



Improving Classification Performance on Imbalanced Stroke Datasets Using Oversampling Techniques

Muhammad Innuddin^{1*}, Hairani Hairani², M. Thonthowi Jauhari³, Lalu Zazuli Azhar Mardedi⁴

^{1,2,4}Department of Computer Science, Faculty of Engineering, Universitas Bumigora, Mataram, Indonesia

³Department of Nutrition, Faculty of Health, Universitas Bumigora, Mataram, Indonesia

¹inn@universitasbumigora.ac.id, ²hairani@universitasbumigora.ac.id, ³thonthowijauhari@universitasbumigora.ac.id,

⁴zazuli@universitasbumigora.ac.id

Abstract

Stroke is the second leading cause of death globally and significantly contributes to long-term disability. While machine learning techniques have been increasingly used for early stroke detection, their performance is often limited by imbalanced data distributions that bias classification outcomes. This study aims to investigate the effectiveness of three oversampling techniques—SMOTE, Borderline-SMOTE, and SVM-SMOTE—in improving stroke classification performance on imbalanced datasets. Oversampling methods are applied to balance class distributions, followed by the implementation of Random Forest and XGBoost classifiers for stroke prediction. Experimental results demonstrate that oversampling techniques substantially improve classification performance, particularly in the Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC) metrics. Among the tested methods, Borderline-SMOTE yields the best performance, achieving accuracies of 96.45% with Random Forest and 96.41% with XGBoost. Moreover, it increases MCC by 87.51% and AUC by 45.40% for Random Forest, and MCC by 76.52% and AUC by 41.81% for XGBoost, compared without oversampling. The results demonstrate that Borderline-SMOTE effectively addresses data imbalance, enhances model robustness, and improves the detection of minority stroke cases in classification tasks.

Keywords: borderline-SMOTE; imbalance data; SMOTE; stroke prediction; XGBoost method

How to Cite: M. Innuddin, Hairani, M. T. Jauhari, and L. Z. A. Mardedi, “Improving Classification Performance on Imbalanced Stroke Datasets Using Oversampling Techniques”, *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 5, pp. 1193 - 1199, Oct. 2025.

Permalink/DOI: <https://doi.org/10.29207/resti.v9i5.6859>

Received: June 29, 2025

Accepted: October 18, 2025

Available Online: October 26, 2025

This is an open-access article under the CC BY 4.0 License

Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The World Health Organization (WHO) reports that stroke ranks as the second leading cause of death globally and significantly contributes to long-term disability. [1]. Early detection and prompt medical treatment are crucial for reducing the risk of serious complications and death [2]. Recently, machine learning methods have demonstrated promising potential in aiding the medical diagnosis process [3] - [5]. However, the implementation of machine learning still encounters challenges, one of which is data imbalance; In particular, the proportion of non-stroke patients is substantially higher than that of stroke patients [1]. This imbalance causes the predictive model to exhibit a bias in favor of the majority class, resulting in suboptimal performance when predicting cases in the minority class.

An effective strategy for mitigating data imbalance involves employing oversampling techniques, which

enhance the representation of the minority class to achieve a more balanced class distribution [6]. This method reduces the tendency of machine learning models to favor the majority class and improves their capacity to identify patterns within the minority class by augmenting its data representation [7]. By achieving a more balanced class distribution, the model's performance in identifying minority classes—particularly recall and AUC—can be significantly improved. Commonly used oversampling methods include Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE, and SVM-SMOTE, each of which employs distinct strategies to generate synthetic samples.

Previous research [8] performed stroke classification using the Support Vector Machine (SVM) algorithm, attaining an accuracy of 78.86%, a precision of 73.98%, and a recall of 56.75%. However, this study did not apply any resampling methods to address data imbalance. In contrast, another study [9] utilized the

SMOTE oversampling technique in stroke classification with the Random Forest algorithm, resulting in improved accuracy of 93.6%, precision of 91.4%, recall of 96.1%, and an F1-score of 93.7%. Additionally, research [10] employed the Naïve Bayes algorithm for stroke classification, achieving an accuracy of 92.48%, but it did not investigate methods for addressing imbalanced data.

Research [11] addresses the issue of stroke data imbalance by applying SMOTE to improve the performance of the Multivariate Adaptive Regression Splines (MARS) model. The findings indicate that integrating SMOTE with the MARS model effectively improves stroke prediction accuracy, thereby supporting early detection and prevention efforts. In a related study [12], the effect of oversampling on SVM performance in stroke classification was investigated. The results demonstrated that oversampling significantly enhances the predictive accuracy of the SVM model for stroke classification.

