p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

Comparative Evaluation of Machine Learning Algorithms for Breast Cancer Classification on a Curated Public Dataset

Resad Setyadi[*1], Aedah Abd Rahman[2]
[*1]Information Systems Study Program, Telkom University, Purwokerto Campus, Indonesia
[2]School of Graduate Studies, Faculty of ICT, Asia e University, Malaysia
E-mail: [*1]**resads@telkomuniversity.ac.id**, [2]aedah.abdrahman@aeu.edu.my

Abstract

Breast cancer classification using machine learning has been widely studied, particularly with the Breast Cancer Wisconsin Diagnostic dataset. Therefore, the main issue is not merely to report high accuracy, but to present a reproducible and clinically cautious comparative evaluation that prioritizes malignant-case performance, prevents data leakage, reports model configurations, and provides consistent interpretation. This study compares Logistic Regression, Support Vector Machine with radial basis function kernel, Gradient Boosting, and Random Forest using 569 instances and 30 numerical features extracted from digitized fine-needle aspiration cell nuclei. Standardization for scale-sensitive algorithms was placed inside the cross-validation pipeline. The malignant class was treated as the clinically critical positive class, and the models were evaluated using accuracy, malignant precision, malignant recall, malignant F1-score, specificity, balanced accuracy, Matthews Correlation Coefficient, and AUC. SVM RBF achieved the strongest overall test performance with accuracy of 0.9825, malignant recall of 0.9762, malignant F1-score of 0.9762, balanced accuracy of 0.9812, MCC of 0.9623, and AUC of 0.9977. A Wilcoxon signed-rank comparison across ten folds showed no significant difference between SVM RBF and Logistic Regression, while SVM RBF was significantly better than Gradient Boosting for malignant F1-score. Permutation importance applied to the selected SVM RBF model indicated that worst smoothness, worst texture, worst area, radius error, and worst radius contributed strongly to the predictions. The findings are limited to one curated public dataset and do not establish clinical validity.

Keywords—breast cancer classification, machine learning, malignant recall, permutation importance, statistical comparison

Submission: June 14, 2026 *Accepted: June 25, 2026* *Published: July 15, 2026*
DOI Article: https://doi.org/10.25134/ilkom.v20i2.621

1. INTRODUCTION

The development of data science has made decision-making increasingly supported by patterns learned from historical data. In classification problems, machine learning algorithms transform structured attributes into predictive outputs, such as benign and malignant classes in a medical dataset. Healthcare classification requires careful reporting because predictive results may influence how readers understand risk, even when the model is only used for computational evaluation.

Breast cancer classification based on diagnostic attributes is a suitable case for algorithm comparison because the Breast Cancer Wisconsin Diagnostic dataset contains numerical features extracted from digitized fine-needle aspiration cell nuclei. The dataset has been widely used in previous studies, so high classification accuracy alone is not sufficient as a scientific contribution. Earlier works have compared several algorithms on the same dataset and reported strong performance, including sensitivity and specificity values [19]. Therefore, a new implementation must clearly state its difference and avoid implying clinical validation from a single curated dataset.

Studies published in Nuansa Informatika show that algorithm implementation has become increasingly diverse, including Support Vector Machine for sentiment analysis [1], PCA and clustering for higher education quality analysis [2], K-Means for poverty clustering [3], and Random Forest Regressor for shipping tariff prediction [4]. These studies indicate that algorithmic and data mining topics are relevant to the journal scope. However, implementation-oriented studies can be strengthened when they report validation procedures, model parameters, class-specific evaluation, and limitations instead of relying only on a single accuracy value.

The research gap addressed in this study is the need for a reproducible comparative pipeline that emphasizes clinically critical malignant-class evaluation, reports important algorithm settings, checks performance stability through cross-validation, applies a statistical comparison

p-ISSN :1858-3911 , e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

across folds, and provides interpretation from the selected best-performing model. This scope differentiates the present study from basic comparisons that primarily focus on overall accuracy or on feature importance from a model that is not the selected predictor.

The contributions of this study are threefold. First, the study presents a leakage-aware comparative pipeline in which standardization is performed within the model pipeline and within each cross-validation fold. Second, the study evaluates the malignant class as the positive class and adds specificity, balanced accuracy, and Matthews Correlation Coefficient to complement accuracy, precision, recall, F1-score, and AUC. Third, the study applies permutation importance to the selected SVM RBF model, while explicitly acknowledging that interpretation is computational and dataset specific.

