

Optimization of Machine Learning Algorithms in Breast Cancer Classification: A Performance Based Analysis

Arie Setya Putra^{*1}, Agus Wantoro², Ochi Marshella Febriani³

¹Department of Information Technology, Universitas Mitra Indonesia, Indonesia

²Department of Computer Science, Universitas Aisyah Pringsewu, Indonesia

³Department of Information System, IIB Darmajaya, Indonesia

Article Info

Article history:

Received Desember 26, 2025

Revised March 26, 2026

Accepted April 27, 2026

Keywords:

Machine Learning;
Imbalance Class;
Feature Selection;
Breast Cancer;
Classification;

ABSTRACT

Timely identification of breast cancer recurrence is closely associated with patient survival and the effectiveness of treatment. Inaccurate detection can contribute to greater disease severity, higher treatment costs, longer recovery, and reduced quality of care. For Machine Learning (ML)-based decision-support systems, two important challenges are the unequal distribution of medical-data classes and the large number of features, both of which may affect model accuracy and computational efficiency. This study evaluates an approach that combines feature selection with class-imbalance handling to improve breast cancer detection performance. Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and Relief-F are used to rank features according to their weights, while the Synthetic Minority Over-Sampling Technique (SMOTE) is applied to improve representation of the underrepresented class. Seven ML classifiers, namely k-Nearest Neighbor (k-NN), Tree, Support Vector Machine (SVM), Naive Bayes, AdaBoost, Random Forest (RF), and Neural Network (NN), are tested and assessed using confusion-matrix-based accuracy, precision, recall, and computational time. The experimental results indicate that incorporating class-imbalance handling improves the predictive performance of the ML algorithms. Among the evaluated combinations, Information Gain with Random Forest (IG+RF) provides the optimal result in this case. These findings highlight the value of integrating class-balancing and feature-selection procedures when developing machine-learning systems for breast cancer detection.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Arie Setya Putra

Department of Information Technology, Universitas Mitra Indonesia, Bandar Lampung, 35145, Indonesia

Email: ariesetyaputra@umitra.ac.id ; Orchid ID: 0000-0001-6929-6717

1 INTRODUCTION

Breast cancer continues to represent a major global health burden and ranks among the most prominent cancers affecting women and contributing to cancer-related mortality worldwide [1]. It is also recognized as one of the leading causes of cancer death in the female population [2]. Recent evidence indicates that its incidence is still rising, placing breast cancer among the most frequently diagnosed malignancies [3]. In cases of recurrence, timely identification is important for improving both survival and treatment outcomes. Patients whose disease is identified at an earlier stage generally have a substantially greater probability of recovery than those diagnosed at a more advanced stage [4].

Rapid progress in Machine Learning (ML) during recent years has supported the development of breast cancer detection systems that are increasingly accurate and computationally efficient [5]. ML techniques can process diverse medical information, including genetic profiles, medical images, and clinical histories, and can uncover complex relationships that may be difficult to recognize through visual examination alone. Previous studies have demonstrated the potential of these approaches to improve detection accuracy and estimate disease risk [6].

Despite these advantages, two important issues remain when ML is applied to the early detection of breast cancer: selecting informative features and dealing with imbalanced classes. The Breast Cancer Dataset (BCD) contains structured medical information that is commonly employed for research, statistical evaluation, and the training of machine-learning models for breast cancer diagnosis and detection. It is also frequently used as a benchmark for assessing breast cancer detection models [7]. The dataset contains nine features that must be examined to determine their contribution to model accuracy. Features that provide little or no relevant information may reduce predictive performance, increase computational demands, and encourage overfitting [8]. Consequently, an appropriate feature-selection procedure is required to retain the most informative variables while limiting unnecessary noise [9].

Another characteristic of the BCD is its unequal class distribution, in which the number of "No-Recurrence-Events" instances exceeds the number of "Recurrence-Events" instances. Such an imbalance can cause an ML model to favour the majority category and, as a result, perform less effectively when identifying the minority category that is of primary interest [10]. The adverse effect of class imbalance on ML performance has been widely reported, together with several strategies for mitigating it, including oversampling methods [8].

One option for addressing this issue is the Synthetic Minority Over-sampling Technique (SMOTE), which increases the representation of the minority class so that the learning algorithm receives a more balanced set of training examples [11]. Previous evidence also indicates that combining effective feature selection with class-imbalance treatment can improve the performance of ML models [12]. Even so, a broader evaluation of different feature-selection strategies together with imbalance-handling procedures is still required to determine which combination is most appropriate for this problem.

Accordingly, this study performs a performance-oriented comparison of feature-selection approaches and class-imbalance handling for breast cancer recurrence detection. The evaluated ML algorithms include Neural Network (NN), Tree, Support Vector Machine (SVM), Naive Bayes, AdaBoost, Neural Network (NN), and Random Forest (RF). By examining different combinations of these methods, the study seeks to identify an approach that can provide accurate, sensitive, and specific breast cancer detection and thereby support improved patient survival [5].

