BAB III

ARTIKEL KARYA ILMIAH

Imbalanced Data Handling for Strokeee Prediction Using Oversampling and Cost-Sensitive Learning

1st Triyan Agung Laksono
Magister Teknologi Informasi
Universitas Teknologi Digital Indonesia
Bantul, Indonesia
triyan31@stiesbi.ac.id

2nd Widyastuti Andriyani
Magister Teknologi Informasi
Universitas Teknologi Digital Indonesia
Bantul, Indonesia
widya@utdi.ac.

***Abstract*—Predictive analysis of stroke using machine learning (ML) is a promising approach for early detection and reducing the number of stroke patients. However, the inherent class imbalance in medical datasets poses a significant challenge, often causing models to fail to detect certain minority cases, such as stroke. This study aims to evaluate and compare two popular techniques for addressing class imbalance: oversampling using the Synthetic Minority Oversampling Technique (SMOTE) and cost-sensitive learning, within the context of stroke prediction. Using the public Kaggle stroke dataset, three ML algorithms (Random Forest, Support Vector Machine, and XGBoost) were trained and tested in three scenarios: baseline (without balancing), SMOTE, and cost-sensitive learning. The results show that both balancing techniques significantly improve recall for the minority class, particularly in the SVM model, but at the cost of reduced precision and accuracy across the entire model. Feature importance analysis using SHAP identified age and hypertension as the most important factors in predicting stroke, consistent with previous research findings. Despite these improvements, this study highlights the trade-off between sensitivity and precision, which must be considered for practical application in medical decision support systems. Future research should explore hybrid approaches and validate results on larger and more diverse datasets.**

***Keywords*—Stroke prediction, class imbalance, SMOTE, cost-sensitive learning, machine learning, feature importance, SHAP**

I. INTRODUCTION

Stroke is one of the leading causes of death and permanent disability worldwide, posing a significant social and economic burden, especially in developing countries [1]. Early prevention of stroke risk is essential to reduce mortality and improve the quality of patients' health. The development of machine learning (ML) methods has shown promising results in predicting stroke risk based on medical data [2], [3].

In addition to the model used in this study, various machine learning approaches such as Decision Tree, Logistic Regression, and ensemble techniques (stacking, bagging) have also been widely applied in stroke prediction using clinical data [2], [3]. The studies report promising results, but still face challenges related to detecting minority cases when the data is imbalanced [4], [5].

However, the effectiveness of ML-based prediction models is greatly influenced by the characteristics of the data used. One of the main challenges in medical diagnosis using ML is the problem of class imbalance, where the number of stroke cases is much less than the number of non-stroke instances [5]. This imbalance often makes the model biased towards the majority class and unable to detect minority cases optimally[4].

Various techniques have been proposed to overcome this problem, especially

oversampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) [6], [7] and [8], [9] cost-sensitive learning. SMOTE aims to improve model performance by creating mock data on minority classes, while cost-sensitive learning adjusts classification errors to make the model more sensitive to minority classes.

Many studies have explored these techniques in the context of stroke prediction, but few studies have directly compared the effectiveness of oversampling learning and cost-sensitive learning using public stroke datasets. This study aims to fill that gap by evaluating both methods using several ML algorithms, as well as providing a comparative analysis of their impact on stroke prediction.
The main contributions of this study are as follows:

- Implement and compare oversampling (SMOTE) and cost-sensitive learning techniques for stroke prediction using Random Forest, SVM, and XGBoost.Presents a comprehensive evaluation using public datasets from Kaggle, highlighting the trade-offs between recall, precision, and overall model accuracy.
- Analyzes the importance of features and discusses their implications for real-world medical decision support systems.

II. METHODOLOGY

This study uses the Kaggle public stroke dataset, which consists of clinical data and patient demographics. The research stages include data pre-processing, classification model training, data imbalance problem handling, and model performance evaluation.

A. Data Pre-processing

Data processing is done to prepare the dataset that will be used for the machine learning model. The steps include:

- Deletion of ID columns that are not relevant for prediction.
- Missing value handling by calculating the missing BMI data based on the average value.
- Transforming category variables using one-hot encoding to convert categories into binary features.
- Normalizing numeric features using StandardScaler to provide uniform feature values, especially for models like Support Vector Machine (SVM).