Existing research on stroke classification has primarily focused on the application of a single classification algorithm or resampling method, thus limiting a comprehensive understanding of the influence of different oversampling techniques on improving the performance of classification methods. This has resulted in a crucial research gap: the lack of a systematic investigation into the use of multiple resampling methods to address class imbalance in stroke prediction tasks. To bridge this research gap, the present study proposes a novel framework that integrates three oversampling techniques: the SMOTE, SVM-SMOTE, and Borderline-SMOTE, with two classifiers, namely Random Forest and XGBoost, to deeply examine their effectiveness and contribution to predictive performance.

2. Methods

The research flow is essential for ensuring that each process functions systematically, starting with data collection and concluding with the final evaluation of the model. Presenting this process as a flowchart aims to enhance understanding of the workflow and visually depict the relationships among the various steps. In this study, the process comprises several stages: data collection and preprocessing, handling data imbalance, selecting appropriate classification algorithms, training and testing the model, and evaluating its performance. These stages are comprehensively depicted in Figure 1.

This study uses a health dataset from Kaggle.com to predict stroke. The dataset comprises 5110 instances and 11 attributes that capture demographic information and medical conditions potentially associated with stroke risk factors. The features utilized in this study are presented in Table 1.

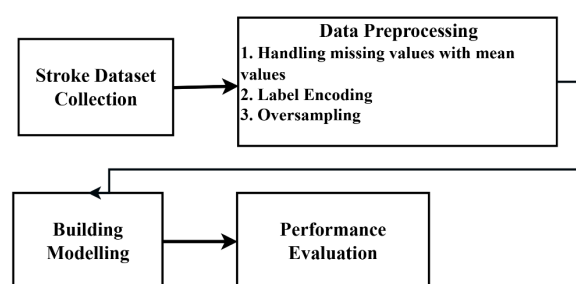


Figure 1. Research Flow

Table 1. Sample Dataset Stroke

No	Gender	Age	Hyperten sion	...	BMI	Stroke
1	Male	67	0	...	36.6	1
2	Female	61	0	...	NA	1
...	...	...	...	...	...	...
5109	Male	51	0	...	25.6	0
5110	Female	44	0	...	26.2	

The next stage involves preprocessing, which includes handling missing values, converting categorical data using label encoding, and balancing the dataset through oversampling. Addressing missing values is crucial to preventing distortion in the results. Label encoding helps in processing non-numeric attributes effectively. Oversampling is employed to mitigate class imbalance between the majority and minority classes within the dataset.

The stroke dataset comprises only 4.87% of cases, leading to class imbalance. To address this issue, researchers applied oversampling techniques using SMOTE, Borderline-SMOTE, and SVM-SMOTE concurrently. SMOTE generates synthetic samples through interpolation among minority instances [13], [14]. Borderline-SMOTE concentrates on the border regions that are susceptible to misclassification [15], while SVM-SMOTE leverages the SVM hyperplane to identify areas for generating synthetic data [16]. After achieving balance in the dataset, the researchers moved on to the classification phase, employing the Random Forest and XGBoost algorithms with default hyperparameters. These algorithms were selected for their capability to handle high-dimensional data and to construct accurate, robust classification models. Random Forest mitigates overfitting by combining predictions from multiple trees. In contrast, XGBoost is known for its speed and performance, particularly with large datasets containing many features. This study examines the effectiveness of integrating oversampling techniques with classification algorithms to improve stroke prediction accuracy, using stratified sampling (80% training, 20% testing) to maintain consistent class distribution.

Researchers evaluated combinations of oversampling techniques and classification algorithms using a confusion matrix and calculated accuracy, MCC, and AUC to assess model performance. AUC was highlighted as a key metric for measuring classification effectiveness under imbalanced data conditions. The

Matthews Correlation Coefficient (MCC) quantifies the correlation between the predicted and actual class labels [17]. Since it accounts for all components of the confusion matrix, MCC is considered one of the most reliable indicators for assessing classification models, particularly in scenarios with imbalanced class distributions. MCC values range from -1 to +1, where +1 signifies perfect predictions, 0 denotes random predictions, and -1 indicates entirely incorrect or opposite predictions relative to the actual conditions. This evaluation enabled researchers to identify the most effective combination of methods for enhancing classification performance on imbalanced data. The formulas for accuracy [18], [19], MCC, and AUC are detailed in Equations (1), (2), and (3). TP and TN represent correct predictions for stroke and non-stroke cases, respectively, while FP and FN indicate incorrect predictions where the model misclassifies stroke presence or absence.

