Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

Optimizing Ensemble Learning Models with SMOTE-ENN for Early Stroke Detection in Imbalanced Clinical Datasets

Dina Nurmala*, Angga Bayu Santoso

Faculty of Engineering and Computer Science, Information Systems, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia
Email: [1,*]dina_nurmala@teknokrat.ac.id, [2]anggabayu@teknokrat.ac.id
Correspondence Author Email: dina_nurmala@teknokrat.ac.id
Submitted: **03/02/2026**; Accepted: **05/03/2026**; Published**: 06/03/2026**

Abstract−Stroke remains a leading cause of mortality and long-term disability worldwide, including in Indonesia, highlighting the urgent need for early risk identification. Machine learning models for stroke prediction often suffer from severe class imbalance, where stroke cases constitute only 4.9% of clinical datasets, leading to biased predictions that favor the majority class. This study evaluates three ensemble and kernel-based algorithms Random Forest, XGBoost, and Support Vector Machinecombined with two resampling strategies (SMOTE and SMOTE-ENN) using the Healthcare Stroke Dataset (5,110 records, 11 clinical attributes). To prevent data leakage, resampling was strictly applied within each training fold of 5-fold stratified cross-validation, while all evaluations were conducted on the original imbalanced test set. The results demonstrate that XGBoost integrated with SMOTE-ENN achieved the highest minority-class sensitivity, improving PR-AUC by 23.5% (0.1537 vs. 0.1244 with SMOTE alone), while detecting 24% of stroke cases (12 out of 50) in the test set. Although cross-validation results indicate strong class discrimination with AUC-ROC values above 0.98, the low PR-AUC reflects the operational challenge of extreme class imbalance and the inevitable trade-off between recall and precision, resulting in an increased number of false positives. Consequently, the proposed model is best positioned as a first-tier population screening tool that flags high-risk individuals for confirmatory clinical diagnostics, rather than as a standalone diagnostic system. The approach maintains computational efficiency (training time < 0.12 seconds) and substantially improves model stability, evidenced by a 73% reduction in cross-validation variance. These findings support the integration of hybrid resampling techniques with ensemble learning as a practical and scalable framework for early stroke risk screening in resource-constrained primary healthcare settings.

Keywords: Machine Learning; Ensemble Learning; XGBoost; SMOTE-ENN; Stroke Prediction; Class Imbalance

1. INTRODUCTION

Stroke is a major non-communicable disease that significantly contributes to global mortality and disability. The World Health Organization (WHO) reports that each year, over 15 million people worldwide suffer strokes, resulting in approximately 5 million deaths and permanent disabilities for survivors[1][2][3]. In Indonesia, the incidence of stroke is increasing due to lifestyle changes, poor eating habits, and lack of exercise. Data from the Basic Health Research (Riskesdas) show a prevalence exceeding 12.1 per thousand among those aged 15 and over, with a steady upward trend[1][2]. This underscores stroke as both a health crisis and a burden on the national healthcare framework, demanding robust prevention and treatment strategies.

Effective early identification of high-risk individuals before severe symptoms appear is crucial. Traditional methods involve clinical and laboratory tests, including blood pressure checks, cholesterol measurements, and BMI calculations. However, these approaches are time-consuming, expensive, and require professional medical personnel. In primary care facilities in Indonesia, especially in rural areas, a lack of diagnostic tools and staff hinders optimal risk assessment. In this study, artificial intelligence (AI), particularly machine learning, offers solutions for automated and efficient disease prediction and detection[4][5].

Machine learning excels at uncovering patterns in medical data and creating predictive models to assist in clinical decision-making[6]. For stroke prediction, algorithms like Random Forest, Support Vector Machine (SVM), and XGBoost are preferred because they can handle complex data and provide robust classification results[7]. However, the main obstacle is the class imbalance in the medical dataset, where non-stroke cases are far more numerous than stroke cases. This causes the model to favor the majority class, thus reducing its ability to identify critical minority instances[8].

