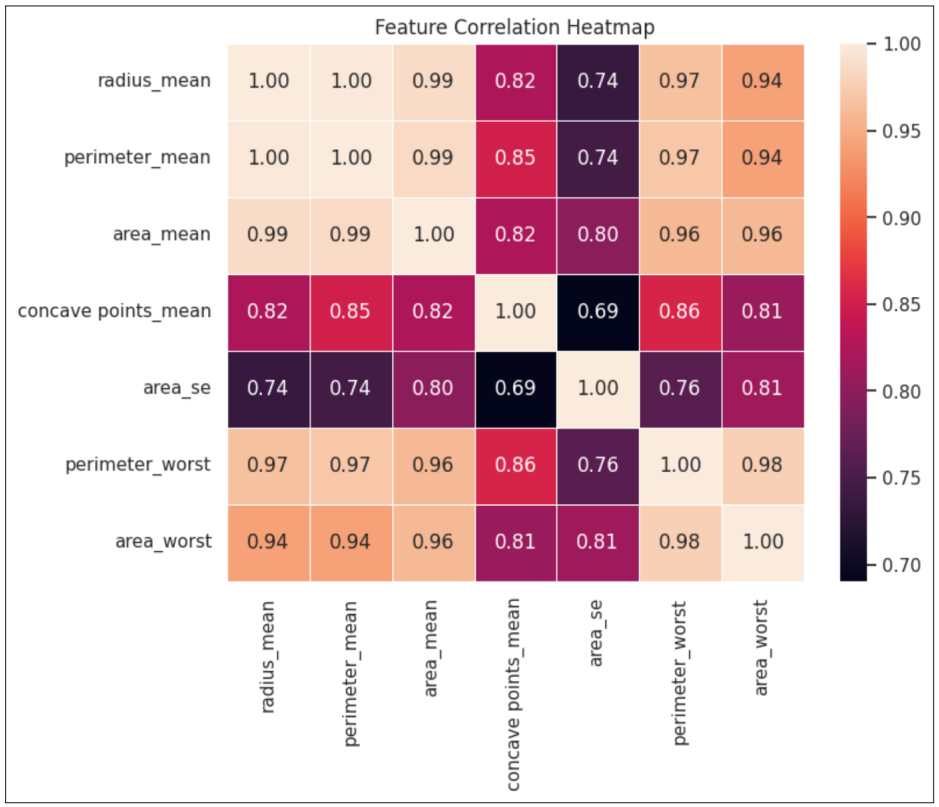

Figure 5: Correlation Heatmap Reveals multicollinearity in size measurements.

Paper introduction and basic concepts

Breast Cancer Classification using XGBoost" (2024) - Hoque et al

Paper objective:
To use the XGBoost algorithm for binary classification of breast cancer

Whats is XGBoost?

- Short for: Extreme Gradient Boosting
- Family: Ensemble Learning $\rightarrow$ Gradient Boosting family
- Main idea: Sequentially correct errors
- Strengths: Speed, accuracy, and automatic feature selection

Paper introduction and basic concepts (cont.)

How it works?