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

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2)$$

$$AUC = \frac{Sensitivity + Specificity}{2} \quad (3)$$

3. Results and Discussions

This section presents the results of the modeling process based on stroke data that underwent preprocessing, along with an analysis of the model's performance. Researchers addressed 202 missing values in the BMI attribute through mean imputation to preserve data quality. Categorical variables were label-encoded for model training, and the dataset was imbalanced, with stroke cases representing only 4.87% of the data. The study addressed the data imbalance—where the model favored the majority class—by applying SMOTE, Borderline-SMOTE, and SVM-SMOTE, as illustrated in Figure 2. Figure 2 depicts the outcomes of the SMOTE, SVM-SMOTE, and Borderline-SMOTE oversampling techniques. Both SMOTE and Borderline-SMOTE yield balanced class distributions, whereas SVM-SMOTE emphasizes the quality of synthetic data near the hyperplane, leading to an unbalanced distribution. Following oversampling, researchers evaluated the performance of Random Forest and XGBoost using an 80% training and 20% test split. In this study, each classification method has specific parameters that serve to optimize model performance. Parameters were determined by adjusting the data characteristics. Details of the classification method parameters used are shown in Table 2. Then, performance metrics were assessed using accuracy, MCC, and AUC.

According to Figure 3, the Random Forest algorithm demonstrates high accuracy in classifying non-stroke patients, achieving 968 true negatives and four false positives. However, it struggles to detect stroke cases, identifying only one true positive and producing 49

false negatives. Such behavior suggests that while the model is reliable for the majority class, it is less effective at recognizing the minority class. Similarly, Figure 4 illustrates the performance of the XGBoost algorithm, which accurately classifies 962 non-stroke patients and five stroke patients but still results in 45 false negatives and 10 false positives. The difficulties in identifying the stroke class in both models highlight the challenges associated with data imbalance, indicating the need for oversampling techniques to enhance the accurate detection of minority classes.

Table 2. Parameters in the Classification Method Used

Parameter of Random Forest		Parameter of XGBoost	
Parameter	Values	Parameter	Values
n_estimators	100	eval_metric	logloss
max_depth	none	max_leaves	0
criterion	gini	max_depth	6
Max_features_leaf	sqrt	learning_rate	0.3
min_samples_split	2	use_label_enc	False
		order	

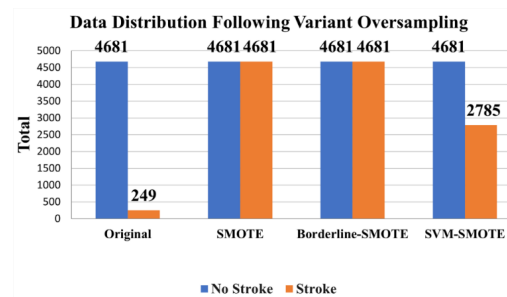


Figure 2. Data Distribution After Oversampling

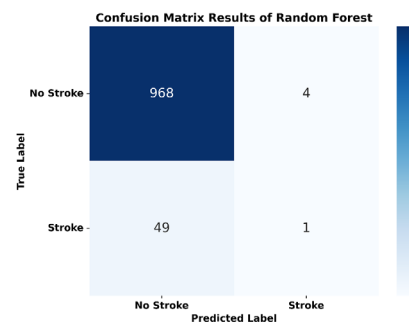


Figure 3. Random Forest Confusion Matrix on Original Data

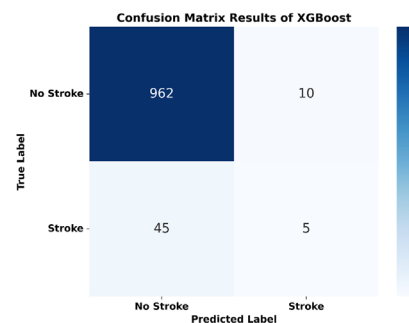


Figure 4. XGBoost Confusion Matrix on Original Data

Based on Figure 5, which displays the confusion matrix of the classification results using the Random Forest algorithm with the SMOTE oversampling technique, the model demonstrates excellent performance in

classifying both classes. A total of 899 instances of the “no stroke” class were accurately classified (true negative), and 941 instances of the stroke class were correctly identified (true positive). There were only 31 cases of stroke misclassified as “no stroke” (false negative) and 74 instances of “no stroke” predicted as stroke (false positive). These findings indicate that using SMOTE significantly enhances the model's ability to detect minority classes, leading to better balance between sensitivity and specificity and more reliable classification of imbalanced datasets.