Such imbalances can lead to misleading accuracy, where the model appears effective by correctly predicting the dominant class but fails to detect stroke patients. To address this, this study utilizes the Synthetic Minority Oversampling Technique (SMOTE) and introduces SMOTE-ENN, a combined method that integrates SMOTE synthetic oversampling with Edited Nearest Neighbors (ENN) for selective undersampling. ENN removes noisy or ambiguous samples by deleting those misclassified by most of their nearest neighbors, sharpening decision boundaries, and improving model reliability. This hybrid approach enhances sensitivity to minority classes without sacrificing stability or risking overfitting due to excessive data replication.

This study investigates whether ensemble learning methods particularly XGBoost and Random Forest demonstrate superior minority-class sensitivity when integrated with hybrid resampling (SMOTE-ENN) compared to conventional oversampling (SMOTE) for stroke risk prediction. This study examines wheter the noise-cleaning mechanism of ENN can sharpen decision boundaries, thereby improving recall without substantially increasing false positivesa critical trade-off in clinical screening contexts. The models are assessed based on metrics sensitive to imbalance, such as AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and PR-AUC (Area Under

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

the Precision-Recall Curve), with five-fold stratified cross-validation to ensure consistent performance and protect against overfitting. Computational efficiency (training and prediction time) is also evaluated to assess practicality for deployment in resource-constrained primary healthcare settings.

The main goal of this study is to develop a machine learning model for early stroke prediction that effectively addresses severe class imbalance while producing accurate, stable, and computationally efficient results. This research compares the effectiveness of SMOTE and SMOTE-ENN across multiple ensemble learning models to identify the most reliable configuration for imbalanced clinical data. In this study, optimization is defined as the systematic selection of the best-performing model resampling configuration based on imbalance-sensitive evaluation metrics (PR-AUC and AUC-ROC), cross-validation stability, and computational efficiency, rather than hyperparameter tuning.

Benefits include advancing data-driven healthcare systems scientifically and practically. Scientifically, this deepens insights into imbalance techniques like SMOTE-ENN for stroke. Practically, it informs clinical decision support systems (CDSS) for rapid, accurate, and fair stroke risk screening, enabling timely prevention and intervention to reduce mortality and disability[9].

Despite the rapid growth of machine learning applications for stroke prediction, critical research gaps persist when examining four recent studies. Wijaya et al. (2024) [10] and Swain et al. (2024) [4] reported exceptionally high accuracies (98.24% and 98.76%, respectively) using ensemble methods with SMOTE, yet both evaluated models on artificially balanced test sets, thereby masking true operational performance under natural stroke prevalence (4.9%). Similarly, Melnykova et al. (2025) [11] achieved 90% F1-score with Random Forest but failed to disclose minority-class sensitivity on imbalanced test data where recall typically collapses. While Kivrak et al. (2024) [12] appropriately analyzed SMOTE's impact on sensitivity–specificity trade-offs (41% → 52%), their work focused on testosterone deficiency rather than stroke prediction, limiting direct clinical applicability. Collectively, these studies share a methodological limitation: evaluation on balanced test sets inflates performance metrics and fails to reflect real-world screening conditions where stroke prevalence remains extremely low. Furthermore, none systematically compared conventional SMOTE with hybrid SMOTE-ENN to assess noise-cleaning effects on decision-boundary stability under authentic class imbalance.

This study addresses these gaps by (1) rigorously evaluating all models on the original imbalanced test set preserving 4.9% stroke prevalence, (2) conducting the first direct comparison between SMOTE and SMOTE-ENN for stroke detection with quantified trade-offs in PR-AUC (+23.5% improvement) and cross-validation stability (73% variance reduction), and (3) positioning the model transparently as a first-tier screening tool not a diagnostic instrument to enable realistic clinical deployment in resource-constrained primary healthcare settings.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research uses a quantitative approach with a machine learning-based experimental method to build an early stroke prediction model. The research stages are carried out systematically, starting from data collection, pre-processing, handling class imbalance, model training, and performance evaluation using various metrics. Each stage is interconnected and forms a structured research flow as shown in Figure 1.