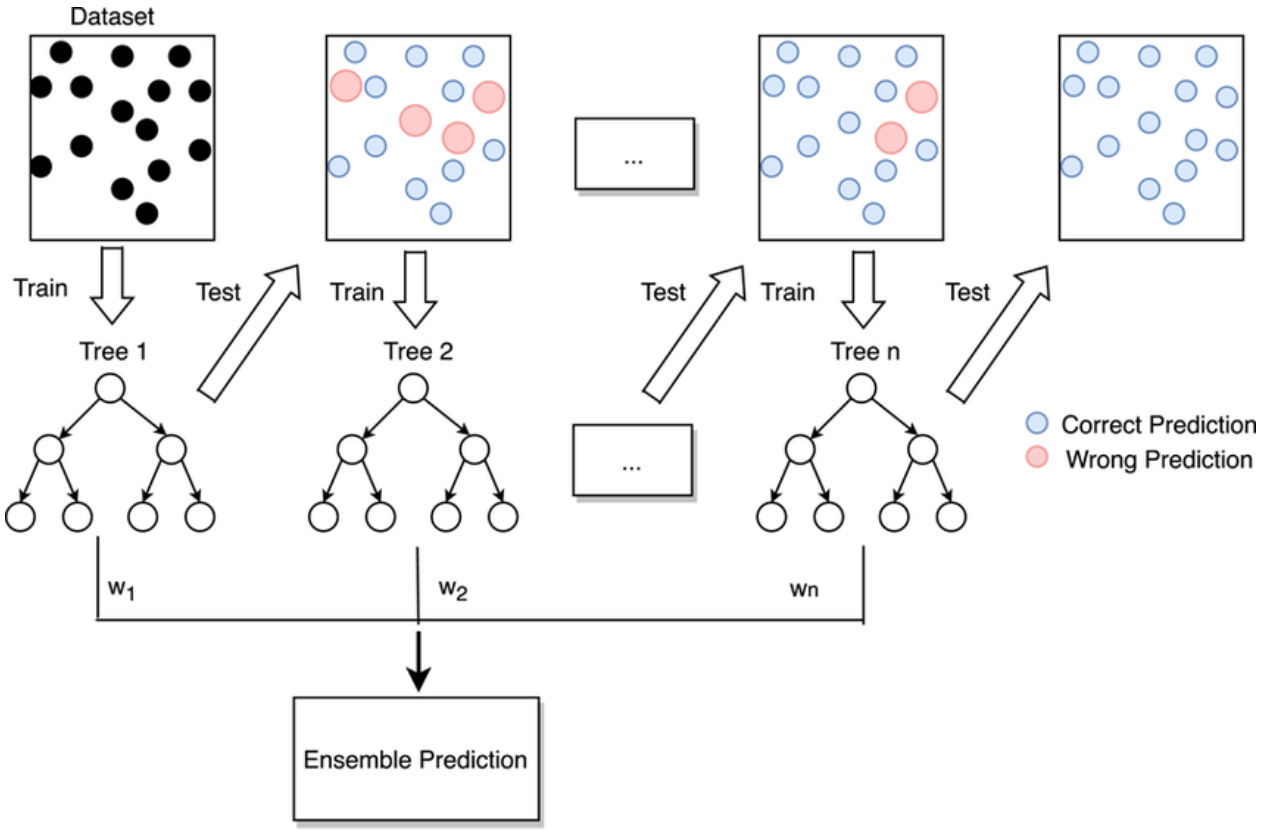

Figure 6: Gradient boosting algorithm

Paper introduction and basic concepts (cont.)

Whats the Extreme in gradient boosting?

- **Gradient:** Uses gradient descent to quickly reduce errors when building the next tree, making the process efficient.
- **Extreme:** Engineering optimizations make it very fast and capable of handling large datasets with many features.

Advantages for this dataset:

- Handles complex patterns
- Resistant to multicollinearity
- Provides feature importance

Feature Selection in action

Methods used

combined approach

- **Filter Method (initial):** Used boxplots and heatmaps to spot strong and correlated features
- **Embedded Method (main):** XGBoost assigns importance scores (F-score) to features after training

Feature Selection in action (cont.)

Feature Importance Results

- Key features: `concave points_worst` and `area_worst` (F-score = 5).
- This matches EDA results highlighting the importance of cell size.