In Figure 6, the XGBoost model with SMOTE also shows strong classification results, with a high number of correct predictions in both classes. A total of 911 instances of the “no stroke” class were accurately predicted, while only 62 instances were misclassified. The model effectively recognized 939 instances of the stroke class and made only 33 false predictions (false negatives). This result reflects a very high recall rate for the minority class after applying SMOTE to address the imbalance in data distribution. With this performance, the model demonstrates high effectiveness in classification tasks that require sensitivity to the positive class, particularly in the medical field.

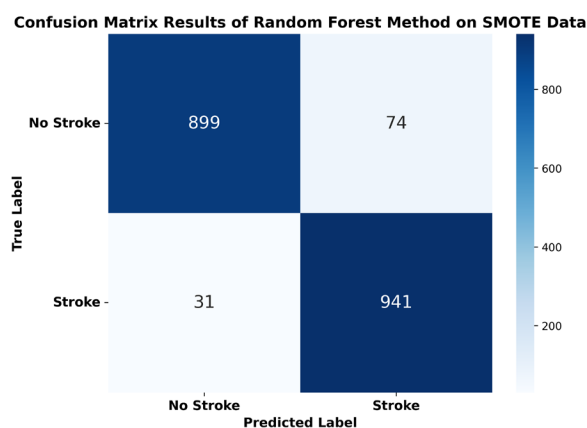


Figure 5. Random Forest Confusion Matrix on SMOTE Data

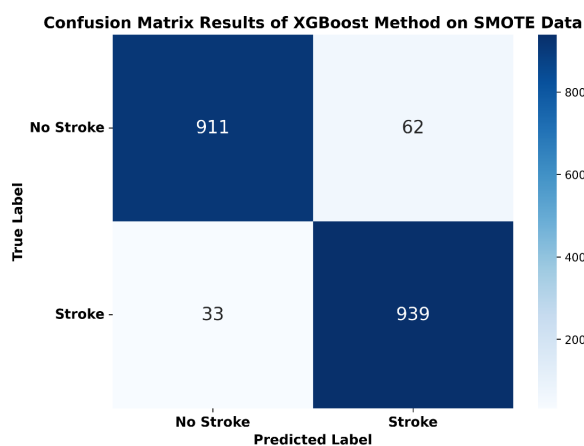


Figure 6. XGBoost Confusion Matrix on SMOTE Data

The Borderline-SMOTE model combined with the Random Forest algorithm, as illustrated in Figure 7, demonstrates excellent classification performance,

achieving high accuracy in detecting both classes. Specifically, 924 instances of the “no stroke” class were correctly classified, while 49 instances were incorrectly categorized as belonging to the stroke class. Additionally, the model successfully recognized 947 instances in the stroke class, with 25 instances mispredicted as false negatives. These results indicate a commendable balance between precision and recall, particularly when addressing imbalanced data. The application of Borderline-SMOTE enhances the model's capability to identify minority samples near the classification boundary, thereby strengthening its generalization ability and rendering this combination reliable for classification tasks involving imbalanced data distributions.

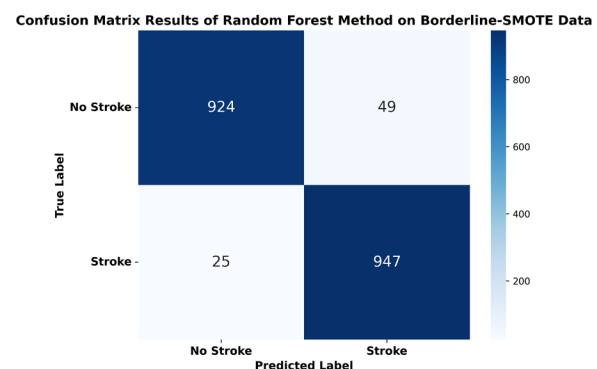


Figure 7. Random Forest Confusion Matrix on Borderline-SMOTE Data

In contrast, the Borderline-SMOTE model utilizing the XGBoost algorithm, as depicted in Figure 8, exhibits even more optimal performance. This model correctly classified 927 instances of the “no stroke” class, with only 46 being misclassified. It also accurately identified 946 instances in the stroke class, with 26 instances mispredicted. This combination achieves excellent levels of accuracy, precision, and recall, bolstered by XGBoost's proficiency in managing complex data and mitigating overfitting. Overall, these findings confirm that the integration of Borderline-SMOTE with XGBoost represents a superior strategy for addressing classification challenges in datasets with imbalanced class distributions.