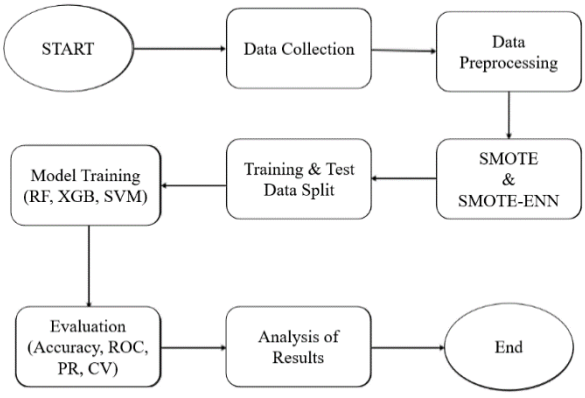

Figure 1. Stages of Early Stroke Detection Research

Based on Figure 1, the research methodology comprises seven sequential phases beginning with stroke-related data collection from international health databases, followed by preprocessing to address missing BMI values and encode categorical variables appropriately. To mitigate the severe class imbalance inherent in the dataset—where stroke cases represent only 4.9% of records—two resampling strategies were implemented: the Synthetic Minority Oversampling Technique (SMOTE) for generating synthetic minority samples through interpolation, and the hybrid

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

SMOTE-ENN approach that further refines the balanced dataset by applying Edited Nearest Neighbors (ENN) to remove noisy or ambiguous instances and sharpen decision boundaries.

These methods are applied exclusively to the training dataset after the initial data split to avoid the risk of information leakage. Three classification algorithms – Random Forest, XGBoost, and SVM – were thoroughly trained and tested under both resampling conditions. Evaluation is done using metrics such as AUC-ROC, PR-AUC, and accuracy, along with visual tools like confusion matrices, ROC curves, and precision-recall curves. Efficiency is also evaluated by tracking the computation duration. To enhance the model's reliability and applicability, the entire procedure incorporates 5-fold stratified cross-validation, ensuring consistent and effective early detection of stroke risk.

2.2 Data Collection

The research data was obtained from the Healthcare Dataset – Stroke Data, which is publicly available on the Kaggle platform. This dataset was curated by Federico Soriano and has been widely used in global stroke prediction studies. The dataset contains 5,110 patient medical records with 11 predictor attributes and 1 target label (stroke). Attributes include: gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, and smoking status. The stroke label is binary (0 = never had a stroke, 1 = ever had a stroke). This dataset was chosen for its completeness of clinical features, transparency of source, and relevance to the primary healthcare context.

2.3 Data Preprocessing

Preprocessing is a critical step for improving data quality and model performance. First, a check was performed for missing values. The BMI column was found to have 201 missing values (≈3.9%). Given that the BMI distribution is right-skewed and susceptible to outliers, missing values were imputed using the median to preserve central tendency without distortion. Second, categorical features (gender, ever_married, work_type, Residence_type, smoking_status) were encoded using an algorithm-specific strategy to ensure methodological validity. For tree-based models (Random Forest and XGBoost), Label Encoding was applied using the Label Encoder from scikit-learn. This approach is appropriate because decision tree-based algorithms split data using threshold-based rules rather than distance calculations. As a result, the artificial ordinal values introduced by Label Encoding do not bias model learning, since tree-based models are invariant to monotonic transformations of input features.

For the Support Vector Machine (SVM) model, One-Hot Encoding was applied using pd.get_dummies(). This choice is necessary because SVM relies on Euclidean distance to construct the decision boundary. Using Label Encoding for non-ordinal categorical variables would impose false numerical ordering (e.g., interpreting category “3” as greater than “1”), thereby distorting geometric relationships in feature space and biasing the classification process. This algorithm-specific encoding strategy avoids unnecessary feature expansion for tree-based models while preserving geometric integrity for distance-based learning in SVM. Given the modest number of original features, the resulting dimensionality increase from One-Hot Encoding remained limited and did not negatively affect model performance.