2 METHOD

The research procedure begins with collecting the data and then preparing it before model training. During preprocessing, the dataset is cleaned and arranged for subsequent analysis. Class distribution is handled with SMOTE, while feature selection is conducted using Information Gain (IG), Gain Ratio (GR), Gini Decrease (GD), and Relief-F. The resulting selected features are subsequently supplied to the breast cancer classifiers, namely K-Nearest Neighbours (k-NN), Tree, SVM, Naive Bayes, AdaBoost, Neural Network (NN), and Random Forest (RF). Figure 1 presents the overall structure of the proposed procedure.

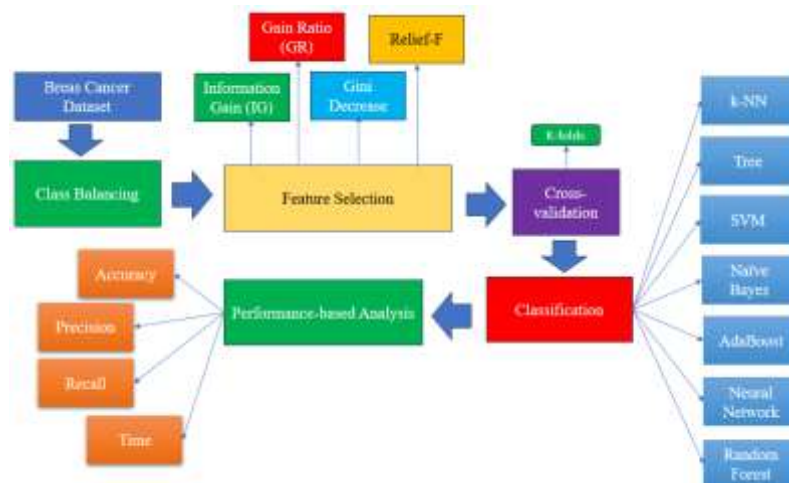


Figure 1. Workflow of the proposed method

Once breast cancer classification has been completed, the models are compared using accuracy, precision, recall, and the computational time required to construct each ML model.

2.1 Dataset

The experiments employ a breast cancer dataset containing 286 records and 10 features. The class distribution of these records is summarized in Table 1.

Table 1. Distribution of class labels

Recurrences-Events	No-Recurrence-Events
201	85

As shown in Table 1, the numbers of “Recurrence-Events” and “No-Recurrence-Events” are unequal. In this dataset, “Recurrence-Events” forms the majority class, whereas “No-Recurrence-Events” represents the minority class. This imbalance can prevent an ML classifier from reaching optimal accuracy because its predictions may be dominated by the majority category [23]. A balancing procedure is therefore required so that the algorithms can provide more optimal performance [24]. Table 2 lists the features contained in the dataset.

Table 2. Dataset features and attributes

ID	Feature	Information
f1	Age	Patient age range
f2	Menopause	Natural biological transition marking the end of a woman's menstrual cycles
f3	Tumor-size	Dimensions or measurements of the tumor considered during cancer diagnosis and treatment
f4	Inv-nodes	Number of axillary lymph nodes (range 0-39) that show metastatic breast-cancer involvement
f5	Node-caps	Indicates spread of the cancer to a lymph node beyond the original tumor site, where it may remain
f6	Deg-maling	Cancer is a major cause of mortality, and breast cancer is the most common cancer among women
f7	Breast	Location or side of the breast
f8	Breast-quad	Breast location divided into four quadrants with the nipple used as the central reference
f9	Irradiat	Irradiation, or radiation therapy, is a commonly used treatment for breast cancer
f10	Class	Target attribute indicating whether cancer recurrence occurs or does not occur

2.2 Over-sampling Technique

The class-balancing method adopted in this study is the Synthetic Minority Over-sampling Technique (SMOTE), which is commonly used to address imbalanced class distributions [25]. Rather than merely duplicating existing minority observations, SMOTE extends conventional oversampling by creating additional minority-class samples, thereby moving the class proportions toward a more balanced distribution [26]. Combining an effective feature-selection strategy with treatment of class imbalance has previously been reported to enhance ML model performance [27]

2.3 Feature Selection

Selecting an appropriate subset of variables is an important part of classification because the chosen features can directly influence predictive accuracy [27]. ML datasets often include attributes that are redundant or irrelevant and therefore do not contribute to improved accuracy [28]. Such variables may have little effect on learning or can even reduce model performance [29]. For this reason, the relevance of the available features needs to be assessed before classification

2.3.1 Information Gain (IG)

Information Gain measures the reduction in class entropy that occurs after the value of an attribute is considered, and in this study, it is used to represent the relevance of individual features [30]. The method originates from decision-tree induction, where information gain serves as a criterion for selecting attributes. Compared with several alternative feature-selection procedures, it can be computed relatively quickly. Features that contribute more information receive higher rankings, whereas less informative features are assigned lower positions [31].