The dataset was then divided into training and testing sets using stratified sampling with a ratio of 80:20, to maintain the class distribution of stroke and non-stroke cases.
Ahead of the model-building process, exploratory data analysis (EDA) was conducted to assess the distribution of key variables such as age, glucose level, and BMI, and detect outliers or missing values that may affect the prediction results. The relationship between key features was analyzed to reduce the possibility of multicollinearity, as recommended by previous research [3]. The selection of input features also considers features that are considered important based on the results of SHAP analysis, which has been widely used in stroke prediction research to ensure transparent results [10].

B. Classification Model

The three main algorithms used are Random Forest, Support Vector Machine (SVM), and XGBoost, which have been proven effective in medical data prediction in previous studies. The selection of this model was also based on previous research trends, highlighting that ensemble models such as Random Forest and XGBoost consistently provide high accuracy in health data prediction [1], [2]. Nevertheless, SVM is still widely used for binary classification in disease diagnosis, despite its higher sensitivity to imbalance compared to ensemble models [3]. To ensure optimal results, each model was trained on data that had undergone value replacement, encoding, and normalization processes in accordance with best practices from previous studies.

C. Handling Data Imbalance

Two main approaches are applied:

- SMOTE (Synthetic Minority Oversampling Technique) [11]: Adding synthetic samples to minority classes.
- Cost-Sensitive Learning [8], [9]: Adjusting the classification error to

make the model more sensitive to stroke cases.

D. Model Evaluation

Evaluation was performed using accuracy, precision, recall, F1-score, and AUC (area under the curve) metrics in the baseline, oversampling, and cost-sensitive learning scenarios. Feature evaluation was performed by utilizing SHAP (Shapley Additive exPlanations).

III. RESULTS AND DISCUSSION

A. Experiment Results

Table 1 shows a comparison of model performance for baseline, SMOTE oversampling, and cost-sensitive learning. In the baseline, all models had high accuracy but low recall in the minority class (stroke), where SVM was unable to detect stroke cases (recall = 0%). SMOTE learning and cost-sensitive learning improved recall, especially for SVM (from 0% to 50% and 68%, respectively), however, the improvement was accompanied by lower precision and accuracy, which is a classic trade-off that occurs in imbalanced medical classification problems. In particular, cost-sensitive learning yields more stable AUC and precision across all models, making it more advisable to use in situations where minimizing false negatives is critical.

Table 1 Comparison of Model Performance on Test Data

Model	Imbalance Handling	Accuracy	Precision	Recall	F sc
Random Forest	Baseline	0.949	0.25	0.02	0.
SVM	Baseline	0.951	0.00	0.00	0.
XGBoost	Baseline	0.943	0.25	0.08	0.
Random Forest	SMOTE	0.915	0.15	0.16	0.15
SVM	SMOTE	0.815	0.13	0.50	0.209
XGBoost	SMOTE	0.920	0.12	0.10	0.109
Random Forest	Cost-Sensitive	0.949	0.00	0.00	0.00
SVM	Cost-Sensitive	0.774	0.14	0.68	0.227
XGBoost	Cost-Sensitive	0.926	0.16	0.12	0.136

Figure 1 ROC Curve Model Baseline

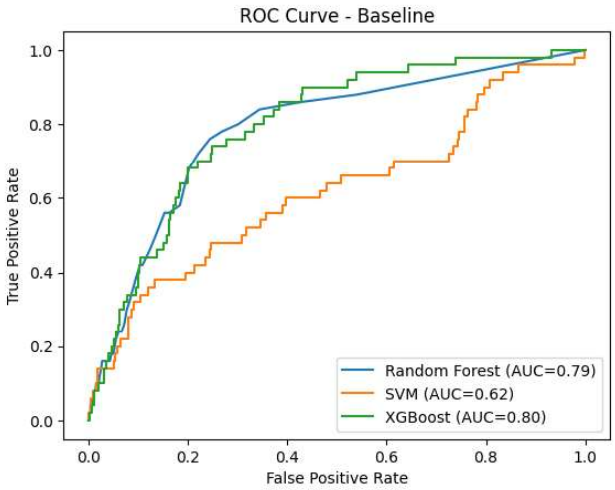


Figure 1 shows the ROC curves of the baseline models. Random Forest and XGBoost have quite high AUC values (0.79 and 0.80), compared to SVM, which has a low performance with an AUC of only 0.62, indicating that SVM has difficulty in recognizing stroke patients in unbalanced data conditions.