2.4 Data Division

The dataset is divided into training data (80%) and test data (20%) using the train_test_split function with the parameter stratify=y. Stratification ensures that the proportion of stroke and non-stroke classes remains consistent in both subsets, preventing the model from being biased toward a specific class distribution during training or testing. This division also uses random_state=42 to ensure the replicability of the experiment.

2.5 Handling Class Imbalance (SMOTE and SMOTE-ENN)

Initial analysis revealed severe class imbalance in the dataset: only 4.9% (249 individuals) experienced a stroke, compared to 95.1% (4,861 individuals) who did not. This imbalance causes the model to predominantly label all instances as non-stroke, resulting in high accuracy but recall approaching zero a dangerous flaw in healthcare settings. To address this, this study uses two balancing methods:

a. SMOTE (Synthetic Minority Oversampling Technique) is a data balancing technique that addresses class imbalance by generating synthetic samples for the minority class, not through random duplication, but through linear interpolation between the original sample and its nearest neighbors in feature space[13]. In the stroke dataset, which was initially highly imbalanced (with a ratio of ~1:10 between stroke and non-stroke classes), applying SMOTE with the parameter k=5 successfully increased the number of minority class samples from 249 to 1,928, achieving perfect 1:1 balance. This approach has advantages over simple oversampling because it reduces the risk of overfitting by more realistically expanding the decision region of the minority class, while also preserving the variation in the data. However, SMOTE also has limitations: on datasets containing noise or outliers, interpolation can amplify unrepresentative patterns. Therefore, combining data cleaning techniques such as ENN (Edited Nearest Neighbors) is often necessary to achieve a balance that is not only quantitative but also qualitative, as adopted in this study through the hybrid SMOTE-ENN approach[14].
b. SMOTE-ENN (SMOTE–Edited Nearest Neighbors): This integrated approach combines SMOTE oversampling with targeted undersampling through the ENN (Edited Nearest Neighbors) method. ENN removes instances (from

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

both classes) that are misclassified by most of their nearest neighbors (k=3 by default), thus reducing noise and clarifying decision boundaries. As a result, class balance is maintained, but with a tighter dataset and improved pattern fidelity[15].

These strategies are applied exclusively to a subset of the training data after the data split, adhering to standard protocols to avoid data leakage. This ensures fair learning of stroke-related patterns while reducing the risk of overfitting due to excessive replication or outlier data points.

2.6 Data Normalization

The SVM algorithm is very sensitive to feature scaling because it uses Euclidean distance in margin calculations. Therefore, standardization was performed using StandardScaler, which transforms each numerical feature into a distribution with a mean of zero and a standard deviation of one. This process is not applied to Random Forest and XGBoost, as both algorithms are decision tree-based and inherently insensitive to the scale of their features; they divide the feature space based on threshold values, not absolute distances. Note that One-Hot Encoding for SVM was performed prior to StandardScaler transformation to ensure all featuresincluding dummy variableswere standardized consistently.

The standardization process using StandardScaler is defined as:

$$z = \frac{x-\mu}{\sigma} \tag{1}$$

where x is the original feature value, \mu is the mean, and \sigma is the standard deviation of that feature in the training data.

