Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 2, May 2026, pp. 192-197
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v5i2.459

EVALUATION OF IMBALANCE CLASS HANDLING STRATEGIES ON MACHINE LEARNING MODEL PERFORMANCE

Arry Verdian*[1], Agus Wantoro[2]

[1]Department of Informatics Education, STKIP Rosalia, Metro, Indonesia
[2]Department of Informatics Engineering, Faculty of Technology dan Informatics, Universitas Aisyah Pringsewu, Pringsewu, Indonesia
Email: [1]verdian.2637@gmail.com, [2]aguswantoro@aisyahuniversity.ac.id

(Received: April 28, 2026; Revised: May 2, 2026; Published: May 5, 2026)

Abstract

Breast Cancer Dataset (BCD) represents a critical health problem due to the increasing prevalence of breast cancer and the importance of early detection of recurrence. Machine Learning (ML) approaches have been widely applied to support diagnosis and prediction; however, class imbalance remains a major challenge, where the majority class ("no-recurrence-events") significantly outnumbers the minority class ("recurrence-events"). This imbalance can lead to biased models that fail to accurately detect recurrence cases. This study aims to evaluate the effectiveness of class imbalance handling using the Synthetic Minority Over-sampling Technique (SMOTE) on several ML models, including Decision Tree, Naïve Bayes, k-Nearest Neighbors (k-NN), and Random Forest. The dataset used consists of 286 records with 9 features obtained from the UCI Machine Learning repository. Data preprocessing was performed, including handling missing values and outliers, followed by class balancing using SMOTE. Model evaluation was conducted using 10-fold cross-validation and performance metrics such as accuracy, precision, recall, and F1-score. The results show that the application of SMOTE significantly improves model performance, with an average accuracy increase of 11.85%. Among the evaluated models, Random Forest combined with SMOTE achieved the best performance, with an accuracy of 79.79%. In contrast, models such as Naïve Bayes and k-NN demonstrated relatively lower performance. Overall, this study confirms that handling class imbalance using SMOTE can enhance classification performance, particularly in improving the detection of minority classes in breast cancer recurrence prediction tasks.

Keywords: breast cancer; imbalance class; smote; machine learning.

1. INTRODUCTION

Breast Cancer Dataset (BCD) is a global health problem whose prevalence continues to increase [1]. Accurate early detection of cancer recurrence is very important to prevent the serious complications that this disease can cause [2]. In recent years, Machine Learning (ML) based approaches have been widely used to assist in the breast cancer diagnosis process [3]. However, a significant challenge that is often faced in implementing this classification model is class imbalance in the dataset, where the amount of data in the majority class, for example "no-recurrence-events" is much greater than in the minority class "recurrence-events"

This class imbalance can cause the classification model to be biased towards the majority class, thereby reducing the model's ability to detect cases in the minority class [4]. Several techniques have been developed to overcome this problem, such as Synthetic Minority Over-sampling Technique (SMOTE) [5]. This technique aims to balance the class distribution in the dataset before the model training process is carried out [6]. BCD is a dataset that is often used in research related to early detection of breast cancer recurrence [7]. However, this dataset also suffers from significant class imbalance problems. Several studies have shown that applying class imbalance handling techniques can improve the performance of classification models on these datasets [8]. For example, research by [5] shows that the use of Oversampling techniques can increase the accuracy of the model in detecting breast cancer case.

Research related to breast cancer classification using ML has experienced rapid development in recent years. Various models such as Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Network (CNN), and Vision Transformer (ViT) have been compared to find the model with the best performance. Research conducted by Ramadhan and Adhinata (2021) used the SMOTE technique for unbalanced data and Gini score for feature ranking. The classification models used are random forest and naïve Bayes. The results obtained by the RF classification model are superior to naïve Bayes [9]. Apart from that, research by Nurjanah dkk (2023) developed an application and compared the performance of ML models using SVM and Linear Logistics. Test results show that SVM + SMOTE shows the best accuracy. Apart from that, the accuracy obtained on the system is 90% [10]. Research by Oktaviani et al (2023) shows that the accuracy performance is different for the three models. The holdout validation

scheme with a ratio of 75%:25% produces the best accuracy for SVM, namely 98.89%. The Random Forest model achieved the best accuracy at a data split ratio of 55%:45%, namely 95.85%. However, Naïve Bayes has better accuracy performance when using k-fold cross validation with an accuracy of 93.85%. The holdout method with a ratio of 75:25 is proven to produce the best accuracy for classifying breast cancer data using SVM [11].