Figure 2 ROC Curve Model with SMOTE

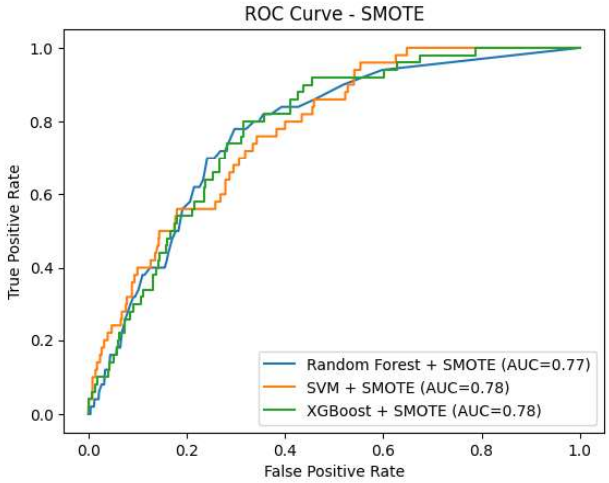


Figure 2 displays the effect of SMOTE oversampling on the ROC curve. The AUC of SVM significantly increases to 0.78, indicating an increase in sensitivity in minority cases. Meanwhile, Random Forest and XGBoost can maintain stable AUC values, but the overall increase in recall is followed by a decrease in precision, which is a common trade-off that occurs with oversampling techniques.

Figure 3 ROC Curve Model Cost-Sensitive

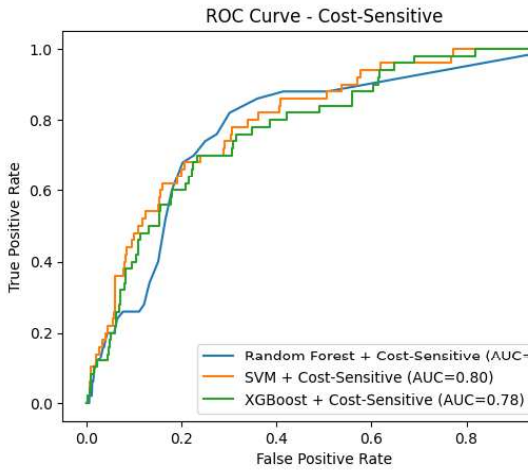


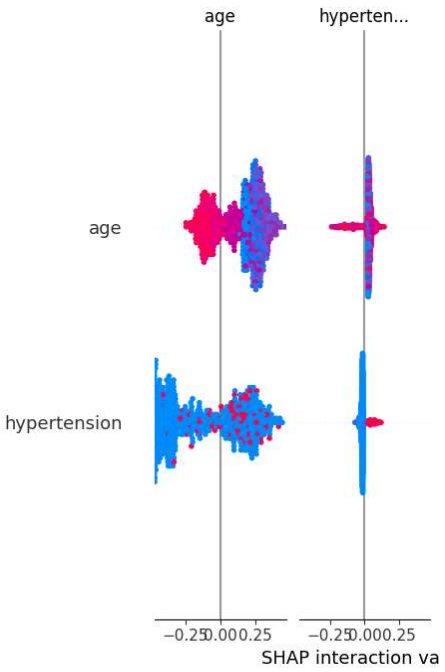
Figure 3 shows the ROC curve of cost-sensitive learning. In this model, SVM obtained the highest AUC of 0.80, and the balance between recall and precision was better across models. This supports the assertion that cost-sensitive approaches are very effective in minimizing false negatives without sacrificing overall performance, making them very suitable for medical decision support.

These findings emphasize the importance of selecting an appropriate data balancing method in stroke prediction. While SMOTE and cost-sensitive learning can both improve the detectability of minority classes, cost-sensitive learning offers a stronger trade-off between recall, precision, and AUC, especially for SVM methods, making it more implementable in scenarios where missing stroke cases can have fatal consequences.
In addition to quantitative performance, model capability was assessed using SHAP analysis. The results show that age and hypertension are the main factors that have the most influence on stroke prediction, which is consistent with previous studies. The ability to explain is critical in building trust among healthcare professionals and aiding the implementation of machine learning-based decision-making systems in the real world.