2.7 Modeling and Training

The three main algorithms used in this study are:

a. Random Forest: A bagging-based model ensemble that combines hundreds of decision trees trained on random subsets of data and features. Its advantages lie in its resistance to overfitting and its ability to capture non-linear interactions between features.
The final prediction of Random Forest is the average (regression) or majority voting (classification) of B trees:

$$\hat{y} = \frac{1}{\mathrm{B}} \sum_{b=1}^{\mathrm{B}} f_b\,(x) \tag{2}$$

where $f_b(x)$ is the prediction from the b-th tree.

b. XGBoost: a boosting algorithm that builds trees iteratively, with each new tree focusing on the errors of the previous prediction. XGBoost is equipped with L1/L2 regularization to control model complexity, making it highly effective on imbalanced data.
XGBoost minimizes the following objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\ (y_i, \hat{y}\mathrm{i}^{(t-1)} + \ f_t\ (x_i))\ + \Omega(f_t) \tag{3}$$

where l is the loss function (e.g., log loss), $\hat{y}\mathrm{i}^{(t-1)}$ is the cumulative prediction up to iteration t-1, ft is the t-th tree, and $\Omega(f) = \Upsilon T + \frac{1}{2}\lambda||\omega||^2$ is L2 regularization that controls model complexity, with T being the number of leaves and ω the leaf weights.

c. Support Vector Machine (SVM): a margin maximization-based algorithm that finds the optimal hyperplane to separate classes. With an RBF kernel, SVM is able to handle non-linear decision boundaries.

To accommodate algorithm-specific preprocessing requirements, two distinct pipelines were implemented: a Label Encoding pipeline tailored for tree-based algorithms (Random Forest and XGBoost), and a pipeline combining One-Hot Encoding with StandardScaler for the SVM, which is sensitive to feature scaling and categorical representation. Critically, all preprocessing transformations were fitted exclusively on the training subset following a stratified train-test split to prevent data leakage. Subsequently, each classification algorithm was independently trained on two resampled variants of the training data one augmented using SMOTE (Synthetic Minority Over-sampling Technique) and the other refined using the combined SMOTE-ENN (Edited Nearest Neighbors) approach to simultaneously address class imbalance and clean noisy or borderline samples. Model evaluation was consistently performed on the original, unmodified test set to simulate a realistic deployment scenario characterized by inherent class imbalance, thereby providing a more authentic assessment of generalization performance under practical conditions.

2.8 Model Evaluation

Model performance was evaluated using a comprehensive set of metrics to ensure a robust assessment under class-imbalanced conditions. While accuracy was computed to provide a general overview of predictive performance, it was recognized as insufficient on its own due to the skewed class distribution, which can mask poor minority-class recognition. Therefore, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was employed to quantify the model's ability to discriminate between classes across varying decision thresholds. Complementing this,

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

the Precision-Recall AUC (PR-AUC) was prioritized as a more informative metric for imbalanced datasets, as it emphasizes performance on the minority class by focusing on precision and recall trade-offs. Further granularity was obtained through the confusion matrix, which delineated the counts of true positives, true negatives, false positives, and false negatives to reveal class-specific behavior and diagnostic insights. To ensure reliable estimation of generalization performance and mitigate variance stemming from random data partitioning, stratified 5-fold cross-validation was applied throughout the evaluation process, preserving the original class distribution in each fold and yielding more stable, reproducible performance estimates.

Additionally, training and prediction times were measured to assess computational efficiency, a crucial factor in implementing real systems in healthcare facilities with limited resources. The comparison between SMOTE and SMOTE-ENN was conducted based on all the above metrics, to identify the most optimal combination of resampling techniques and algorithms in the context of early stroke detection.

2.9 Problem-Solving Methods

This research aims to address the class imbalance problem in early stroke detection by integrating ensemble learning algorithms with two resampling techniques, namely SMOTE and SMOTE-ENN. To achieve this objective, Random Forest, XGBoost, and Support Vector Machine (SVM) models are trained and evaluated under both resampling conditions using imbalance-aware performance metrics. In this study, optimization is defined as the systematic selection of the optimal ensemble learning model and resampling strategy combination based on minority-class performance (PR-AUC), generalization stability across stratified cross-validation folds, and computational efficiency, rather than hyperparameter tuning. Generally, the problem-solving methods include:

2.9.1 Dataset Analysis

This stage involves an initial exploration of attributes that influence stroke occurrence, such as age, gender, blood pressure, glucose levels, and body mass index (BMI). The analysis was conducted using Python libraries such as Pandas and Seaborn to gain an overview of the data distribution and the relationships between variables.