Various recent studies show that the application of SMOTE is generally able to improve the performance of classification models, especially in the metrics of recall, F1-score, and accuracy. Studies by Machine Learning show that the combination of SMOTE with models such as Random Forest, Support Vector Machine (SVM), and Logistic Regression provide significant improvements compared to no data balancing. However, there has not been much research comparing the performance of SMOTE on ML models such as Decision Tree, Naïve Bayes, k-NN, and Random Forest.

Although various techniques have been developed and applied, there is still a need to conduct a comprehensive evaluation of the effectiveness of each technique in the context of predicting breast cancer recurrence. This evaluation is important to determine the best strategy for dealing with class imbalance, to increase the accuracy and reliability of the classification model in detecting breast cancer recurrence.

2. RESEARCH METHODS

This section describes the procedures and stages of the research. The research stage begins with data collection. Before the data is entered into the model, it is necessary to pre-process the data. At this stage, the data is cleaned and adjusted to be processed to the next step of performing class balance using the SMOTE technique. The next stage is selecting the number of k-fold validations for classification of ML models such as Naive Bayes, Tree, k-NN, and Random Forest. The research framework design is illustrated in Figure 1.

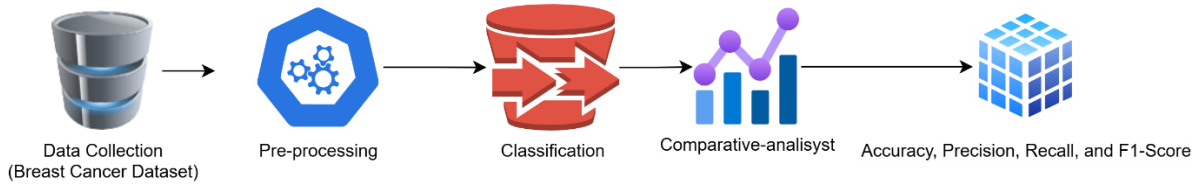


Figure 1. Research Framework Design

This research uses the BCD taken from the UCI Machine Learning dataset. The dataset has 286 records, 9 features, and 2 classes. The amount of data in each class is shown in Table 1.

Table 1. Number of Class

No-recurrence-events	Recurrence-events
188	76

Table 1 displays the number of "no-recurrence-events" and "recurrence-events" classes where the number of classes is not balanced. The number of "no-recurrence-events" classes is the majority, and the "recurrence-events" classes are the minority. This results in the classification accuracy of the ML model not being optimal because it tends to be in the majority class [12]. Some of the features used in the BCD dataset are presented in Table 2.

Table 2. Fitur, Role, dan Type Dataset

Feature	Role	Type
Class	Target	Binary
Age	Feature	Categorical
Menopause	Feature	Categorical
Tumor-size	Feature	Categorical
Inv-nodes	Feature	Categorical
Node-caps	Feature	Binary
Deg-malig	Feature	Integer
Breast	Feature	Binary
Breast-quad	Feature	Categorical
Irradiat	Feature	Binary

Next, we pre-processed the data by handling missing values and outlier data. This is done to fill in missing data and make corrections to inappropriate data

2.1. Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a development of the oversampling method, where the way this method works is by generating new samples from the minority class to make the class proportions more balanced by re-sampling the minority class samples [13]. This technique aims to balance the class distribution in the dataset before the model training process is carried out [6]

2.2. Performance Measurement

This is used to measure the performance of the ML model using the calculation of the number of correct predictions divided by the total amount of data [8]. Calculation of ML model performance accuracy (acc), precision (pre), recall (rec), and F1-Score (f1) using the equation (1-4)

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

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

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

$$\text{F1} - \text{Score} = 2\text{x}\,\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

2.3. Cross-validation (k-fold)

Classification is a form of data mining technique that is currently popular [14]. This strategy uses various methods to assess available data to produce breast cancer predictions [15]. The classification model will be validated using k-fold cross-validation. The cross-validation method is generally used for training sets [16]. Figure 2 displays the cross-validation procedure