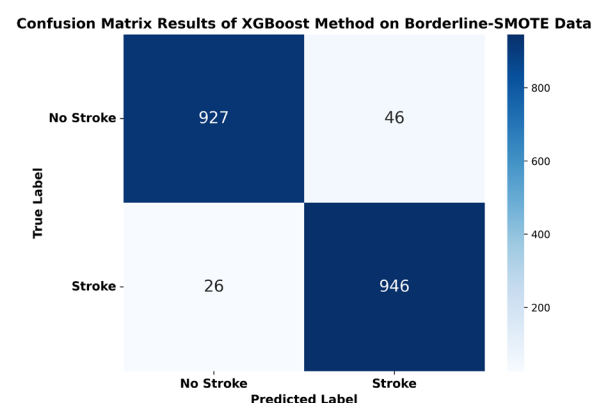


Figure 8. the XGBoost Confusion Matrix on Borderline-SMOTE Data

The SVM-SMOTE model combined with the Random Forest algorithm, as illustrated in Figure 9, demonstrates excellent classification performance. A total of 931 instances in the “no stroke” class were accurately classified, with only 42 instances misclassified. This model also correctly classified 526 instances in the stroke class, while 31 instances were misclassified. The integration of the SVM-SMOTE technique with Random Forest has proven effective in addressing data imbalance, resulting in a precise and stable model that maintains a balance between sensitivity and specificity.

In contrast, the SVM-SMOTE model utilizing the XGBoost algorithm, shown in Figure 10, also exhibits strong performance. It accurately classified 932 instances in the “no stroke” class, with only 41 instances misclassified. Additionally, it successfully classified 526 instances in the stroke class, with 31 instances misclassified. This advantage indicates that the combination of SVM-SMOTE with XGBoost produces a balanced model in terms of accuracy, precision, and recall. With low rates of false positives and false negatives, this model is particularly well-suited for application in medical prediction systems where data may be imbalanced.

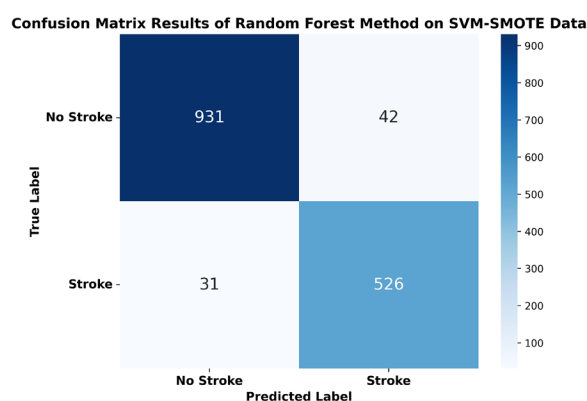


Figure 9. Random Forest Confusion Matrix on SVM-SMOTE Data

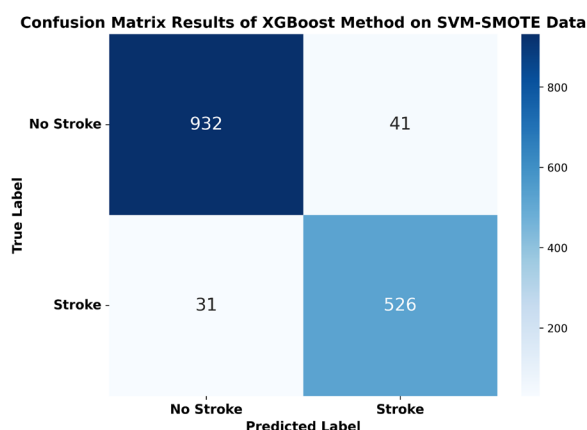


Figure 10. XGBoost Confusion Matrix on SVM-SMOTE Data

The results in Table 3 and Table 4 show that the Random Forest method has a fairly high tendency towards overfitting. This is demonstrated by the

performance on training data, which consistently achieves 100% accuracy, MCC, and AUC across all oversampling variants. Achieving perfect scores indicates that the model tends to memorize training data patterns without adequate generalization capabilities. This condition is increasingly evident in the test results, where the performance of Random Forest decreased, especially on data without oversampling, which only obtained an MCC of 4.91% and an AUC of 50.79%. Although oversampling techniques such as SMOTE, Borderline-SMOTE, and SVM-SMOTE effectively enhanced performance on the test data, the substantial gap between training and testing results provides strong evidence of overfitting.