B. Feature Analysis and Practical Implications

Examination of the salient features using SHAP, as shown in Figure 4, indicates that age and hypertension are the two most important factors in stroke prediction. This is in line with previous medical research that emphasizes age and history of chronic diseases as the main risk factors for stroke [1], [2]. Other factors, such as glucose levels and BMI, also contributed, although not as much as the two main factors.
Previous studies have also shown that oversampling techniques such as SMOTE can significantly improve recall in minority classes, especially in the case of stroke prediction, but a decrease in precision is a consequence that must be considered. This decrease has the potential to increase the number of false positives, which can impact medical workload and patient anxiety [4], [5]. For this reason, some researchers suggest evaluating medical utility models using net benefit or decision curve analysis, so that not only the statistical aspects can be measured, but also the practical impact in the medical setting [12].

Figure 4 SHAP-based Feature Importance



The SHAP visualization in Figure 4 also clarifies the strong relationship between age and hypertension in influencing stroke prediction outcomes. These findings further reinforce that the combination of these two factors is a key indicator for stroke risk, as reported in previous studies[4], [5]. From a medical practice perspective, understanding these most influential features is crucial to helping healthcare professionals detect and intervene more effectively in high-risk patient

groups. Additionally, the use of explainability visualizations like SHAP can enhance transparency and strengthen doctors' trust in the machine learning-based prediction systems they use[6].

A limitation of this study is the very small number of stroke cases in the dataset (~5%); therefore, the effectiveness of the balancing model cannot be fully generalized to the entire population. In addition, the decrease in precision after balancing requires special attention in the implementation of medical decision support systems, as the risk of false positives may impact resource allocation and patient concerns [3], [4].

IV. CONCLUSION

This study compares the effectiveness of oversampling (SMOTE) and cost-sensitive learning techniques in dealing with imbalanced data problems in machine learning-based stroke prediction. The results show that both approaches can improve the model's ability to detect stroke cases (recall), especially in SVM algorithms, although it is often followed by a decrease in precision and accuracy. Based on the feature importance analysis using SHAP, age and hypertension are the main risk factors for stroke, consistent with the findings of previous case studies.

The limitation of this study lies in the limited amount of stroke data, so the results of this generalization still need to be tested on a larger and more diverse dataset. For future research, researchers are expected to explore hybrid techniques (e.g., a combination of SMOTE and cost-sensitive learning) and test their effectiveness in real-world applications

REFERENCES

[1] T. Vu *et al.*, "Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study," *J Cardiovasc Dev Dis*, vol. 11, no. 7, p. 207, Jul. 2024, doi: 10.3390/jcdd11070207.

[2] P. Chakraborty *et al.*, "Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing," *BMC Bioinformatics*, vol. 25, no. 1, p. 329, Oct. 2024, doi: 10.1186/s12859-024-05866-8.

[3] W. Heseltine-Carp *et al.*, "Machine learning to predict stroke risk from routine hospital data: A systematic review," *Int J Med Inform*, vol. 196, p. 105811, Apr. 2025, doi: 10.1016/j.ijmedinf.2025.105811.

[4] K. Moulaei, L. Afshari, R. Moulaei, B. Sabet, S. M. Mousavi, and M. R. Afrash, "Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models," *Sci Rep*, vol. 14, no. 1, p. 31392, Dec. 2024, doi: 10.1038/s41598-024-82931-5.

[5] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*, vol. 2, p. 100116, Nov. 2022, doi: 10.1016/j.health.2022.100116.

[6] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf Sci (N Y)*, vol. 465, pp. 1–20, Oct. 2018, doi: 10.1016/j.ins.2018.06.056.

[7] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci Rep*, vol. 11, no. 1, p. 24039, Dec. 2021, doi: 10.1038/s41598-021-03430-5.

[8] F. Feng, K.-C. Li, J. Shen, Q. Zhou, and X. Yang, "Using Cost-Sensitive Learning and Feature Selection Algorithms to Improve the Performance of Imbalanced Classification," *IEEE Access*, vol. 8, pp. 69979–69996, 2020, doi: 10.1109/ACCESS.2020.2987364.

[9] B. Krawczyk, M. Woźniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Appl Soft Comput*, vol. 14, pp. 554–562, Jan. 2014, doi: 10.1016/j.asoc.2013.08.014.

[10] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," *J Healthc Eng*, vol. 2021, pp. 1–12, Nov. 2021, doi: 10.1155/2021/7633381.

[11] A. Saad Hussein, T. Li, C. W. Yohannese, and K. Bashir, "A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE," *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, p. 1412, 2019, doi: 10.2991/ijcis.d.191114.002.

[12] F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, "The most efficient machine learning algorithms in stroke prediction: A systematic review," *Health Sci Rep*, vol. 7, no. 10, Oct. 2024, doi: 10.1002/hsr2.70062.