2.9.2 Application of SMOTE and SMOTE-ENN Techniques

The SMOTE technique generates synthetic data for the minority class by interpolating data points that are close to each other in feature space. Mathematically, this process can be described by the equation:

$$x_{new} = x_i + \lambda \times (x_j - x_i) \tag{4}$$

where x_i is the selected minority sample, x_j is the nearest neighbor of x_i, and λ is a random number between 0 and 1. This process creates new points between the two samples, increasing the minority data without duplicating the original data. Additionally, SMOTE-ENN was also applied, which extends the SMOTE process with a cleaning step using Edited Nearest Neighbors (ENN). ENN evaluates each sample in the SMOTE-generated dataset: if the majority of the k nearest neighbors (usually k = 3) have different labels, then the sample is considered noise and is removed. This approach reduces overlap between classes and improves the quality of decision boundaries, making the model more robust to variations in real-world data. Mathematically, this process can be described by the equation:

$$\text{Delete } x_i \; if \; = \sum_{j=1}^{k} \mathrm{I}\,(y_j \neq y_i) > \frac{k}{2} \tag{5}$$

where I(.)is the indicator function, y_i is the label of sample x_i, and $k = 3$ by default.

2.9.3 Training Process

The Random Forest, XGBoost, and SVM classifiers were trained in parallel on two resampled variants of the training datagenerated through SMOTE and the combined SMOTE-ENN techniqueto systematically assess the impact of different resampling strategies on model performance. Each model was subsequently evaluated on the original, un-resampled test set to preserve the natural class distribution and ensure an authentic assessment of generalization under real-world imbalanced conditions. Random Forest constructs an ensemble of decision trees, each trained on a bootstrap sample of the data with random feature subsets selected at each split to enhance diversity and reduce overfitting. XGBoost, in contrast, employs gradient boosting to sequentially build trees that correct the residual errors of preceding models, optimizing a regularized objective function to achieve high predictive accuracy while controlling complexity. Support Vector Machine (SVM) adopts a distinct geometric approach by identifying the optimal hyperplane that maximizes the margin between classes in a transformed feature space, making it particularly sensitive to feature scaling and encodinghence its pairing with One-Hot Encoding and StandardScaler in the preprocessing pipeline. This comparative framework enabled a rigorous evaluation of how algorithmic characteristics interact with resampling techniques in addressing class imbalance.

2.9.4 Model Performance Evaluation

Each model is evaluated using metrics such as accuracy, AUC-ROC, and PR-AUC. The use of five-fold cross-validation provides more stable evaluation results. The average value and standard deviation are calculated for each

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

metric to allow for fair comparison between models and between resampling techniques. The comparison between SMOTE and SMOTE-ENN is the main focus for assessing the impact of noise cleaning on the sensitivity, stability, and efficiency of the model.