2.3.2 Gain-Ratio (GR)

Gain Ratio (GR) evaluates the entropy associated with the probability distribution of a data subset after a split and normalizes the corresponding information-gain value [27]. During feature selection, GR also considers the number and relative size of the resulting subsets [32]. This adjustment reduces the bias inherent in information gain. Attributes are therefore assessed according to both the branching structure and the intrinsic information produced by a split, allowing GR to correct the original information-gain measure

2.3.3 Gini Decrease (GD)

Gini Decrease, also referred to as Mean Decrease Gini or Gini Importance, is a Random Forest-based approach for feature selection. It estimates the contribution of each variable to improving node purity across the decision trees that compose the forest. Because the importance is aggregated over multiple trees rather than derived from a single tree, Mean Decrease Gini can reduce the risk of relying on an overfitted individual decision tree [33]. A low Gini value (close to 0) represents a highly homogeneous set of samples. Accordingly, a large reduction in Gini indicates that a feature is effective at separating observations into their appropriate classes [34].

2.3.4 Relief-F

Relief-F is a filter-based feature-selection method that ranks variables by assigning a weight to each feature according to its relevance [35]. Its assessment is based on how well a feature differentiates nearby observations that belong to different classes [36]. The resulting Relief-F ranking can then be combined with a classification algorithm to support improved predictive performance

2.4 Cross-validation

Classification is a widely used data-mining task in which learning methods, including decision-tree and neural-network approaches, evaluate the available observations to generate predictions [27]. After the feature-ranking stage identifies the variables associated with the class label, classification is performed. Accuracy, precision, recall, error rate, and processing time are then compared for Naive Bayes, J48, AdaBoost, Random Tree, Random Forest, and Support Vector Machine (SVM). Model validation is conducted with k-fold cross-validation, a procedure commonly applied to training data [37]. The k-fold validation process used in this study is illustrated in Figure 2

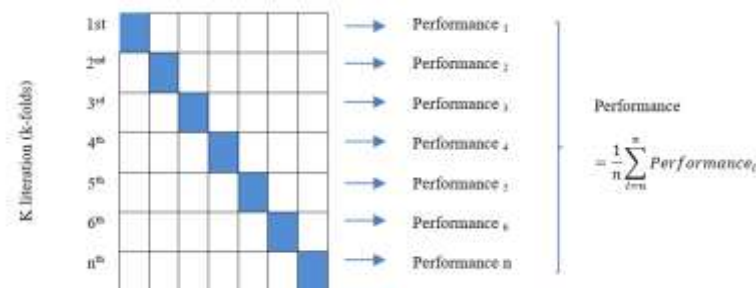


Figure 2. Procedure for k-fold cross-validation

2.5 Performance Analysis

Model performance is assessed using a confusion matrix to examine correct and incorrect classifications. For a classifier with n classes, an $n \times n$ confusion matrix places the observed class outcomes against the corresponding predicted outcomes [38]. The binary case, where $n=2$, is presented in Table 3.

Table 3. Binary confusion matrix

Class	Predicted positives	Predicted negatives
Observed positive instances	True positive instances (TP)	False negative instances (FN)
Observed negative instances	False positive instances (FP)	True negative instances (TN)

Accuracy, precision, and recall are also calculated to compare the predictive behaviour of the classifiers. These measures are obtained from the entries of the confusion matrix in Table 3 using equations (1), (2), and (3)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

3 RESULTS AND DISCUSSION (10 PT)

The outcomes of the experiments are discussed through both qualitative interpretation and quantitative measurement. To make the comparisons clear, the experimental findings are summarized with tables and graphical representations, while the graphical presentation follows the same general format as the other figures and diagrams.

3.1 Imbalance Class with SMOTE

Class balancing was performed with SMOTE oversampling using k-fold=5 because this configuration produced the best classification results in our tests. The changes in class counts after balancing are reported in Table 4.

Table 4. Class counts before and after SMOTE

Class	Non-SMOTE	SMOTE
No-recurrence-Events	201	201
Recurrence-Events	85	170
Total	286	371

Table 4 demonstrates that SMOTE changes both the class distribution and the total number of records. After oversampling, the dataset grows by 29.72%. The next analysis compares model accuracy, precision, and recall. Cross-validation again uses k-fold=5 because this setting gave the best results in our experiments. The corresponding performance values are presented in Table 5.

Table 5. Accuracy, precision, and recall for the original and SMOTE datasets

Algorithm	Dataset non-SMOTE				Dataset SMOTE			
	Accuracy	Precision	Recall	Average	Accuracy	Precision	Recall	Average
k-NN	0.741	0.722	0.741	0.7347	0.701	0.701	0.701	0.7009
Tree	0.626	0.614	0.626	0.6220	0.706	0.706	0.706	0.7060
Random Forest	0.713	0.687	0.713	0.7044	0.720	0.719	0.720	0.7195
SVM	0.703	0.654	0.703	0.6867	0.668	0.675	0.668	0.6707
Naive Bayes	0.713	0.713	0.713	0.7133	0.685	0.684	0.685	0.6843
AdaBoost	0.633	0.652	0.633	0.6391	0.709	0.713	0.709	0.7101
Neural Network	0.671	0.660	0.671	0.6677	0.687	0.692	0.687	0.6888
Average	0.686	0.672	0.686	0.6811	0.697	0.698	0.697	0.6972

The results indicate that applying SMOTE improves the overall performance of the evaluated ML algorithms, with an average increase of 2.35%. We then compare the computational time needed to construct the models, as summarized in Figure 3.