In contrast, the XGBoost method demonstrated more stable performance and better generalization capabilities. On training data, this model achieved near-perfect results with an accuracy of around 99% and an MCC above 95%, but did not reach 100%. This indicates a mechanism for controlling model complexity through regularization and limiting tree depth. Performance on test data was also relatively consistent, with an accuracy ranging from 94–96%, an MCC between 90–93%, and an AUC above 91%. The relatively small difference between training and testing results confirms that XGBoost can balance learning on training data with generalization on test data, resulting in a lower risk of overfitting compared to Random Forest.

Table 3. Performance Training Results of Classification Methods on Each Oversampling Variant

Methods	Oversampling Variant	Accuracy	MCC	AUC
Random Forest	Without Oversampling	100%	100%	100%
	SMOTE	100%	100%	100%
	Borderline-SMOTE	100%	100%	100%
	SVM-SMOTE	100%	100%	100%
XGBoost	Without Oversampling	99.63%	95.97%	96.23%
	SMOTE	99.27%	98.54%	99.98%
	Borderline-SMOTE	99.33%	98.66%	99.98%
	SVM-SMOTE	99.31%	98.52%	99.98%

Table 4 presents the evaluation results of the Random Forest and XGBoost classification methods, which show that the oversampling technique has a significant effect on model performance, especially in overcoming the problem of imbalanced data. Without using oversampling, both Random Forest and XGBoost recorded high accuracy, namely 94.81% and 94.62%, respectively. However, the resulting MCC and AUC values were very low, namely only 4.91% and 50.79% for Random Forest and 16.10% and 54.49% for XGBoost. These results indicate that both models tend to predict the majority class and fail to identify the minority class well.

Table 4. Performance Test Results of Classification Methods on Each Oversampling Variant

Methods	Oversampling Variant	Accuracy	MCC	AUC
Random Forest	Without Oversampling	94.81%	4.91%	50.79%
	SMOTE	94.59%	89.29%	94.60%
	Borderline-SMOTE	96.41%	92.42%	96.19%
	SVM-SMOTE	95.29%	89.75%	95.06%
	Without Oversampling	94.62%	16.10%	54.49%
XGBoost	SMOTE	95.06%	90.27%	91.11%
	Borderline-SMOTE	96.45%	92.62%	96.30%
	SVM-SMOTE	95.26%	89.88%	95.11%
	SMOTE			

The implementation of oversampling methods, including SMOTE, SVM-SMOTE, and Borderline-SMOTE on Random Forest managed to increase the MCC and AUC values significantly. Among the three techniques, Borderline-SMOTE gave the best results with an accuracy of 96.41%, MCC of 92.42%, and AUC of 96.19%. The SVM-SMOTE technique also showed competitive performance with an MCC of 95.29% and an AUC of 95.06%. In the XGBoost algorithm, The three oversampling techniques resulted in significant improvements in model performance, particularly in MCC and AUC. Among these methods, Borderline-SMOTE achieved the highest accuracy of 96.45%, an MCC of 92.62%, and the highest AUC value of 96.30%. Overall, this technique proved the most effective across both algorithms, maintaining high accuracy while simultaneously improving F1-score and AUC. Therefore, the application of oversampling techniques has been shown to substantially improve the effectiveness of classification models, consistent with previous research [20]–[23] highlighting the effectiveness of such techniques in improving machine learning methods.

Based on the evaluation results using the MCC and AUC metrics on the training data (Table 4) and testing data (Table 5), especially on the original data, both classification methods used showed strong indications of overfitting. This can be seen from the significant difference between performance on training data, which achieved MCC and AUC values above 98%, with performance on testing data which was only below 16% for MCC and 50% for AUC. This significant difference indicates that the model overfits to patterns in the training data but fails to generalize to new data. This condition occurs due to an imbalanced class distribution, where the model prioritizes predictions for the majority class (No Stroke) and ignores the minority class (Stroke). As a result, despite high accuracy, MCC, and AUC values on the training data, the model's performance on the testing data is low, reducing the validity of the classification results.