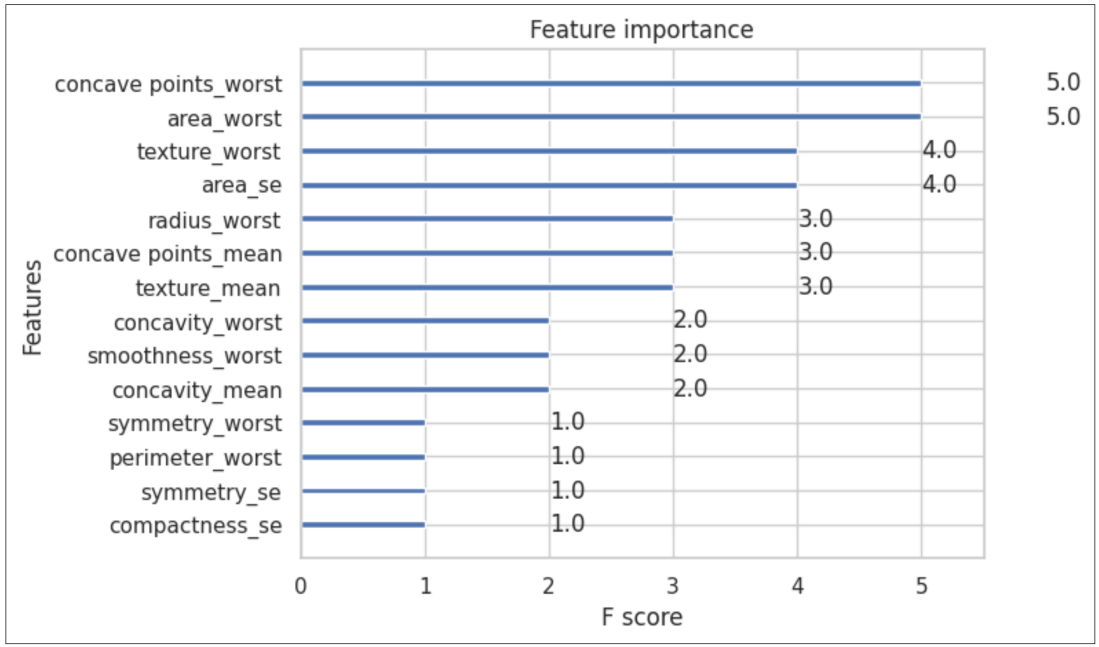

Figure 7: Feature ranking and F score

Feature Selection in action (cont.)

DimReduction

Although classic dimensionality reduction (like PCA) was not used, this feature selection helped reduce model complexity.

Advantages:

- Less noise
- Lower risk of overfitting
- Faster training

Model Training and Results

Model setup:

- Data split: 80% train, 20% test
- XGBoost hyperparameters: `learning_rate=0.3`, `max_depth=4`

Test set evaluation:

- Accuracy: 94.74%
- Precision: 90.91%
- Recall: 95.24% $\rightarrow$ critical in medicine to avoid missing patients
- F1-Score: 93.02%

Comparison and Summary

Comparison with previous works

Model	Accuracy (similar datasets)
SVM	83.3%
Random Forest	75%
Logistic Regression	94.4%
XGBoost (our paper)	94.74%

Table 1: Comparison of model accuracy on similar breast cancer datasets.

Conclusion

- XGBoost is an excellent choice for this classification task
- The data mining process (EDA $\rightarrow$ Feature Selection $\rightarrow$ Model Training) was correctly executed
- Future work suggested by the paper: use larger datasets and extend the model to predict cancer stage.

REFERENCES

Cruz, J. A. and Wishart, D. S. (2006).
Applications of machine learning in cancer prediction and prognosis.
Cancer informatics, 2:59–77.

Hoque, R., Das, S., Hoque, M., and Haque, E. (2024).
Breast Cancer Classification using XGBoost.
World Journal of Advanced Research and Reviews, 21(02):1985–1994.

Repository, U. M. L. (2024).
Breast Cancer Wisconsin (Diagnostic) Data Set.
Accessed: February 2024.

Tan, A. C. and Gilbert, D. (2003).
Ensemble machine learning on gene expression data for cancer classification.
In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 145–150.