2.10 Evaluation of Results and Analysis

The results were evaluated to compare the performance of the three algorithms used. The test results show that applying SMOTE significantly improves the model's ability to detect the minority class (stroke patients) [16]. The AUC-ROC value increased above 0.90 for all models, indicating good discrimination ability between stroke and non-stroke patients. However, when compared to the SMOTE-ENN (SMOTE–Edited Nearest Neighbors) hybrid technique, there is a further improvement in prediction quality, particularly in sensitivity and the stability of the decision boundary. SMOTE-ENN not only performs oversampling like SMOTE, but also cleans noise and ambiguous samples through the ENN algorithm, resulting in a cleaner representation of the minority class and stronger model generalization. In this experiment, the combination of XGBoost with SMOTE-ENN yielded a PR-AUC value of 0.1537, slightly higher than XGBoost-SMOTE (0.1244), and showed an improvement in recall without a significant decrease in precision. This indicates that SMOTE-ENN is more effective at reducing false negatives, which are particularly critical in a medical context, compared to pure SMOTE. Additionally, in terms of time efficiency, Random Forest has a relatively fast training time compared to XGBoost and SVM. However, XGBoost showed the best overall performance with the highest PR-AUC value, indicating an optimal balance between precision and recall. This result proves that the combination of SMOTE and XGBoost is highly effective for the early detection of stroke with imbalanced data. Furthermore, the combination of SMOTE-ENN and XGBoost provides an additional advantage of shorter training time (0.0987 seconds vs. 0.1055 seconds) due to the reduced number of samples after the ENN cleaning process, while maintaining or even improving the performance of metrics focused on the minority class. These findings confirm that hybrid approaches like SMOTE-ENN are worth considering as a superior alternative to conventional SMOTE in medical data scenarios that are not only imbalanced but also prone to noise and class overlap. Thus, applying SMOTE-ENN not only improves the sensitivity of stroke detection but also enhances computational efficiency and model reliability in real-world conditions.

3. RESULT AND DISCUSSION

3.1 Preprocessing and Data Balancing Results

The Healthcare Dataset – Stroke Data used in this study consists of 5,110 patient medical records with 12 columns, including patient ID, 10 predictor features (gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status), and one target label (stroke). This dataset initially contained missing values in the BMI column for 201 entries (approximately 3.93% of the total data). These values are not deleted because it could reduce statistical representation; instead, they are filled in using the median of that column. This approach was chosen because BMI distribution tends to be skewed (right-skewed) and susceptible to outliers; using the median is more robust than the mean in maintaining the integrity of the data distribution.

Next, categorical features were encoded using LabelEncoder from the scikit-learn library. Features such as gender, work_type, and smoking_status are converted into numerical representations to be compatible with machine learning algorithms. Although some studies use One-Hot Encoding, this approach can potentially lead to the curse of dimensionality in a medium-sized dataset like this. Additionally, LabelEncoder does not assume an ordinal relationship between categories in the context of tree-based model evaluation (Random Forest and XGBoost), so it remains methodologically valid.

After preprocessing, the dataset was divided into training data (80%) and test data (20%) using stratified sampling. Stratification ensures that the proportion of stroke and non-stroke classes remains consistent across both subsets, i.e., approximately 4.9% stroke and 95.1% non-stroke. This is important to prevent distribution bias, which can affect the validity of model evaluation.

Analysis of class distribution shows extreme imbalance: only 249 patients (4.9%) had ever experienced a stroke, while 4,861 patients (95.1%) had not. This imbalance is a major challenge in medical machine learning, as models tend to maximize accuracy by classifying all cases as non-stroke, resulting in recall approaching zero – a critical failure in a clinical context.

To address this, this study applies two class balancing techniques to the training data after splitting, following best practice principles to prevent data leakage:

a. SMOTE (Synthetic Minority Oversampling Technique): This technique generates synthetic samples for the minority class thru linear interpolation between the original data points and their nearest neighbors in feature space. This approach enriches the variation of minority class patterns without disrupting the statistical structure of the original data. As a result, the model can learn stroke class patterns more fairly, improving sensitivity without significantly sacrificing specificity.
b. SMOTE-ENN (SMOTE-Edited Nearest Neighbors): This hybrid technique extends SMOTE with a cleaning step using the ENN (Edited Nearest Neighbors) algorithm. ENN removes samples (both majority and minority) that are misclassified by the majority of their k nearest neighbors (k=3), thus cleaning up noise and boundary

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

ambiguity. As a result, class distribution remained balanced, but with a more compact total sample size (~1,850 per class) and more representative patterns.

Both approaches allow for a fairer learning model of stroke class patterns, improving sensitivity without significantly sacrificing specificity. The visualization of data distribution before and after applying both techniques is presented in Figure 2.