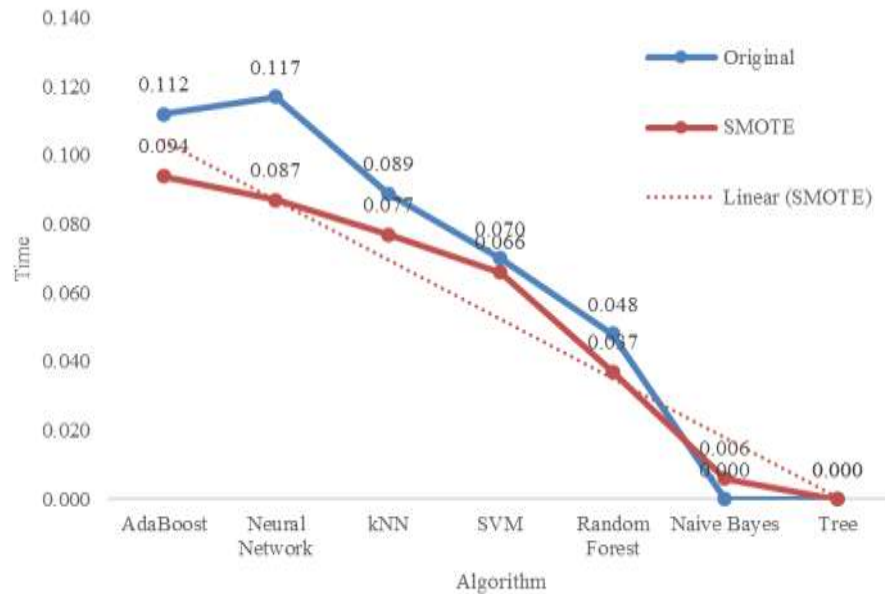


Figure 3. Comparison of computational time

Figure 3 compares the processing time of the different ML algorithms. Among the tested methods, Decision Tree consistently requires the least computation and remains fast during both training and prediction. Its relatively low complexity also makes it efficient for handling large datasets. In contrast, Neural Network and AdaBoost show the longest times on both the original and SMOTE data. The mean computation time for the original dataset is 0.062, whereas the SMOTE dataset requires 0.052. Thus, the use of SMOTE reduces the computational time by 0.010, corresponding to an improvement of 15.82%.

3.2 Comparison of Feature Selection Performance

Feature selection is applied to the SMOTE-balanced dataset in order to identify the variables that have the greatest influence on breast cancer classification. Table 6 reports the rankings obtained from the different selection techniques.

Table 6. Feature rankings produced by the selection techniques

Information Gain (IG)		Gain Ratio (GR)		Gini Decrease (GD)		Relief-F	
Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight
Tumor-size	0.096	Node-caps	0.077	Deg-malig	0.060	Breast-quad	0.105
Deg-malig	0.089	Deg-malig	0.058	Tumor-size	0.055	Node-caps	0.086
Inv-nodes	0.071	Inv-nodes	0.050	Inv-nodes	0.047	Deg-malig	0.076
Node-caps	0.062	Irradiat	0.036	Node-caps	0.042	Tumor-size	0.072
Irradiat	0.030	Tumor-size	0.033	Irradiant	0.021	Irradiat	0.046
Age	0.026	Breast	0.014	Age	0.017	Menopause	0.044
Breast-quad	0.021	Age	0.013	Breast-quad	0.014	Breast	0.042

The feature-selection results reported in Table 7 show that the different techniques generate different rankings. With IG, seven top-ranked features are obtained and “Tumor-size” receives the largest weight with respect to the class. GR assigns the greatest influence to “Node-cap”, GD ranks “Deg-malig” highest, and Relief-F identifies “Breast-quad” as its leading feature. Using these selected variables, the ML algorithms are compared again on the basis of the seven features with the highest weights. Figure 4 displays the resulting performance comparison.

