Jurnal Kajian
Komputasi Inovatif Vol.17 No.2, Februari 2026

BREAST CANCER CLASSIFICATION USING MACHINE LEARNING

I Gede Wahyu Surya Dharma[1], I Gede Karang Komala Putra[2]
Universitas Bali Internasional
E-mail: suryadharma@iikmpbali.ac.id[1],
igdkarang@iikmpbali.ac.id[2]

Abstract

Breast cancer remains one of the most prevalent malignancies, where early and accurate diagnosis is critical to improve patient outcomes. This study investigates the performance of five supervised machine learning algorithms—Support Vector Machine (SVM), Decision Tree (DT), k-Nearest Neighbors (KNN), Random Forest (RF), and Logistic Regression (LR)—for automated breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The dataset contains 569 samples with 30 numerical features extracted from digitized fine needle aspirate (FNA) images, labeled as benign or malignant. The experimental protocol employs an 80/20 stratified train–test split, feature standardization for scale-sensitive models, and RandomizedSearchCV with 5-fold cross-validation for hyperparameter optimization. Models are evaluated using accuracy, precision, recall (sensitivity), specificity, F1-score, ROC-AUC, confusion matrices, and cross-validation statistics, complemented by approximate 95% confidence intervals and McNemar's test for pairwise comparison. The optimized SVM with radial basis function kernel achieves test accuracy of 98.25%, precision of 100%, recall of 95.24%, specificity of 100%, and ROC-AUC of 0.9960, outperforming other models with statistically significant improvements over DT and KNN. Feature importance analysis from tree-based models highlights "worst" size and shape descriptors (area_worst, perimeter_worst, radius_worst, concave_points_worst) as dominant predictors, aligning with cytopathological understanding of malignant nuclei. The results demonstrate that properly tuned traditional models can provide robust and interpretable performance for tabular medical data, and establish a reproducible baseline for future research in breast cancer classification.

Keywords: *Breast Cancer, Wisconsin Diagnostic Breast Cancer (WDBC), Support Vector Machine, Random Forest, Hyperparameter Optimization, Statistical Validation.*

1. INTRODUCTION

Cancer is a disease that continues to infect many people and is often difficult to recognize at early stages, making early detection crucial to reduce mortality and treatment failure. Breast cancer, in particular, is one of the most common cancers among women and remains a leading cause of cancer-related death worldwide. Diagnostic workflows typically rely on radiological imaging and histopathological assessment, which are subject to inter-observer variability and workload constraints. Hence, computer-aided diagnosis (CAD) systems based on machine learning have emerged as promising tools to support clinicians by providing quantitative and reproducible assessments.

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the UCI Machine Learning Repository is a widely used benchmark for evaluating CAD models (Wolberg et al., n.d.; Breast Cancer Wisconsin (Diagnostic) data documentation, n.d.). It consists of 569 FNA cases with 30 numerical features describing radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, each summarized as mean, standard error, and “worst” (largest) values (Breast Cancer Wisconsin (Diagnostic) data documentation, n.d.). A large body of work has applied classical classifiers such as SVM, KNN, DT, RF, and LR, as well as more recent ensemble and deep learning approaches, often reporting high accuracy (typically 92–99%) on WDBC and related datasets (Aamir et al., 2022; Aamir et al., 2025; Asri et al., n.d.; Enhancing breast cancer detection and classification using machine learning-based computer-aided diagnosis systems, 2023). However, many studies differ in preprocessing, feature selection, hyperparameter tuning, and validation strategies, making direct comparison difficult. In addition, some works rely mainly on accuracy and do not fully analyze sensitivity, specificity, false negatives, or model stability across folds (Asri et al., n.d.; Eftekharian et al., 2025).

This study aims to provide a rigorous and methodologically oriented comparison of five supervised learning algorithms—SVM, DT, KNN, RF, and LR—on the WDBC dataset under a unified experimental protocol. The contributions of this work are as follows:

1. Apply systematic hyperparameter optimization using RandomizedSearchCV with stratified 5-fold cross-validation for all models.
2. Evaluate models using a rich set of metrics, including accuracy, precision, recall, specificity, F1-score, ROC-AUC, confusion matrices, and approximate 95% confidence intervals.
3. Perform statistical analysis using cross-validation distributions and McNemar’s test to compare SVM with other classifiers on the test set.
4. Analyze feature importance in tree-based models to provide interpretable insights aligned with clinical understanding (Eftekharian et al., 2025).

The rest of this paper is organized as follows. Section 2 reviews related work on breast cancer classification and CAD systems. Section 3 describes the WDBC dataset and preprocessing. Section 4 presents the methodology and implementation details. Section 5 reports experimental results and statistical analysis. Section 6 discusses the findings, limitations, and future work, followed by the conclusion in Section 7.

