

An Intelligent Machine Learning Framework for Early Breast Cancer Classification

Naga Deepti Ponnaganti¹, Vema Manisha²

¹Professor, ²PG Scholar,

Department of Computer Science and Engineering, Sir C. R. Reddy College of Engineering,
Eluru, Andhra Pradesh, India

Email id : deepthi.viggeswarapu@gmail.com¹ , vemmanisha29@gmail.com²

Article Received: 02 May 2025

Article Accepted: 30 May 2026

Article Published: 20 June 2026

Citation

Naga Deepti Ponnaganti, Vema Manisha, "An Intelligent Machine Learning Framework for Early Breast Cancer Classification", Journal of Next Generation Technology (ISSN: 2583-021X), 6(3), pp. 44-53. June 2026.

Abstract

Breast cancer classification is an important machine-learning application in which computational models can be used to classify diagnostic observations into benign and malignant categories. This study presents an intelligent machine learning framework for early breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The framework integrates data validation, exploratory analysis, preprocessing, model development, comparative evaluation, and feature-level interpretation. Six supervised learning approaches—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting—are evaluated using a stratified 80:20 train-test split and five-fold stratified cross-validation. Performance is assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Permutation-based feature importance is employed to provide a model-agnostic view of influential variables. On the held-out test set, SVM and Random Forest achieved 0.9737 accuracy, while Logistic Regression produced the highest test ROC-AUC of 0.9960. Five-fold cross-validation produced a mean ROC-AUC of 0.9953 for Logistic Regression and 0.9945 for SVM. The results demonstrate the feasibility of a reproducible comparative machine-learning workflow on the WDBC dataset. The framework is intended as an academic decision-support prototype and does not replace pathological examination or professional clinical judgment.

Keywords: Breast cancer classification, Machine learning, Supervised Learning, Wisconsin Diagnostic Breast Cancer, Explainable Machine Learning, Decision Support.

I. Introduction

Breast cancer classification is a representative biomedical machine-learning problem in which structured diagnostic measurements are mapped to a target class. In this study, the target classes are benign and malignant. The objective is to investigate how standard supervised learning algorithms can process diagnostic features and provide reproducible classification

outputs rather than to establish a clinical diagnosis. Earlier work on breast cancer diagnosis established benchmark datasets and demonstrated that numerical measurements extracted from fine-needle aspirates can support automated classification [1–3]. Logistic Regression provides a simple probabilistic baseline, while Decision Trees provide directly inspectable decision rules. Random Forest extends tree-based learning through randomized ensembles [5]. Support Vector Machines provide maximum-margin classification and can model nonlinear boundaries through kernels [6]. K-nearest neighbours provides an instance-based alternative [8], and Gradient Boosting constructs an additive sequence of weak learners [7].

Scikit-learn provides a widely used implementation environment for supervised learning and reproducible preprocessing pipelines [4]. For health-related machine learning, interpretability is also important because a predictive score alone does not explain how a model arrived at its output. In this study, permutation importance is used as a model-agnostic interpretation method [10].

The gap addressed here is therefore not a new mathematical classifier, but a reproducible comparative framework. Many student-level studies focus on a single algorithm or a single accuracy value. The present work evaluates six models under the same data partition, reports multiple metrics, uses five-fold stratified cross-validation, and includes feature-level interpretation.

Machine learning is useful for pattern recognition because it can estimate relationships between multiple input variables and a known target using previously labelled observations. A well-designed academic framework should separate data preparation, model fitting, evaluation, and interpretation. Such separation reduces accidental data leakage and makes model comparison more meaningful [16].

The Wisconsin Diagnostic Breast Cancer dataset contains numerical measurements computed from digitized images of fine-needle aspirate samples. The observations are labelled as benign or malignant, allowing a supervised classification experiment without requiring an image-processing pipeline. The problem can therefore be represented as a mapping $f(X) \rightarrow y$, where X is a vector of diagnostic features and y is the binary class.

The contribution of this work is methodological: it combines data validation, exploratory analysis, algorithm comparison, held-out testing, cross-validation, ROC analysis, confusion-matrix analysis, and permutation-based feature interpretation in one reproducible workflow. It does not introduce a new classifier architecture.

II. Frame work

A. Data Set

The WDBC dataset used in the implementation is distributed through the scikit-learn dataset interface. It contains 569 observations, 30 numerical input features, and two target classes. There are 357 benign observations and 212 malignant observations. The features are grouped

around radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, with mean, standard-error, and worst-case measurements.