To overcome this challenge, the present study introduces oversampling techniques to reduce overfitting in stroke classification. Oversampling is

employed to balance the class distribution between the majority and minority classes, ensuring that the model does not solely learn patterns from the majority class but can also effectively recognize patterns in the minority class. The application of oversampling has been shown to enhance performance on metrics sensitive to class imbalance, such as the Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC), achieving MCC and AUC values of at least 89% and 91%, respectively, on the test data. MCC offers a more balanced evaluation of model performance by accounting for both positive and negative classes, whereas AUC reflects the model's overall ability to distinguish between these classes. Consequently, the use of oversampling techniques can improve model generalization, provide a more reliable performance assessment, and reduce the likelihood of overfitting. The results indicate that Borderline-SMOTE outperforms both SMOTE and SVM-SMOTE. Its strength lies in strategically creating synthetic samples around the decision boundary between minority and majority classes[24], [25]. By focusing on these key boundary areas, the method enhances the model's ability to detect patterns in the minority class while reducing noise and data overlap. Consequently, Borderline-SMOTE consistently boosts performance metrics—especially the AUC—and achieves more precise class differentiation. Additionally, XGBoost has demonstrated superiority over Random Forest when classifying imbalanced data, such as stroke. This advantage arises from its boosting approach, which incrementally builds the model while concentrating on correcting errors from previous iterations [26], [27]. Furthermore, XGBoost includes regularization features and is adept at managing class imbalance, leading to more accurate and stable predictions, particularly with complex medical data. Ultimately, the combination of Borderline-SMOTE and XGBoost yields the most effective classification results in this study. This study utilized only a small dataset, necessitating further research with multiple larger datasets to test the oversampling method more reliably. The findings of this study have real-world healthcare implications. By improving the ability of classification models to handle imbalanced stroke datasets, the proposed method could support more accurate identification of patients at risk of stroke, particularly in minority groups often overlooked by conventional models. Furthermore, this approach can be integrated into hospital decision support systems or telemedicine platforms, enabling doctors to make more informed decisions and allocate medical resources more effectively. In the long term, applying these techniques across a wide range of healthcare datasets could improve personalized care strategies, optimize preventive care, and ultimately contribute to better public health outcomes.

4. Conclusions

The research results demonstrate that the application of the oversampling technique significantly enhances the

performance of machine learning methods in stroke classification scenarios. This technique effectively addresses the imbalance in data distribution across classes, making the model more sensitive to the minority class and enabling it to generate predictions with higher accuracy and AUC values. The findings reveal that Borderline-SMOTE outperforms the SMOTE and SVM-SMOTE techniques across all classification methods assessed. The study finds that Borderline-SMOTE significantly improves model accuracy and class differentiation in imbalanced datasets, confirming its effectiveness. Future research should combine oversampling with feature selection and validate results on larger, more diverse datasets.

Acknowledgements

The author would like to thank the Kemdiktisaintek for funding research under the 2025 Beginner Lecturer Research (PDP) scheme.