The research questions are as follows: how can a reproducible machine learning pipeline be applied to breast cancer classification on a curated public dataset, which algorithm provides the strongest performance when malignant-case metrics are prioritized, and which features most influence the selected model according to permutation importance. The study is limited to a public structured dataset and does not use local patient data, external validation cohorts, or clinical deployment testing. Thus, the results must be interpreted as computational evidence only, not as a diagnostic recommendation.

2. RESEARCH METHOD

2.1. Research design

This study uses a computational experimental design with a quantitative approach. The workflow adapts CRISP-DM because it provides systematic stages for problem formulation, data understanding, preprocessing, modeling, validation, and interpretation [5]. The design was revised to emphasize reproducibility and leakage prevention, especially because the dataset is relatively small and the difference between leading algorithms can be narrow.

The research stages include dataset inspection, stratified splitting, preprocessing inside the pipeline, model training, cross-validation, holdout testing, statistical comparison, and interpretability analysis. Figure 1 summarizes the revised workflow.

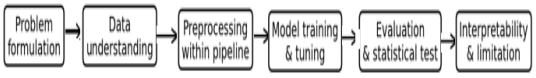


Figure 1. Revised CRISP-DM-based comparative pipeline

2.2. Dataset and research variables

The dataset used is the Breast Cancer Wisconsin Diagnostic dataset, which contains 569 instances and 30 numerical features describing cell nucleus characteristics, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension [6], [17]. The target classes consist of malignant and benign. In the dataset coding, class 0 represents malignant and class 1 represents benign.

The dataset contains 212 malignant instances and 357 benign instances. Although the imbalance is moderate, the study uses stratified splitting to preserve class proportions in training, testing, and cross-validation folds. For clinical caution, malignant is treated as the positive class in class-specific precision, recall, and F1-score calculations because false-negative malignant predictions may delay further examination.

Table I: Summary of the research dataset

Component	Description
Data source	Breast Cancer Wisconsin Diagnostic dataset
Instances	569
Features	30 numerical features
Target classes	Malignant (0) and benign (1)
Class distribution	212 malignant; 357 benign
Evaluation scheme	80:20 stratified split and 10-fold CV
Positive class	Malignant for class-specific metrics

2.3. Preprocessing and data leakage control

The preprocessing stage begins with checking data structure, feature type, missing values, and class distribution. The dataset contains numerical features and no missing values; therefore, imputation and categorical encoding are not required. Standardization is required for Logistic Regression and SVM RBF because these algorithms are sensitive to feature scale.

To prevent data leakage, standardization is placed inside the scikit-learn pipeline. In the holdout test, the scaler is fitted only on the training data and then applied to the testing data. In cross-validation, the scaler is refitted within each training fold. This procedure prevents information from the validation or testing set from influencing the training transformation [15], [16]. Tree-based models are evaluated without standardization because their split mechanisms are relatively insensitive to scale.

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

2.4. Implemented algorithms and configurations

Four supervised algorithms were compared to represent different model families: Logistic Regression as a linear probabilistic baseline, SVM RBF as a margin-based non-linear classifier, Gradient Boosting as a sequential ensemble, and Random Forest as a bagging ensemble [7], [8]. The main configurations are reported to improve reproducibility. A limited grid search using five-fold validation on the training data was used to select key parameters for the final comparison. The search space was intentionally moderate to avoid over-optimistic tuning on a small dataset [9].

Table II: Final algorithm configurations

Model	Preprocess	Key parameters
Logistic Reg.	StandardScaler	C=0.1; solver=lbfgs; max_iter=1000
SVM RBF	StandardScaler	C=10; gamma=0.01; kernel=rbf
Grad. Boosting	None	n_estimators=100; lr=0.1; max_depth=2
Random Forest	None	n_estimators=300; class_weight=balanced

2.5. Evaluation of metrics and statistical comparison

The evaluation uses accuracy, malignant precision, malignant recall, malignant F1-score, specificity, balanced accuracy, Matthews Correlation Coefficient, and AUC. Specificity is reported as benign-class recall. Balanced accuracy is used to summarize sensitivity and specificity, while MCC provides a balanced correlation-based measure for binary classification [14].

Because the leading models show small numerical differences, statistical comparison was added using the Wilcoxon signed-rank test on malignant F1-score across ten stratified folds. This test is used to examine whether paired fold-level differences are likely to be meaningful rather than only sampling variation [20]. The result is interpreted cautiously because ten folds from one dataset do not replace external validation.