For this study, the target is encoded as 1 for malignant and 0 for benign. The explicit positive-class encoding simplifies the interpretation of precision, recall, ROC-AUC, and confusion-matrix cells.

Property	Value
Dataset	Wisconsin Diagnostic Breast Cancer (WDBC)
Observations	569
Input features	30
Target classes	2
Malignant cases	212
Benign cases	357
Missing values	None in the loaded dataset
Feature type	Numerical
Learning task	Binary supervised classification
Train/test split	80% / 20%, stratified
Random seed	42

Table 1. Dataset characteristics

B. Data Set Class Distribution

Exploratory analysis was conducted before model interpretation to understand class distribution, feature scale differences, and relationships among variables. The dataset is not perfectly balanced because benign observations outnumber malignant observations. Stratification was therefore used during both train-test partitioning and cross-validation.

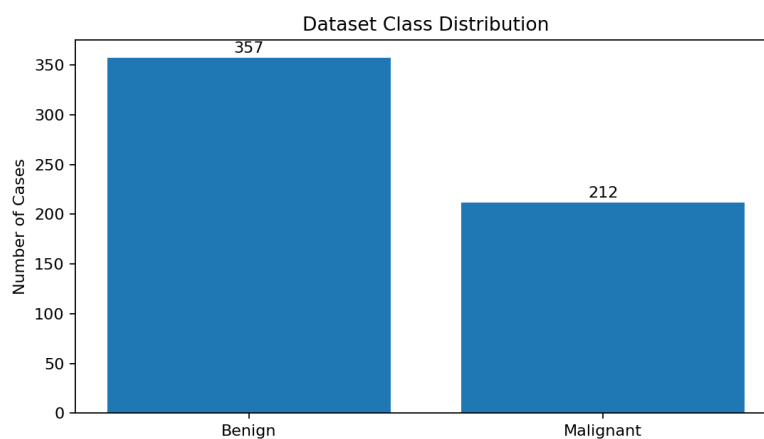


Fig. 1. Class distribution of the WDBC dataset.

The feature correlation matrix indicates that several variables are strongly correlated because multiple measurements describe related geometric properties. High correlation does not necessarily imply that a feature is redundant for every model, but it means that feature importance should be interpreted cautiously.

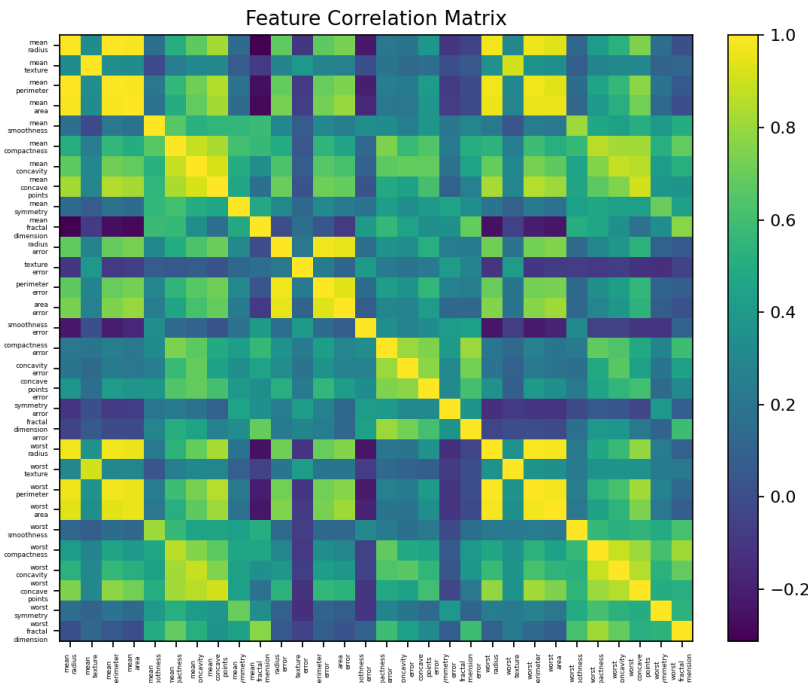


Fig. 2. Feature correlation matrix.