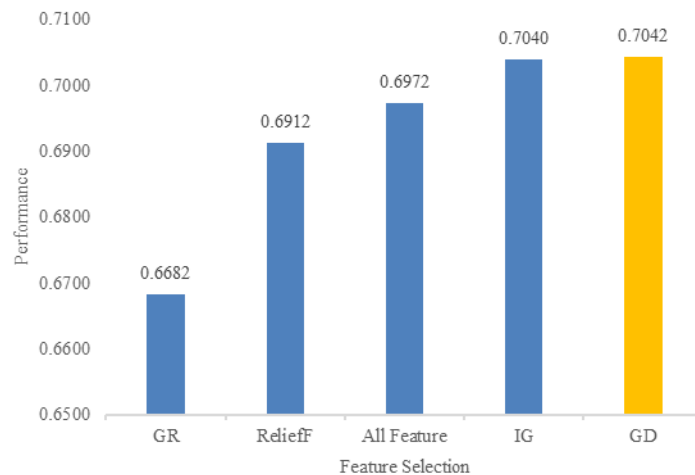


Figure 4. Performance comparison among feature-selection techniques

The evaluated feature-selection procedures produce different levels of performance, with Gini Decrease (GD) giving the strongest result. GD ranks a feature according to how much it contributes to reducing impurity throughout the trees. A larger decrease in Gini therefore indicates that the variable is more effective at separating the data into purer class groups. Within Random Forest, this mechanism is also useful for representing dependencies and interactions among complex variables.

In contrast, Gain Ratio (GR) provides the lowest performance in this experiment. One reason is that GR may favour attributes associated with very small split-information values, which can lead to the selection of variables that are not the strongest structural separators. The computation time required by each feature-selection technique is compared next in Figure 5.

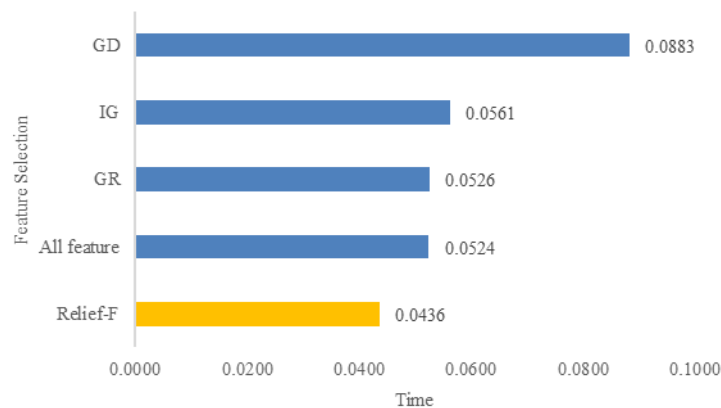


Figure 5. Computation-time comparison of feature-selection techniques

Although GD gives the best predictive performance in Figure 5, it requires the longest computation time among the evaluated feature-selection methods. This result shows a trade-off between the accuracy achieved by GD and its processing cost. Relief-F has the shortest computation time, followed by GR and IG.

3.3 Comparison of ML Algorithm Performance

A broader comparison is performed to obtain a more complete view of ML algorithm performance. The evaluation includes the confusion matrices presented in Figure 6.

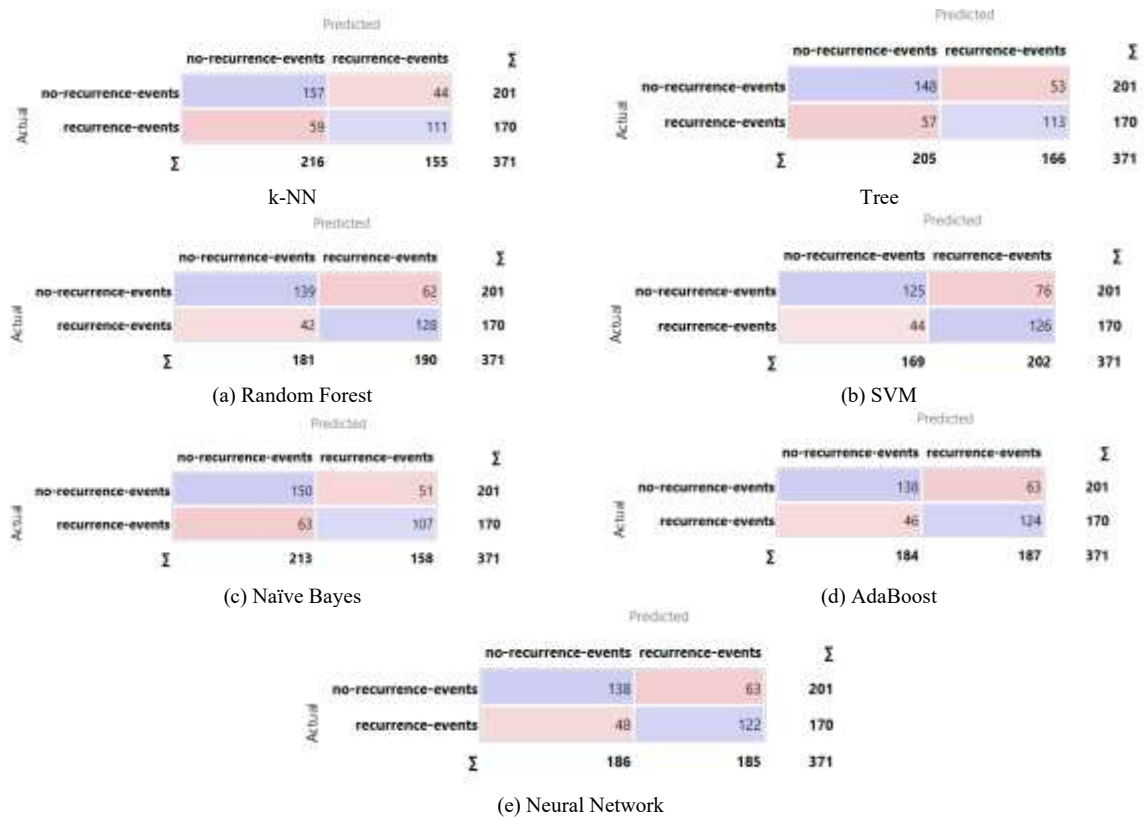


Figure 6. Confusion matrices of the machine-learning algorithms

Table 7. Average ML performance (accuracy, precision, recall) with feature-selection techniques