Table III: Main evaluation focus

Metric	Purpose
Malignant recall	Sensitivity to malignant cases; reduces false-negative risk
Specificity	Benign recall; reduces false-positive burden
Balanced acc.	Average of malignant recall and specificity
MCC	Balanced binary classification correlation
AUC	Discrimination across thresholds
Wilcoxon test	Paired fold-level comparison of malignant F1

2.6. Interpretability method

The reviewer noted that Random Forest feature importance does not directly explain the predictions of Logistic Regression or SVM RBF. Therefore, the revised analysis uses permutation importance on the selected SVM RBF model. Permutation importance measures the decrease in malignant F1-score after each feature is randomly permuted, so the interpretation is more consistent with the selected predictive model [11], [12]. The analysis remains global and does not provide individual patient-level explanations.

3. RESEARCH RESULTS

3.1. Dataset profile

The data exploration results show that the dataset consists of 569 instances, including 212 malignant and 357 benign instances. Figure 2 shows that the benign class is larger, but the imbalance is not extreme. This distribution justifies the use of stratified splitting and class-specific metrics rather than relying only on overall accuracy.

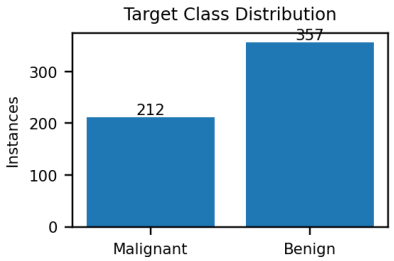


Figure 2. Target class distribution

3.2. Cross-validation results

Table IV presents the stratified ten-fold cross-validation results. SVM RBF and Logistic Regression produced the highest mean performance and relatively low standard deviations. SVM RBF achieved a malignant F1-score of 0.9687 ± 0.0269 and balanced accuracy of 0.9740 ± 0.0232, while Logistic Regression achieved a malignant F1-score of 0.9680 ± 0.0287 and balanced accuracy of 0.9712 ± 0.0261. The differences between these two leading models are small, indicating that both are stable for this dataset.

Table IV: Stratified ten-fold cross-validation results

p-ISSN :1858-3911 **,** ***e-ISSN : 2614-5405***
https://journal.fkom.uniku.ac.id/ilkom

Model	Acc.	M-F1	BalAcc	MCC	AUC-M
Logistic Reg.	0.9772 ±0.0193	0.9680 ±0.0287	0.9712 ±0.0261	0.9520 ±0.0403	0.9956 ±0.0069
SVM RBF	0.9772 ±0.0193	0.9687 ±0.0269	0.9740 ±0.0232	0.9522 ±0.0405	0.9963 ±0.0062
Grad. Boosting	0.9526 ±0.0260	0.9348 ±0.0374	0.9457 ±0.0305	0.8996 ±0.0565	0.9909 ±0.0119
Random Forest	0.9579 ±0.0296	0.9430 ±0.0409	0.9539 ±0.0327	0.9115 ±0.0627	0.9884 ±0.0182

3.3. Holdout testing results and false-negative analysis

Testing on 20% holdout data was conducted after model training on the stratified training set. Table V shows that SVM RBF achieved the strongest overall holdout performance, with accuracy of 0.9825, malignant recall of 0.9762, malignant F1-score of 0.9762, balanced accuracy of 0.9812, MCC of 0.9623, and AUC of 0.9977. Logistic Regression achieved the same accuracy and malignant F1-score, with slightly lower AUC.

The confusion matrix in Figure 3 shows that SVM RBF produced one false negative for malignant cases and one false positive for benign cases. Although the error count is small, the malignant false negative is clinically more important because it represents a malignant case predicted as benign. This result supports the reviewer's recommendation that malignant-class sensitivity must be emphasized.

Table V: Holdout testing results with malignant as positive class

Model	Acc.	M-Prec	M-Rec	M-F1	Spec.	MCC	AUC-M
Logistic Reg.	0.9737	0.9756	0.9524	0.9639	0.9861	0.9433	0.9957
SVM RBF	0.9825	0.9762	0.9762	0.9762	0.9861	0.9623	0.9977
Grad. Boosting	0.9474	0.9500	0.9048	0.9268	0.9722	0.8864	0.9907
Random Forest	0.9474	0.9286	0.9286	0.9286	0.9583	0.8869	0.9937

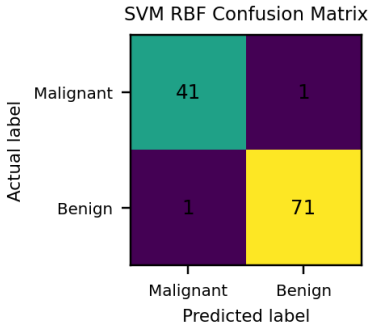


Figure 3. Confusion matrix of SVM RBF on holdout data

3.4. Statistical comparison

The Wilcoxon signed-rank test in Table VI shows that SVM RBF and Logistic Regression do not differ significantly in malignant F1-score across ten folds ($p = 0.7500$). SVM RBF is significantly higher than Gradient Boosting at the 0.05 level ($p = 0.0254$), while the difference between SVM RBF and Random Forest is not significant at the 0.05 level ($p = 0.0938$). These findings suggest that the superiority of SVM RBF over Logistic Regression should not be overstated, even though SVM RBF has the best overall numerical profile.