C. System Analysis

An 80:20 stratified train-test split with random state 42 was used. The training partition was used for model fitting and the held-out test partition was reserved for final comparison. Logistic Regression, SVM, and KNN were implemented in pipelines containing StandardScaler. Scaling parameters were learned only from the training data. Tree-based models were trained directly because their splitting rules do not require feature standardization [14,15].

D. System Architecture

The proposed architecture is organized as a sequence of data, modelling, evaluation, and interpretation stages. The dataset layer supplies the diagnostic observations. The preprocessing layer validates the input and applies scaling where required. The modelling layer contains six independent classifiers. The evaluation layer computes standardized metrics and visualizations, while the explainability layer produces feature-importance information. The final output reports the classification result together with evaluation evidence.

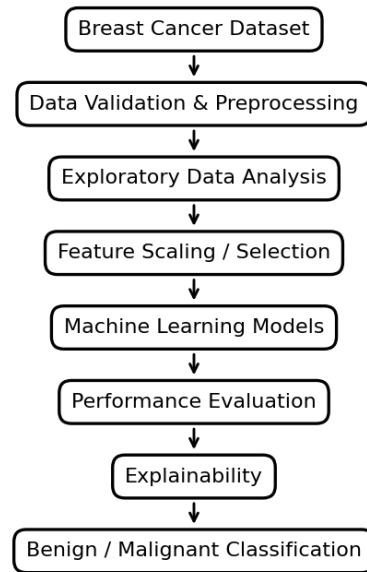


Fig. 3. Proposed intelligent machine learning framework

E. Model Configuration

Model	Configuration	Role
Logistic Regression	StandardScaler; max_iter=5000	Interpretable linear baseline
Decision Tree	max_depth=5; random_state=42	Rule-based nonlinear model
Random Forest	n_estimators=300; random_state=42	Nonlinear ensemble
Support Vector Machine	StandardScaler; RBF kernel; probability=True	Margin-based nonlinear model
K-Nearest Neighbours	StandardScaler; n_neighbors=7	Distance-based classifier
Gradient Boosting	random_state=42	Sequential nonlinear learner

Table 2. Model configurations

Five complementary metrics were used. Accuracy measures the proportion of correct predictions. Precision measures the proportion of predicted malignant observations that are actually malignant. Recall measures the proportion of malignant observations correctly identified. F1-score is the harmonic mean of precision and recall. ROC-AUC summarizes ranking performance over classification thresholds. No single metric is treated as sufficient on its own.

III. Results and Discussion

The reported results are computational results produced from the scikit-learn WDBC dataset using the implementation and parameters described in the source report. They represent benchmark results for the stated academic experiment and are not clinical validation results.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.9649	0.9750	0.9286	0.9512	0.9960
Gradient Boosting	0.9649	1.0000	0.9048	0.9500	0.9947
Support Vector Machine	0.9737	1.0000	0.9286	0.9630	0.9947
Random Forest	0.9737	1.0000	0.9286	0.9630	0.9944
K-Nearest Neighbours	0.9561	0.9744	0.9048	0.9383	0.9825
Decision Tree	0.9211	0.9459	0.8333	0.8861	0.9448

Table 3. Held-out test-set performance

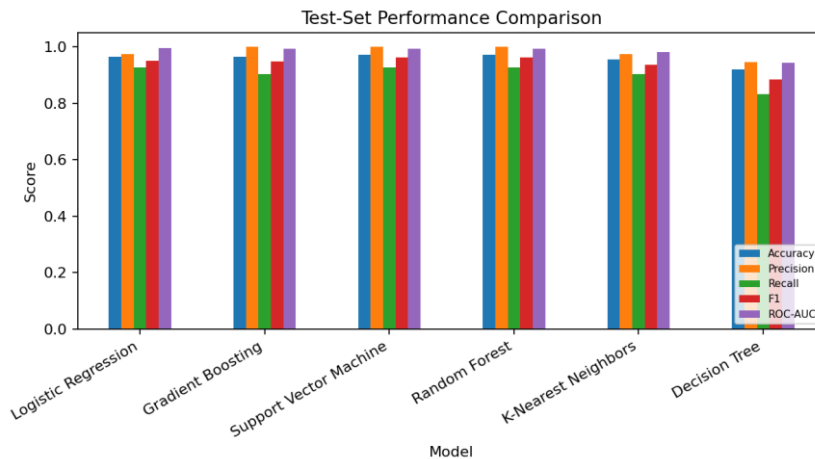


Fig. 4. Test-set performance comparison and ROC curves.