2. METHODOLOGY

A. Evaluated Algorithms

Five supervised learning algorithms are evaluated in this study:

1. Support Vector Machine (SVM) with radial basis function (RBF) kernel.
2. Decision Tree (DT) using the Gini impurity criterion.
3. k-Nearest Neighbors (KNN) using the k-nearest neighbors rule.
4. Random Forest (RF) as an ensemble of decision trees.
5. Logistic Regression (LR) with regularized linear decision boundary.

These models represent a diverse set of hypothesis classes and are widely adopted in breast cancer classification literature (Aamir et al., 2022; Asri et al., n.d.; Eftekharian et al., 2025).

B. Hyperparameter Optimization

For each algorithm, hyperparameters are optimized using RandomizedSearchCV with stratified 5-fold cross-validation on the training set and accuracy as the primary scoring metric (Aamir et al., 2022). The search spaces are defined as follows:

- **SVM (RBF)**: $C \in [10^{-2}, 10^{2}]$, gamma ∈ {‘scale’, ‘auto’, 0.001, 0.01, 0.1, 1}.
- **DT**: max_depth ∈ {3, 5, 7, 9, 11, 13, 15}, min_samples_split ∈ {2, 4, 6, 8, 10}.

- **KNN**: n_neighbors ∈ {3, 5, 7, 9, 11, 13, 15}, weights ∈ {‘uniform’, ‘distance’}, metric ∈ {‘euclidean’, ‘manhattan’}.
- **RF**: n_estimators ∈ {50, 100, 150, 200}, max_depth ∈ {5, 8, 10, 12, 15, 20}, max_features ∈ {‘sqrt’, ‘log2’}.
- **LR**: C ∈ {0.01, 0.1, 1, 10, 100}, penalty ∈ {‘l1’, ‘l2’}, with suitable solver choices.

RandomizedSearchCV is configured with a fixed number of iterations per model, and the best estimator (highest mean cross-validation accuracy) is selected and retrained on the full training set.

C. Evaluation Metrics and Statistical Analysis

Model performance is evaluated on the held-out test set and using cross-validation on the training set. The following metrics are computed:

- Accuracy
- Precision
- Recall (sensitivity)
- Specificity
- F1-score
- Area under ROC curve (ROC-AUC)
- Confusion matrix (TP, TN, FP, FN).

To assess model stability, 5-fold cross-validation accuracies are summarized by mean, standard deviation, range, and approximate 95% confidence interval using normal approximation and bootstrap resampling (Aamir et al., 2022). Additionally, McNemar’s test is applied on paired test set predictions to assess whether SVM significantly outperforms other classifiers (Eftekharian et al., 2025).

D. Implementation Details

All experiments follow a reproducible stack with Python, standard scientific libraries for data handling and visualization, and scikit-learn as the core machine learning framework (Aamir et al., 2022; Eftekharian et al., 2025). Train–test splitting, preprocessing, RandomizedSearchCV, model training, and metric computation are implemented using scikit-learn pipelines to avoid data leakage. The implementation is designed to run on a standard workstation CPU without requiring GPU acceleration, reflecting realistic usage in many academic and clinical environments.

E. Experimental Results

1. Optimal Hyperparameters

After RandomizedSearchCV, the best hyperparameters for each algorithm are summarized in Table 1.

Table 1. Selected Hyperparameters for Each Algorithm

Model	**Selected Hyperparameters (Key)**
SVM	C = 10.0, gamma = ‘scale’, kernel = ‘rbf’
DT	max_depth = 8, min_samples_split = 5
KNN	n_neighbors = 5, weights = ‘distance’, metric = ‘euclidean’
RF	n_estimators = 150, max_depth = 12, max_features = ‘sqrt’

LR	C = 1.0, penalty = ‘l2’

F. Test Set Performance

Table 2 reports the performance of each optimized model on the test set.

Table 2 Test Set Performance for All Models

Model	Accuracy	Precision	Recall	Specificity	F1-score	ROC-AUC
SVM	0.9825	1.0000	0.9524	1.0000	0.9760	0.9960
RF	0.9737	0.9714	0.9286	0.9861	0.9492	0.9929
LR	0.9649	0.9286	0.9286	0.9861	0.9286	0.9960
DT	0.9474	0.8800	0.8810	0.9861	0.8800	0.9605
KNN	0.9386	0.8571	0.8571	0.9861	0.8571	0.9600

SVM achieves the highest accuracy and F1-score while maintaining perfect precision and specificity on the test set. RF and LR also perform competitively, with accuracies above 96% and high ROC-AUC values. In contrast, DT and KNN show lower recall and F1-score, mainly due to higher numbers of false negatives.