Table VI: Wilcoxon signed-rank test on malignant F1-score

Comparison	p-value	Interpretation
SVM vs Logistic Reg.	0.7500	No significant difference
SVM vs Grad. Boosting	0.0254	Significant at 0.05
SVM vs Random Forest	0.0938	No significant difference

3.5. Permutation importance of the selected model

Permutation importance was calculated for the selected SVM RBF model using malignant F1-score as the scoring function. Figure 4 and Table VII show that worst smoothness, worst texture, worst area, radius error, and worst radius are among the most influential features. These results are computationally plausible because the WDBC features are derived from nuclear size, shape, and texture measurements [17]. However, the feature ranking should not be interpreted as a clinical biomarker discovery result without medical validation.

The presence of nuclear morphology-related variables among important features is consistent with the broader pathological understanding that tumor morphology and nuclear characteristics are relevant in breast cancer assessment [18]. Nevertheless, this study only evaluates digitized

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

feature patterns in a public dataset, not direct histopathological diagnosis.

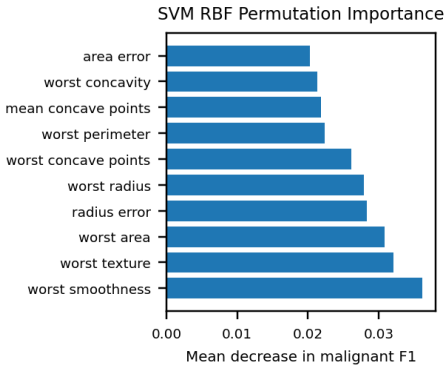


Figure 4. Top permutation importance values for SVM RBF

Table VII: Top ten SVM RBF permutation importance values

Rank	Feature	Mean decrease
1	worst smoothness	0.0362
2	worst texture	0.0321
3	worst area	0.0308
4	radius error	0.0283
5	worst radius	0.0279
6	worst concave points	0.0261
7	worst perimeter	0.0224
8	mean concave points	0.0218
9	worst concavity	0.0213
10	area error	0.0203

4. DISCUSSION

The revised results show that Logistic Regression and SVM RBF are strong models for this curated numerical dataset. Their high performance indicates that the WDBC feature space has clear class-separation patterns after appropriate standardization. This finding is consistent with the idea that more complex ensemble models do not always outperform simpler or margin-based models when the input features are already informative.

Compared with earlier machine learning studies using the WDBC dataset, the contribution of this study is not the introduction of a new algorithm. Instead, the contribution is the corrected evaluation structure: leakage-aware preprocessing, explicit parameter reporting, malignant-class prioritization, additional clinically relevant metrics, statistical comparison, and model-consistent permutation importance. This framing avoids overclaiming novelty while still improving methodological transparency.

The false-negative issue is central in medical classification. In this study, SVM RBF produced one malignant false negative on the holdout data. Although this is numerically small, the consequence of a malignant case being predicted as benign is more serious than a benign case being flagged as malignant. Therefore, future work should consider threshold adjustment, cost-sensitive learning, calibration analysis, and external validation if the model is intended to support decision-making.

The interpretability analysis is also revised. The original Random Forest feature importance can explain the ensemble model but not the selected SVM RBF predictor. Permutation importance on SVM RBF is more consistent because it evaluates the effect of feature perturbation on the selected model's malignant F1-score. However, the method remains global and cannot explain individual cases. SHAP or LIME may be needed in future studies for instance-level explanation [11], [12].

Several limitations remain. First, the dataset is a single curated public dataset, so the findings do not establish clinical validity or generalizability to hospitals with different measurement protocols. Second, the hyperparameter search is limited and may not capture all possible optimal configurations. Third, the statistical test is based on folds from one dataset and should be treated as supporting evidence rather than definitive proof. Fourth, the study does not use external validation, prospective data, or physician assessment.