A. ROC Analysis

For the model selected by the highest test ROC-AUC in the documented experiment, Logistic Regression, the confusion matrix contains 71 true negatives, 1 false positive, 3 false negatives, and 39 true positives. False negatives correspond to malignant observations classified as benign. In a real screening-oriented setting, the consequences of such errors would

require domain-specific validation and cannot be inferred from this academic experiment alone.

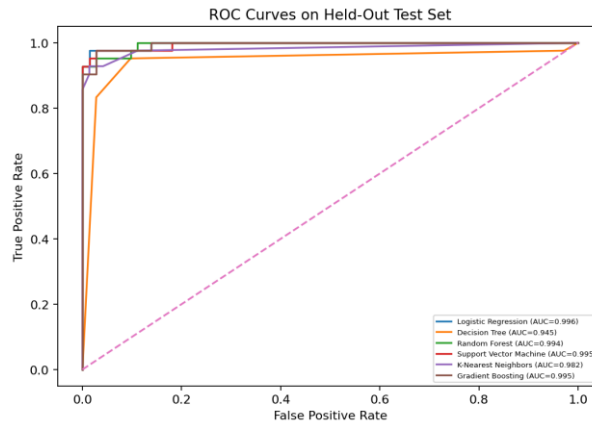


Fig. 5. Confusion matrix for Logistic Regression.

B. Five-Fold Cross Validation

Model	CV Accuracy	CV Precision	CV Recall	CV F1	CV ROC-AUC
Logistic Regression	0.9737	0.9863	0.9436	0.9633	0.9953
Support Vector Machine	0.9772	0.9810	0.9576	0.9688	0.9945
Gradient Boosting	0.9508	0.9569	0.9107	0.9315	0.9927
Random Forest	0.9526	0.9461	0.9295	0.9355	0.9895
K-Nearest Neighbours	0.9649	0.9856	0.9200	0.9505	0.9863
Decision Tree	0.9280	0.9294	0.8779	0.8992	0.9068

Table 4. Five-fold stratified cross-validation results

Under the fixed 80:20 split, Logistic Regression achieved the highest test ROC-AUC (0.9960). In five-fold cross-validation, SVM achieved 0.9772 mean accuracy and 0.9688 mean F1-score, while Logistic Regression achieved 0.9953 mean ROC-AUC. These observations are dataset- and protocol-specific and should not be interpreted as evidence of universal model superiority.

C. Permutation feature importance — Logistic Regression

Permutation importance was estimated by shuffling one feature at a time and measuring the reduction in ROC-AUC. Larger reductions indicate that the tested model relied more heavily on the feature under the experimental data. The source report identifies worst symmetry, worst texture, and worst concavity among the higher-ranked variables in the Logistic Regression permutation analysis. Because many WDBC variables are correlated, importance may be shared across related predictors and does not establish biological causation.

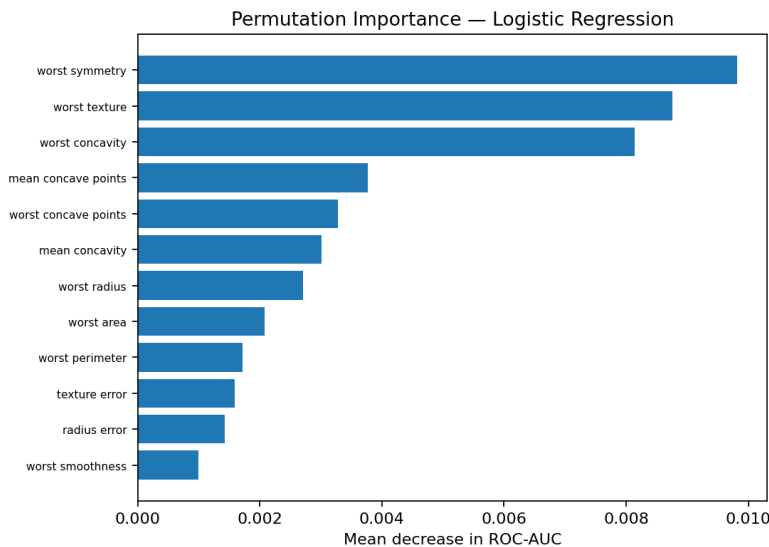


Fig. 6. Permutation feature importance for Logistic Regression.