References

- [1] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021, doi: 10.14569/IJACSA.2021.0120662.
- [2] Z. G. Al-Mekhlafi et al., "Deep Learning and Machine Learning for Early Detection of Stroke and Haemorrhage," *Comput. Mater. Contin.*, vol. 72, no. 1, pp. 775–796, 2022, doi: 10.32604/cmc.2022.024492.
- [3] Y. Chahine, M. J. Magoon, B. Maidu, J. C. del Álamo, P. M. Boyle, and N. Akoum, "Machine Learning and the Conundrum of Stroke Risk Prediction," *Arrhythmia Electrophysiol. Rev.*, vol. 12, 2023, doi: 10.15420/aer.2022.34.
- [4] M. O. Ullah, S. A. Raju, M. I. Nazir, A. Akter, and M. S. Rahman, "An Innovative Machine Learning Pipeline for Stroke Prediction on Imbalanced Data," *2023 Int. Conf. Inf. Commun. Technol. Sustain. Dev. ICICT4SD 2023 - Proc.*, pp. 153–157, 2023, doi: 10.1109/ICICT4SD59951.2023.10303319.
- [5] T. Zuo, F. Li, X. Zhang, F. Hu, L. Huang, and W. Jia, "Stroke classification based on deep reinforcement learning over stroke screening imbalanced data," *Comput. Electr. Eng.*, vol. 114, p. 109069, 2024, doi: <https://doi.org/10.1016/j.compeleceng.2023.109069>.
- [6] A. Vilorio, O. B. Pineda Lezama, and N. Mercado-Caruzo, "Unbalanced data processing using oversampling: Machine Learning," *Procedia Comput. Sci.*, vol. 175, pp. 108–113, 2020, doi: <https://doi.org/10.1016/j.procs.2020.07.018>.
- [7] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Inf. Sci. (Nij.)*, vol. 565, pp. 438–455, 2021, doi: 10.1016/j.ins.2021.03.041.
- [8] S. Rahayu and Y. Yamasari, "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)," *J. Informatics Comput. Sci.*, vol. 5, no. 03, pp. 440–446, 2024, doi: 10.26740/jinacs.v5n03.p440-446.
- [9] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke," *JT J. Tek.*, vol. 12, no. 02, pp. 72–80, 2023.
- [10] A. F. Riany and G. Testiana, "Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes," *J. SAINTEKOM*, vol. 13, no. 1, pp. 42–54, 2023, doi: 10.33020/saintekom.v13i1.352.
- [11] L. R. Safitri, N. Chamidah, and T. Saifudin, "Modeling of Stroke Risk Using Synthetic Minority Oversampling Technique in Multivariate Adaptive Regression Spline Model," *Commun. Math. Biol. Neurosci.*, vol. 2025, pp. 1–10, 2025, doi: 10.28919/cmbn/9099.
- [12] L. A. Martini, G. A. Pradipta, and R. R. Huizen, "Analysis of the Impact of Data Oversampling on the Support Vector Machine Method for Stroke Disease Classification," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 2, pp. 404–421, 2025, doi: 10.35882/jeeemi.v7i2.698.
- [13] Y. B. Wah et al., "Machine Learning and Synthetic Minority Oversampling Techniques for Imbalanced Data: Improving Machine Failure Prediction," *Comput. Mater. Contin.*, vol. 75, no. 3, pp. 4821–4841, 2023, doi: 10.32604/cmc.2023.034470.
- [14] H. Hairani, T. Widiyaningtyas, D. D. Prasetya, and A. Aminuddin, "Addressing Imbalance in Health Datasets: A New Method NR-Clustering SMOTE and Distance Metric Modification," *Comput. Mater. Contin.*, vol. 82, no. 2, pp. 2931–2949, 2025, doi: 10.32604/cmc.2024.060837.
- [15] M. Mujahid et al., "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00943-4.
- [16] R. Qaddoura, A. M. Al-Zoubi, I. Almomani, and H. Faris, "A multi-stage classification approach for iot intrusion detection based on clustering with oversampling," *Appl. Sci.*, vol. 11, no. 7, 2021, doi: 10.3390/app11073022.
- [17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, pp. 1–13, 2020.
- [18] N. A. Azhar, M. S. Mohd Pozi, A. Mohamed Din, and A. Jatowt, "An Investigation of SMOTE based Methods for Imbalanced Datasets with Data Complexity Analysis," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2022, doi: 10.1109/TKDE.2022.3179381.
- [19] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Appl. Soft Comput.*, vol. 143, p. 110415, 2023, doi: 10.1016/j.asoc.2023.110415.
- [20] H. Hairani, T. Widiyaningtyas, and D. Dwi Prasetya, "Addressing Class Imbalance of Health Data: A Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 3, pp. 1310–1318, 2024, doi: <https://dx.doi.org/10.62527/joiv.8.3.2283>.
- [21] L. G. R. Putra, K. Marzuki, and H. Hairani, "Correlation-based feature selection and Smote-Tomek Link to improve the performance of machine learning methods on cancer disease prediction," *Eng. Appl. Sci. Res.*, vol. 50, no. 6, pp. 577–583, 2023, doi: 10.14456/easr.2023.59.
- [22] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, "Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction," *Inf. Softw. Technol.*, vol. 139, no. June, p. 106662, 2021, doi: 10.1016/j.infsof.2021.106662.
- [23] M. Khushi et al., "A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.
- [24] P. Wibowo and C. Fatichah, "An in-depth performance analysis of the oversampling techniques for high-class imbalanced dataset," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 7, no. 1, pp. 63–71, 2021, doi: 10.26594/register.v7i1.2206.
- [25] H. Sun, J. Li, and X. Zhu, "A Novel Expandable Borderline Smote Over-Sampling Method for Class Imbalance Problem," *IEEE Trans. Knowl. & Data Eng.*, vol. 37, no. 05, pp. 2183–2199, May 2025, doi: 10.1109/TKDE.2025.3544284.
- [26] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, pp. 1–40, 2025.
- [27] Z. Shao and M. N. Ahmad, "Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface," *Remote Sens.*, vol. 16, no. 4, pp. 1–18, 2024.