Table VIII: Reviewer-related revision focus

Reviewer concern	Revision made
Novelty and gap	Gap reframed as reproducible, malignant-focused comparison
Model settings	Solver, C, gamma, estimators, depth, and class weight reported
Positive class	Malignant treated as positive class in class-specific metrics
Statistical test	Wilcoxon test added across ten folds
Interpretability	Permutation importance applied to selected SVM RBF model
Clinical validity	Single public dataset limitation stated in abstract, discussion, conclusion

p-ISSN :1858-3911 **,** ***e-ISSN : 2614-5405***
https://journal.fkom.uniku.ac.id/ilkom

5. CONCLUSION

This study revised the comparative evaluation of Logistic Regression, SVM RBF, Gradient Boosting, and Random Forest for breast cancer classification on the Breast Cancer Wisconsin Diagnostic dataset. The revised pipeline prevents standardization leakage, reports important model configurations, evaluates malignant cases as the clinically critical positive class, and adds specificity, balanced accuracy, MCC, AUC, statistical comparison, and permutation importance. SVM RBF achieved the strongest overall holdout profile with accuracy of 0.9825, malignant recall of 0.9762, malignant F1-score of 0.9762, balanced accuracy of 0.9812, MCC of 0.9623, and AUC of 0.9977. However, the Wilcoxon test showed no significant malignant F1-score difference between SVM RBF and Logistic Regression, so both models should be regarded as leading candidates on this dataset. Permutation importance for SVM RBF identified morphology-related features such as worst smoothness, worst texture, worst area, radius error, and worst radius as influential variables. These findings remain computational and dataset-specific; they do not establish clinical validity and must be externally validated before any practical medical use.

REFERENCES

[1] E. Setiana, Marwondo, V. R. Danestiara, and Wiyanudin, "Sentiment Analysis of Online Lecture Implementation Using the Support Vector Machine Algorithm," Nuansa Informatika, vol. 17, no. 2, pp. 66–70, 2023, doi: 10.25134/ilkom.v17i2.11.

[2] R. Rianti, R. Andarsyah, and R. M. Awangga, "Application of PCA and Clustering Algorithms for Higher Education Quality Analysis in LLDIKTI Region IV," Nuansa Informatika, vol. 18, no. 2, pp. 67–77, 2024, doi: 10.25134/ilkom.v18i2.211.

[3] A. Mukhyidin, A. Faqih, and A. R. Rinald, "Using K-Means for District-City Poverty Clustering in Indonesia," Nuansa Informatika, vol. 19, no. 1, pp. 75–81, 2025, doi: 10.25134/ilkom.v19i1.300.

[4] N. H. Harani, M. Y. H. Setyawan, and D. Ferdinan, "Predicting Basic Shipping Tariff Using Machine Learning," Nuansa Informatika, vol. 19, no. 2, pp. 58–66, 2025, doi: 10.25134/ilkom.v19i2.388.

[5] F. Martinez-Plumed et al., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

[6] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019.

[7] G. Biau and E. Scornet, "A Random Forest Guided Tour," TEST, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[8] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and Tuning Strategies for Random Forest," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3, 2019, doi: 10.1002/widm.1301.

[9] L. Yang and A. Shami, "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice," Neurocomputing, vol. 415, pp. 295–316, 2020, doi: 10.1016/j.neucom.2020.07.061.

[10] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature Selection: A Data Perspective," ACM Computing Surveys, vol. 50, no. 6, pp. 1–45, 2017, doi: 10.1145/3136625.

[11] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed. 2022.

[12] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[13] Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018, doi: 10.1145/3233231.

[14] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient over F1 Score and Accuracy in Binary Classification Evaluation," BMC Genomics, vol. 21, no. 1, 2020, doi: 10.1186/s12864-019-6413-7.

[15] G. Varoquaux and V. Cheplygina, "Machine Learning for Medical Imaging: Methodological Failures and Recommendations for the Future," npj Digital Medicine, vol. 5, no. 1, 2022, doi: 10.1038/s41746-022-00592-y.

[16] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-Learning-Based Science," Patterns, vol. 4, no. 9, 2023, doi: 10.1016/j.patter.2023.100804.

[17] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear Feature Extraction for Breast Tumor Diagnosis," in Proceedings of SPIE, vol. 1905, pp. 861–870, 1993, doi: 10.1117/12.148698.

[18] C. W. Elston and I. O. Ellis, "Pathological Prognostic Factors in Breast Cancer. I. The

p-ISSN :1858-3911* , *e-ISSN : 2614-5405
https://journal.fkom.uniku.ac.id/ilkom

Value of Histological Grade in Breast Cancer: Experience from a Large Study with Long-Term Follow-Up," Histopathology, vol. 19, no. 5, pp. 403–410, 1991, doi: 10.1111/j.1365-2559.1991.tb00229.x.

[19] A. F. Agarap, "On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset," arXiv:1711.07831, 2017.

[20] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," Journal of Machine Learning Research, vol. 7, pp. 1–30, 2006.