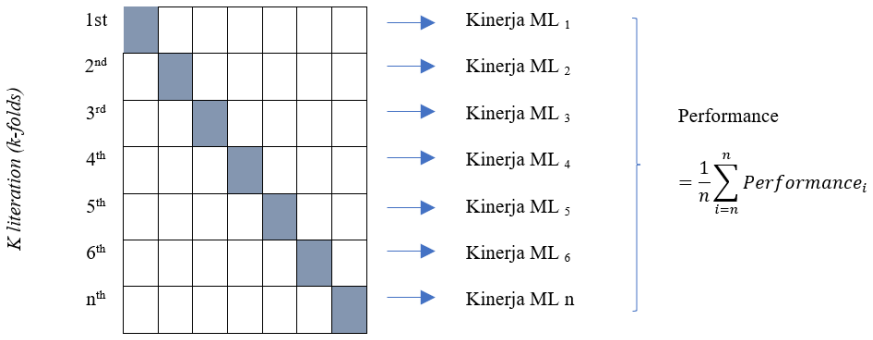


Figure 2. *Procedure k-fold validation*

3. RESULTS AND DISCUSSION

We use RapidMiner software (2026.1.1). This platform simplifies the construction of several data analysis techniques [17]. RapidMiner has the ability to categorize, perform regression, classification, remove features, create association rules, and adjust classes to datasets [18]. Figure 3 shows design implementation of the SMOTE technique.

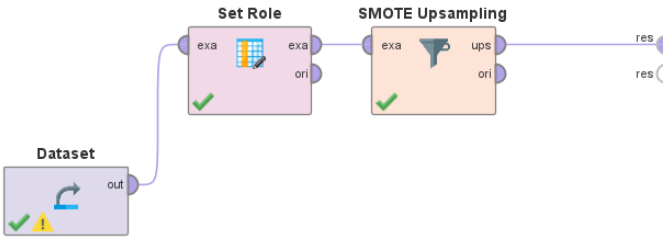


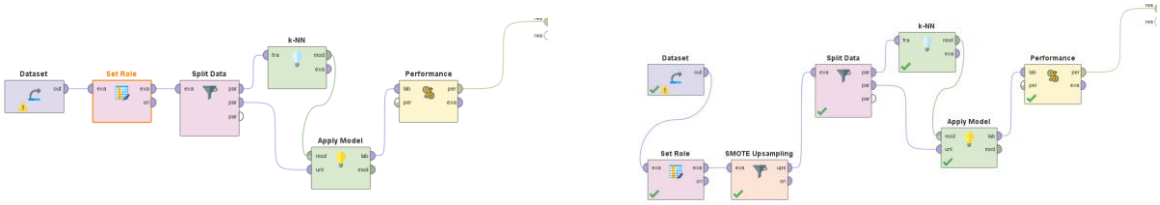
Figure 3. Design implementation SMOTE

We carried out class balancing using an oversampling approach using the SMOTE method using k-fold=10 because we found this option was the best for ML model classification. The results of the differences in the number of classes after class balancing are shown in Table 3.

Table 3. Number of class original, and SMOTE

Class	Original	SMOTE
Recurrence-events	76	188
No-recurrence-events	188	188
Total	**264**	**376**

Table 3 shows that applying the SMOTE method increases the number of classes and records following the number of majority classes. The number of minority classes increased by 40%. Next, we carry out accuracy testing using the SMOTE dataset shown in Figure. We apply a data split of 75% for training data and 25% for test data. Our findings show that splitting using these values is the most optimal for model performance. Figure 4 show design implementation SMOTE, and model performance testing.



(a) Non-SMOTE (b) SMOTE

Figure 4. SMOTE implementation design and model performance testing

After implementing the SMOTE technique, each model has different performance. There are models that experience increased performance, but there are those that experience decreased performance. The performance results of each ML model are shown in Table 4**.**