3. RESULT AND DISCUSSION

Methodological Insights

From a methodological perspective, this study demonstrates that classical models, when equipped with systematic hyperparameter optimization and robust evaluation, can achieve excellent performance on tabular medical datasets such as WDBC (Aamir et al., 2022; Eftekharian et al., 2025). SVM with RBF kernel provides the best trade-off between sensitivity and specificity, with near-perfect diagnostic performance and stable cross-validation behavior. RF and LR offer strong alternatives, with slightly lower recall or precision but similar ROC-AUC and narrow confidence intervals.

The experimental design—stratified train–test splitting, standardized preprocessing, comprehensive metrics, cross-validation, and McNemar’s test—provides a reproducible methodology suitable for comparative studies and for extending to other biomedical datasets (Aamir et al., 2022; Eftekharian et al., 2025).

Practical Implications

In practice, SVM is particularly attractive for CAD systems due to its high precision (zero false positives in the current test split) and strong recall, reducing both unnecessary biopsies and missed malignancies. RF and LR may be favored when interpretability and simplicity of deployment are prioritized, as RF offers feature importance, and LR provides a transparent linear model (Eftekharian et al., 2025).

The feature's important results align with medical knowledge, as extreme morphological changes in nuclei are indicative of malignant transformation, supporting the potential acceptance of these models by clinical experts (Eftekharian et al., 2025).

Limitations

Several limitations should be acknowledged. First, the dataset size is relatively small (569 samples) and originates from a single benchmark source, which may limit generalizability to diverse populations and acquisition protocols (Breast Cancer Wisconsin (Diagnostic) data documentation, n.d.; Eftekharian et al., 2025). Second, the class distribution, although not severely imbalanced, still favors benign cases; recall values may be sensitive to different prevalence in real-world settings (Aamir et al., 2022). Third, no external validation cohort is used, and hyperparameter tuning uses only internal cross-validation (Aamir et al., 2022). Finally, this study focuses on classical models and does not compare against modern gradient boosting and deep learning approaches on raw imaging data (Aamir et al., 2022; Transforming breast cancer prediction: Advanced machine learning models for risk prediction and classification, 2025).

Future Work

Future research can extend this work in several directions. Incorporating additional datasets from multiple centers would allow more reliable assessment of generalization and robustness (Eftekharian et al., 2025). Cost-sensitive learning and customized loss functions could be used to explicitly penalize false negatives, which are critical in cancer screening (Aamir et al., 2022). Integration of explainable AI techniques such as SHAP or LIME would provide local explanations for individual predictions, further improving interpretability (Eftekharian et al., 2025). Finally, ensemble strategies combining SVM, RF, and gradient boosting could be explored as hybrid CAD frameworks, along with comparisons against deep learning models trained on original imaging data (Aamir et al., 2022)

4. CONCLUSION

This paper presents a comprehensive and statistically grounded comparison of five supervised machine learning algorithms for breast cancer diagnosis on the WDBC dataset. By applying systematic hyperparameter optimization, stratified 5-fold cross-validation, and multi-metric evaluation, the study shows that an RBF-kernel SVM achieves the best overall performance, with 98.25% accuracy, 100% precision, 95.24% recall, 100% specificity, and ROC-AUC of 0.9960 on the test set. RF and LR also provide competitive and stable results, while DT and KNN show higher false negative rates and larger variance.

Feature importance analysis confirms that "worst" morphological features dominate predictive performance and are consistent with established histopathological criteria (Eftekharian et al., 2025). Overall, the findings highlight that carefully tuned traditional models remain highly effective for tabular medical data and provide a robust baseline and methodology for future research in breast cancer CAD systems (Aamir et al., 2022; Eftekharian et al., 2025).

5. REFERENCES

Aamir, S., Ali, M., & Hussain, M. Predicting breast cancer leveraging supervised machine learning algorithms. Frontiers in Public Health.

Aamir, S., et al. (2022). An optimized framework for breast cancer classification using machine learning. Computational and Mathematical Methods in Medicine, 2022, 8482022.

Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. Breast cancer prediction using KNN, SVM, Logistic Regression and Random Forest on WDBC dataset. IJRASET.

Breast Cancer Wisconsin (Diagnostic) data documentation. (n.d.). R mclust package manual, wdbc dataset description. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Eftekharian, M., et al. (2025). Accurate and interpretable breast cancer diagnosis using the Breast Cancer Wisconsin (Diagnostic) Dataset. International Journal of Systems and Machine Research, 12(2), 1–15.

Enhancing breast cancer detection and classification using machine learning-based computer-aided diagnosis systems. (2023). Journal of Imaging and Health Informatics, 13(10), 1–15.

Transforming breast cancer prediction: Advanced machine learning models for risk prediction and classification. (2025). International Journal of Scientific & Medical Research, 14(3), 10575.

Wolberg, W. H., Mangasarian, O. L., Street, W. N., & Street, D. Breast Cancer Wisconsin (Diagnostic) Data Set. UCI Machine Learning Repository.