Algorithm	All Feature	Information Gain (IG)	Gain Ratio (GR)	Gini Decrease (GD)	Relief-F
k-NN	0.7009	0.7222	0.6816	0.7222	0.6763
Tree	0.7060	0.7060	0.6359	0.7034	0.6817
Random Forest	0.7195	0.7363	0.6793	0.7222	0.7006
SVM	0.6707	0.6799	0.6546	0.6799	0.6630
Naïve Bayes	0.6843	0.6924	0.6897	0.6924	0.6685
AdaBoost	0.7101	0.6939	0.6824	0.7075	0.7362
Neural Network	0.6888	0.6970	0.6536	0.7019	0.7119
Average	0.6972	0.7040	0.6682	0.7042	0.6912

Table 7 indicates that the IG and GD feature-selection methods can increase ML accuracy, although the magnitude of improvement is small. IG raises performance by 0.97%, whereas GD produces an increase of 1.008%. Among all combinations considered, Random Forest with Information Gain (IG) achieves the best performance. The analysis is then extended to the computational time required by each ML algorithm to construct its model. Figure 7 provides this model-building-time comparison.

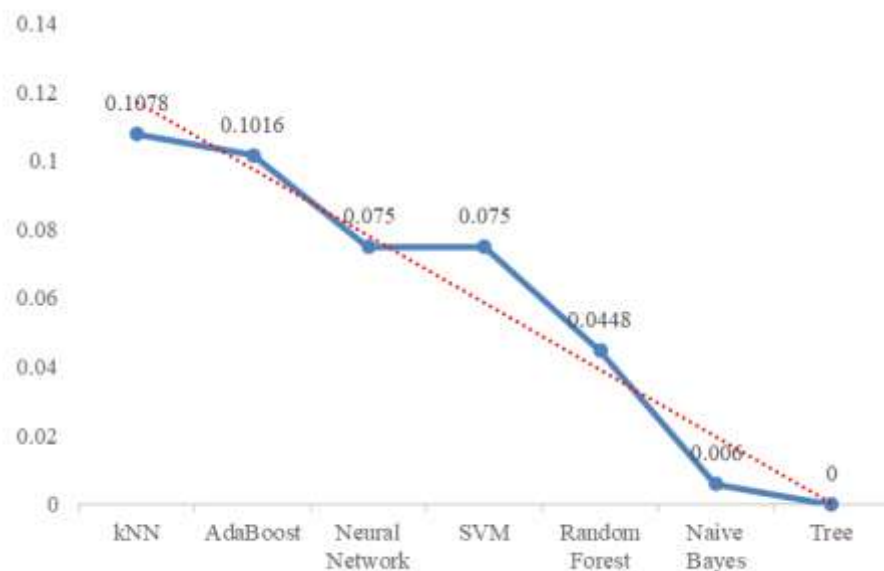


Figure 7. Comparison of model-building time

Although the Tree classifier is not the strongest method in terms of accuracy, precision, or recall, it consistently records the shortest model-building time. Naive Bayes provides the next best computational time.

4 CONCLUSION

Balancing the classes in the Breast Cancer Dataset influences classification performance. The results show that SMOTE can improve the evaluated ML models, although the changes are not large: accuracy increases by 1.568%, precision by 3.974%, and recall by 1.568%. Feature selection is also performed using IG, GR, GD, and Relief-F to identify the variables with the strongest weighted relationship to the class, after which the seven highest-ranked features are retained. Based on the accuracy, precision, and recall tests, GD provides the best feature-selection performance, whereas GR gives the weakest result. Computation time shows a different pattern: Relief-F is the fastest technique and GD is the slowest. Thus, GD offers the strongest predictive performance in this experiment but at a higher computational cost.

The ML classifiers are subsequently evaluated using the selected features to obtain a more comprehensive picture of their behaviour. Random Forest combined with Information Gain (IG) produces the highest accuracy among the tested combinations. For model-construction time, the Tree algorithm is consistently the fastest, while k-NN requires a longer execution time. The number of retained features also affects processing time, and Relief-F requires less time than using all features or the IG, GR, and GD procedures. These findings indicate that class-imbalance treatment and feature selection should be incorporated into the development pipeline of medical classification systems because they can influence both detection accuracy and overall interpretability. Future work should examine other imbalance-handling strategies, including ensemble-based sampling and cost-sensitive learning, and should evaluate these methods on larger and more complex clinical datasets

ACKNOWLEDGEMENTS

The author sincerely thanks the Faculty of Health, Aisyah Pringsewu University (UAP), for its assistance in validating the anonymized breast cancer patient data used in this study. Appreciation is also extended to the Machine Learning (ML) Research Group at UAP for its constructive input during model development and evaluation. Special appreciation is given to the physicians whose clinical perspectives contributed substantially to the interpretation of the breast cancer recurrence prediction results.