D. Comparative Discussion

The comparative results show that different algorithms produce different performance profiles even when trained on the same observations. Logistic Regression provides a strong and relatively transparent baseline. Decision Tree offers direct rule-based structure but can be sensitive to model complexity. Random Forest captures nonlinear interactions through multiple trees. SVM provides a margin-based nonlinear alternative. KNN depends on local neighbourhood structure, and Gradient Boosting provides sequential nonlinear learning.

Accuracy alone can conceal class-specific behaviour, particularly when class counts differ. Therefore, recall for the malignant class, precision, F1-score, ROC-AUC, and the confusion matrix should be considered together. The cross-validation results also demonstrate why a single held-out split should not be treated as the only evidence of model behaviour.

The principal methodological value of the framework is reproducibility. A fixed random seed, explicit model configurations, stratified partitioning, algorithm-specific preprocessing, and documented evaluation metrics allow another researcher or student to rerun the experiment. The framework is modular and can be adapted to independent datasets for external validation.

IV. Conclusion and Future work

The study is limited by the use of a single public dataset and the absence of external validation. The dataset is relatively small and contains correlated predictors. Fixed model parameters were documented rather than exhaustively optimized. The study also contains no patient outcomes, treatment information, or prospective clinical data. Consequently, the reported metrics quantify performance on the specified dataset and experimental protocol rather than clinical utility.

The classifier should not be presented as an autonomous diagnostic tool. Responsible translation would require assessment of data provenance, population differences, potential bias, privacy and security, calibration, decision thresholds, clinical oversight, and prospective validation. The present work is an academic decision-support prototype.

This paper presented an intelligent machine learning framework for early breast cancer classification using the Wisconsin Diagnostic Breast Cancer dataset. The framework integrates validation, exploratory analysis, preprocessing, six supervised classifiers, held-out testing, five-fold cross-validation, ROC analysis, confusion-matrix analysis, and permutation-based feature importance. The experimental results demonstrate the feasibility of a reproducible classification workflow using standard machine-learning methods.

Future work includes evaluation on independent external breast-cancer datasets, systematic hyperparameter optimization using nested validation, calibration and decision-curve analysis, SHAP and other local explanation methods, comparison with imaging-based deep learning, subgroup and fairness analysis, and secure prototype development. Any prospective clinical validation should be performed with appropriate clinical and institutional oversight.

References

- [1]. W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Breast Cancer Diagnosis and Prognosis via Linear Programming," *Operations Research*, vol. 43, no. 4, 1995.
- [2]. W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine Learning Techniques to Diagnose Breast Cancer from Fine-Needle Aspirates," *Cancer Letters*, vol. 77, no. 2–3, 1994.
- [3]. W. H. Wolberg, W. N. Street, O. L. Mangasarian, et al., "Wisconsin Diagnostic Breast Cancer dataset documentation as distributed through the UCI/scikit-learn ecosystem."
- [4]. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [5]. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
- [6]. C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, 1995.
- [7]. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

- [8]. T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
- [9]. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, vol. 7, pp. 179–188, 1936.
- [10]. C. Molnar, Interpretable Machine Learning, online reference text covering model-agnostic interpretation methods including permutation importance.
- [11]. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer.
- [12]. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, Springer.
- [13]. Tirumanadham, N. S. K. M. K., Priyadarshini, V., Praveen, S. P., Thati, B., Srinivasu, P. N., & Shariff, V. (2025g). Optimizing Lung Cancer Prediction Models: A hybrid methodology using GWO and Random Forest. In Studies in computational intelligence (pp. 59–77). https://doi.org/10.1007/978-3-031-82516-3_3
- [14]. S Phani Praveen et al., (2025) "Enhanced Predictive Modeling For Alzheimer's Disease: Integrating Cluster-Based Boosting And Gradient Techniques With Optimized Feature Selection", Journal of Theoretical and Applied Information Technology, vol. 103, no. 8, April 2025
- [15]. S Phani Praveen et al., (2025), "AI- Powered Diagnosis: Revolutionizing Healthcare With Neural Networks", Journal of Theoretical and Applied Information Technology, vol. 103, no. 3, February 2025
- [16]. Rao, B. G., & Vindhya, A. S. (2026). A Multi-Metaheuristic feature selection and ensemble learning framework for intrusion detection systems (pp. 144–150). <https://doi.org/10.1109/icicds70526.2026.11604694>