Table 4. Model performance comparison

Model	Non-SMOTE					SMOTE					
	Acc	Pre	Rec	F1	Avg	Acc	Pre	Rec	F1	Avg	Gap
Decision Tree	65.15	55.75	55.13	55.44	57.87	75.53	75.82	75.53	75.67	75.63	17.77
Naïve Bayes	68.18	58.93	67.28	62.83	64.30	74.47	76.85	74.47	75.64	75.35	11.05
k-NN	77.27	76.29	63.66	69.41	71.66	74.47	75.63	74.47	75.05	74.90	3.25
Random Forest	71.21	63.46	60.97	62.19	64.46	79.79	79.8	79.79	79.79	**79.79**	15.34
Average	**70.45**	**63.61**	**61.76**	**62.47**	**64.57**	**76.07**	**77.03**	**76.07**	**76.54**	**76.42**	**11.85**

The accuracy of the ML model is calculated using equations (1-4). Our findings show that the combination of Random Forest and SMOTE models has the best accuracy. In general, the application of the class balance method can increase accuracy. Implementing SMOTE improves accuracy by 11.85%.

The breast cancer recurrence dataset has an unbalanced class distribution (No-recurrence-events=188, recurrence-events=76). Models tend to be in the majority class. Accuracy appears to be high but recall for minority classes is low. The SMOTE technique creates synthetic data for the minority class, the distribution becomes more balanced, the model learns the patterns of both classes more fairly. Random Forest builds many decision trees from a subset of data. In the minority class the tree is not visible. With the SMOTE technique, each tree has a greater chance of learning minority patterns. As a result, the performance of the Random Forest model increases. Next, we ranked (Figure 5) the models that had the lowest to highest performance.

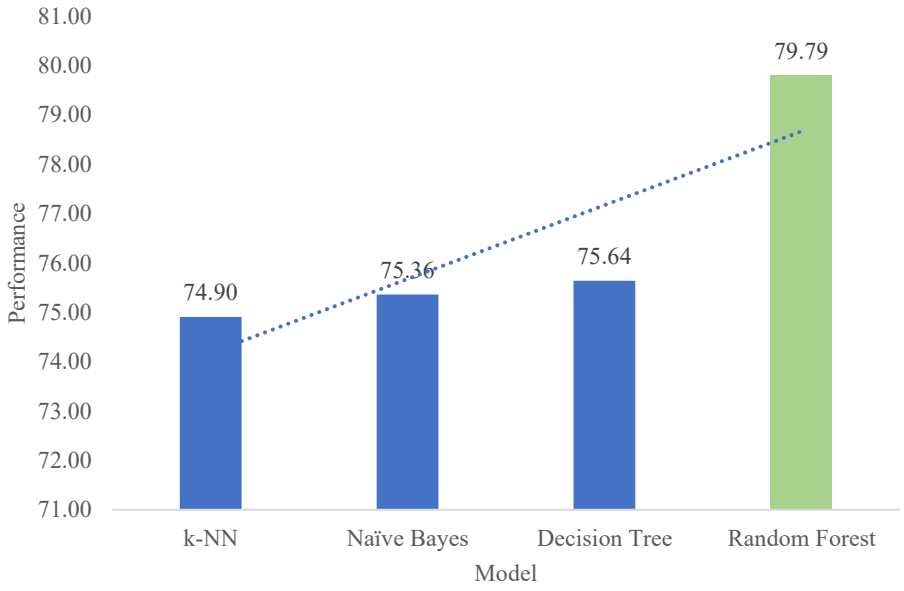


Figure 5. Model performance comparison

The ranking results show that the Random Forest model shows the best performance, followed by Decision Tree. Meanwhile, Naive Bayes and k-NN show the worst performance. This is because Random Forest can reduce the risk of overfitting compared to a single decision tree. This model is very flexible for classification and regression tasks, able to handle high-dimensional data, missing values, and data with noise or outliers effectively.

4. CONCLUSION

Based on the evaluation results, it can be concluded that class imbalance in the breast cancer dataset has a significant impact on the performance of the classification model, especially in terms of the ability to detect positive cases (breast cancer). The SMOTE data balancing technique has proven to be effective in increasing accuracy values. Implementing SMOTE improves accuracy by 11.85%. The Random Forest model consistently provided the best results after applying oversampling techniques, while the k-NN and Naïve Bayes models showed the worst performance. This is because Random Forest can reduce the risk of overfitting compared to a single decision tree. This model is very flexible for classification and regression tasks, able to handle high-dimensional data, missing values, and data with noise or outliers effectively

Class imbalance management strategies must be an integral part of the medical classification system development pipeline, because it can affect the overall interpretability and accuracy of diagnosis. Further research is recommended to evaluate the effectiveness of other imbalance handling techniques such as ensemble-based sampling or cost-sensitive learning, as well as applying these approaches to larger and more complex clinical datasets.