REFERENCES

- [1] A. Schindele *et al.*, "Interpretable machine learning for thyroid cancer recurrence prediction: Leveraging XGBoost and SHAP analysis," *Eur. J. Radiol.*, vol. 186, May 2025, doi: 10.1016/j.ejrad.2025.112049.
- [2] A. H. Barfejani *et al.*, "Predicting overall survival in anaplastic thyroid cancer using machine learning approaches," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 282, no. 3, pp. 1653–1657, 2025, doi:

- 10.1007/s00405-024-08986-2.
- [3] D. W. Chen, B. H. H. Lang, D. S. A. McLeod, K. Newbold, and M. R. Haymart, "Thyroid cancer," *Lancet*, vol. 401, no. 10387, pp. 1531–1544, May 2023, doi: 10.1016/S0140-6736(23)00020-X.
 - [4] A. Kuang, V. L. Kouznetsova, S. Kesari, and I. F. Tsigelny, "Diagnostics of Thyroid Cancer Using Machine Learning and Metabolomics," *Metabolites*, vol. 14, no. 1, 2024, doi: 10.3390/metabo14010011.
 - [5] R. Iacob *et al.*, "Evaluating the Role of Breast Ultrasound in Early Detection of Breast Cancer in Low- and Middle-Income Countries: A Comprehensive Narrative Review," 2024. doi: 10.3390/bioengineering11030262.
 - [6] Y.-M. Huang *et al.*, "Correction: Huang et al. Systemic Anticoagulation and Inpatient Outcomes of Pancreatic Cancer: Real-World Evidence from U.S. Nationwide Inpatient Sample. *Cancers* 2023, 15, 1985," 2024. doi: 10.3390/cancers16061181.
 - [7] I. O. Lixandru-Petre *et al.*, "Machine Learning for Thyroid Cancer Detection, Presence of Metastasis, and Recurrence Predictions—A Scoping Review," *Cancers (Basel)*, vol. 17, no. 8, pp. 1–27, 2025, doi: 10.3390/cancers17081308.
 - [8] S. Li, Z. Tang, L. Yang, M. Li, and Z. Shang, "Application of deep reinforcement learning for spike sorting under multi-class imbalance," *Comput. Biol. Med.*, vol. 164, p. 107253, 2023, doi: <https://doi.org/10.1016/j.compbiomed.2023.107253>.
 - [9] X. Song *et al.*, "Evolutionary computation for feature selection in classification: A comprehensive survey of solutions, applications and challenges," *Swarm Evol. Comput.*, vol. 90, p. 101661, 2024, doi: <https://doi.org/10.1016/j.swevo.2024.101661>.
 - [10] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artif. Intell. Rev.*, vol. 57, no. 6, p. 137, 2024, doi: 10.1007/s10462-024-10759-6.
 - [11] L. C. M. Liaw, S. C. Tan, P. Y. Goh, and C. P. Lim, "A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification," *Inf. Sci. (Ny)*, vol. 686, p. 121193, 2025, doi: <https://doi.org/10.1016/j.ins.2024.121193>.
 - [12] K. E. Setiawan, "Predicting Recurrence in Differentiated Thyroid Cancer: a Comparative Analysis of Various Machine Learning Models Including Ensemble Methods With Chi-Squared Feature Selection," *Commun. Math. Biol. Neurosci.*, vol. 2024, no. Scenario 1, pp. 1–29, 2024, doi: 10.28919/cmbn/8506.
 - [13] M. S. H. Shaon, T. Karim, M. S. Shakil, and M. Z. Hasan, "A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction," *Healthc. Anal.*, vol. 6, p. 100353, 2024, doi: <https://doi.org/10.1016/j.health.2024.100353>.
 - [14] S. Benghazouani, S. Nouh, and A. Zakrani, "Optimizing breast cancer diagnosis: Harnessing the power of nature-inspired metaheuristics for feature selection with soft voting classifiers," *Int. J. Cogn. Comput. Eng.*, vol. 6, pp. 1–20, 2025, doi: <https://doi.org/10.1016/j.ijcce.2024.09.005>.
 - [15] S. Batool and S. Zainab, "A comparative performance assessment of artificial intelligence based classifiers and optimized feature reduction technique for breast cancer diagnosis," *Comput. Biol. Med.*, vol. 183, p. 109215, 2024, doi: <https://doi.org/10.1016/j.compbiomed.2024.109215>.
 - [16] B. Zielosko and A. Dmytrenko, "Selected Methods of Feature Selection - Medical Case Study," *Procedia Comput. Sci.*, vol. 246, pp. 4451–4460, 2024, doi: <https://doi.org/10.1016/j.procs.2024.09.295>.
 - [17] K. Premalatha, D. Prabha Devi, and K. Sivakumar, "Machine learning framework for breast cancer detection with feature selection with L2 ridge regularization: insights from multiple datasets," *J. Transl. Genet. Genomics*, vol. 9, no. 1, pp. 11–34, 2025, doi: 10.20517/jtgg.2024.82.
 - [18] I. Chhillar and A. Singh, "An improved soft voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing," *Discov. Artif. Intell.*, vol. 5, no. 1, p. 4, 2025, doi: 10.1007/s44163-025-00224-w.
 - [19] D. S. Hassan, "The effect of feature selection methods on machine learning model performance: a comparative study for breast cancer prediction," *Sci. J. Univ. Zakho*, vol. 13, no. 1, pp. 102–113, 2025, doi: 10.25271/sjuoz.2025.13.1.1429.
 - [20] A. Yaqoob *et al.*, "SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study," *Sci. Rep.*, vol. 15, no. 1, p. 10944, Mar. 2025, doi: 10.1038/s41598-025-95786-1.
 - [21] G. Alfian *et al.*, "Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method," 2022. doi: 10.3390/computers11090136.
 - [22] W. M. Shaban, "Insight into breast cancer detection: new hybrid feature selection method," *Neural*