5. ACKNOWLEDGMENTS

The author would like to express his deepest gratitude to the Faculty of Health, Aisyah Pringsewu University (UAP) and STMIK Rosalia Metro, for helping validate the anonymous breast cancer patient data used in this research. We also thank the Machine Learning (ML) Research Group at UAP and STKIP Metro for their valuable input during the model development and evaluation phase. Special awards are given to doctors whose clinical insight contributes significantly to the interpretation of breast cancer prediction results.

REFERENCES

[1] M. F. Ak, “A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications,” *Healthc.*, vol. 8, no. 2, 2020, doi: 10.3390/healthcare8020111.

[2] Y. Cakmak and I. Pacal, “Enhancing Breast Cancer Diagnosis : A Comparative Evaluation of Machine Learning Algorithms Using the Wisconsin Dataset,” *J. Oper. Intell.*, vol. 3, no. 1, pp. 175–196, 2025.

[3] J. Li *et al.*, “Predicting breast cancer 5-year survival using machine learning: A systematic review,” *PLoS One*, vol. 16, no. 4 April, pp. 1–23, 2021, doi: 10.1371/journal.pone.0250370.

[4] C. Karima and W. Anggraeni, “Performance Analysis of the Ada-Boost Algorithm For Classification of Hypertension Risk With Clinical Imbalanced Dataset,” *Procedia Comput. Sci.*, vol. 234, pp. 645–653, 2024, doi: https://doi.org/10.1016/j.procs.2024.03.050.

[5] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Appl. Soft Comput.*, vol. 83, p. 105662, 2019, doi: https://doi.org/10.1016/j.asoc.2019.105662.

[6] 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.

[7] O. N. Oyelade, A. A. Obiniyi, S. B. Junaidu, and S. A. Adewuyi, "ST-ONCODIAG: A semantic rule-base approach to diagnosing breast cancer base on Wisconsin datasets," *Informatics Med. Unlocked*, vol. 10, pp. 117–125, 2018, doi: 10.1016/j.imu.2017.12.008.

[8] 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.

[9] G. Ramadhan and F. D. Adhinata, "Teknik SMOTE dan Gini Score dalam Klasifikasi Kanker Payudara," *J. Perad. Sains, Rekayasa, dan Teknol.*, vol. 9, no. 2, pp. 125–134, 2021.

[10] N. Nurjanah *et al.*, "Implementasi Model Klasifikasi Jenis Kanker Payudara Menggunakan Model SVM dan Logistic Regression berbasis Web," *Ris. dan E-Jurnal Manaj. Inform. Komput.*, vol. 7, no. 4, pp. 1739–1750, 2023, doi: http://doi.org/10.33395/remik.v7i4.12817.

[11] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split Data terhadap Performa Klasifikasi Penyakit Kanker Payudara menggunakan Model Machine Learning," *Jurnali Sainsi dan iInformatika*, vol. 9, no. April, pp. 19–28, 2023, doi: 10.34128/jsi.v9i1.622.

[12] 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.

[13] R. Shakil, B. Akter, F. M. J. M. Shamrat, and S. R. H. Noori, "A novel automated feature selection based approach to recognize cauliflower disease," *Bull. Electr. Eng. Informatics*, vol. 12, no. 6, pp. 3541–3551, 2023, doi: 10.11591/eei.v12i6.5359.

[14] F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, *Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques*. London, 2019. doi: 10.1007/979-981-13-8798-2-12.

[15] 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.

[16] 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.

[17] I. P. Adebayo, "Idowu Peter Adebayo. Predictive Model for the Classification of Hypertension Risk Using Decision Trees Algorithm," *Am. J. Math. Comput. Model.*, vol. 2, no. 2, pp. 48–59, 2017, doi: 10.11648/j.ajmcm.20170202.12.

[18] F. A. Ibrahim and O. A. Shiba, "Data Mining : WEKA Software (an Overview)," *J. Pure Appl. Sci.*, vol. 18, no. 3, pp. 54–58, 2019, [Online]. Available: www.Suj.sebhau.edu.ly