- Comput. Appl.*, vol. 35, no. 9, pp. 6831–6853, 2023, doi: 10.1007/s00521-022-08062-y.
- [23] K. Kannadasan, D. R. Edla, and V. Kuppili, “Type 2 diabetes data classification using stacked autoencoders in deep neural networks,” *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 4, pp. 530–535, 2019, doi: 10.1016/j.cegh.2018.12.004.
 - [24] C. Sharma and A. Singla, “Advanced PTSVM Based Breast Cancer Classification with Weighted Feature Selection,” *SN Comput. Sci.*, vol. 6, no. 1, p. 50, 2024, doi: 10.1007/s42979-024-03590-x.
 - [25] G. Husain *et al.*, “SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models,” *Algorithms*, vol. 18, no. 1, pp. 1–16, 2025, doi: 10.3390/a18010037.
 - [26] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, “Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest,” *Appl. Sci.*, vol. 8, no. 8, 2018, doi: 10.3390/app8081325.
 - [27] H. Sulistiani, A. Syarif, K. Muludi, and Warsito, “Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder,” *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1383–1391, 2024, doi: 10.11591/eei.v13i2.6717.
 - [28] J. Wang, S. Zhou, Y. Yi, and J. Kong, “An improved feature selection based on effective range for classification,” *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/972125.
 - [29] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, “Improving Heart Disease Prediction Using Feature Selection Approaches,” in *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 2019, pp. 619–623. doi: 10.1109/IBCAST.2019.8667106.
 - [30] J. Gao, Z. Wang, T. Jin, J. Cheng, Z. Lei, and S. Gao, “Information gain ratio-based subfeature grouping empowers particle swarm optimization for feature selection,” *Knowledge-Based Syst.*, vol. 286, p. 111380, 2024, doi: <https://doi.org/10.1016/j.knosys.2024.111380>.
 - [31] P. Bhat and K. Dutta, “A multi-tiered feature selection model for android malware detection based on Feature discrimination and Information Gain,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, Part B, pp. 9464–9477, 2022, doi: <https://doi.org/10.1016/j.jksuci.2021.11.004>.
 - [32] M. Trabelsi, N. Meddouri, and M. Maddouri, “A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis,” *Procedia Comput. Sci.*, vol. 112, pp. 186–194, 2017, doi: 10.1016/j.procs.2017.08.227.
 - [33] E. A. Algehyne, M. L. Jibril, N. A. Algehainy, O. A. Alamri, and A. K. Alzahrani, “Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia,” 2022. doi: 10.3390/bdcc6010013.
 - [34] S. Tangirala, “Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm *,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020.
 - [35] Alok Kumar Shukla, Sanjeev Kumar Pippal, Srishti Gupta, B. Ramachandra Reddy, and Diwakar Tripathi, “Knowledge discovery in medical and biological datasets by integration of Relief-F and correlation feature selection techniques,” *J. Intell. Fuzzy Syst.*, vol. 38, no. 5, pp. 6637–6648, May 2020, doi: 10.3233/JIFS-179743.
 - [36] P. Cunningham and S. J. Delany, “k-Nearest Neighbour Classifiers - A Tutorial,” *ACM Comput. Surv.*, vol. 54, no. 6, Jul. 2021, doi: 10.1145/3459665.
 - [37] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, “Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm,” *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, 2022, doi: <https://doi.org/10.1016/j.jrmge.2022.03.002>.
 - [38] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, “Confusion-matrix-based kernel logistic regression for imbalanced data classification,” *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1806–1819, 2017, doi: 10.1109/TKDE.2017.2682249