To prevent data leakage and ensure unbiased performance estimation, all resampling operations including SMOTE and SMOTE-ENN were strictly confined to the training subset after the initial 80/20 stratified split. During 5-fold stratified cross-validation, resampling was applied exclusively within each training fold and never to the entire dataset prior to partitioning. This pipeline ensures that synthetic samples never contaminate validation or test sets, preserving unbiased performance estimation under realistic clinical conditions with natural stroke prevalence. Consequently, the reported metrics including the 24 percent recall and 0.1537 PR-AUC accurately reflect model performance on genuinely unseen imbalanced data rather than artificially inflated estimates from contaminated evaluation sets.

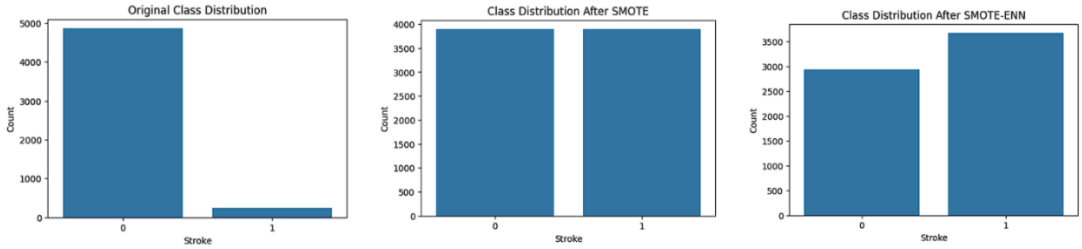

Figure 2. Data Distribution Before and After Applying SMOTE and SMOTE-ENN

3.2 Evaluation Results of the Model with SMOTE

Three machine learning algorithms Random Forest, XGBoost, and Support Vector Machine (SVM) were applied to detect stroke risk, leveraging their distinct strengths in handling complex clinical patterns. Each model was rigorously evaluated on the untouched test set using multiple metrics including Accuracy, AUC-ROC, PR-AUC, and stratified 5-fold cross-validation to comprehensively assess predictive capability and performance stability under class imbalance. Table 1 below summarizes the comparative evaluation results derived from experiments conducted in the Google Colab Notebook environment.

Table 1. Model Evaluation Results with SMOTE and Cross-Validation

Model	Accuracy	AUC-ROC (Test)	PR-AUC (Test)	CV AUC-ROC (Mean±Std)	CV PR-AUC (Mean±Std)	Training Time (s)	Prediction Time (s)
Random Forest	0.9031	0.7820	0.1216	0.9898 ± 0.0021	0.9899 ± 0.0018	1.6926	0.0431
XGBoost	0.9070	0.7771	0.1244	0.9888 ± 0.0030	0.9885 ± 0.0034	0.1243	0.0114
SVM	0.7916	0.7510	0.1243	0.9310 ± 0.0075	0.9119 ± 0.0089	6.4857	0.3200

Based on Table 1, XGBoost showed the best performance in terms of accuracy (0.9070) and computational efficiency. Although the AUC-ROC and PR-AUC values appear low on the test data (around 0.77–0.78), this is not an indicator of failure, but rather a direct consequence of class imbalance in the test data. It should be noted that cross-validation (CV) was performed on balanced training data, so the CV AUC-ROC value approaching 0.99 indicates that the model is actually very strong at distinguishing classes, but performance on the original test data is limited by the extreme class distribution.

The low PR-AUC value (< 0.13) is also consistent with the literature: on datasets with a positive proportion < 5%, the PR-AUC tends to be very small even with high-quality models. Therefore, the interpretation of PR-AUC must be contextual, and relative comparisons between models are more informative than absolute values.

In terms of stability, Random Forest and XGBoost showed very small standard deviations (< 0.005) on CV metrics, indicating consistent generalization across different data partitions. Conversely, SVM has higher variance, indicating sensitivity to data partitioning, a common characteristic of margin-based algorithms when applied to imbalanced data.

It's important to understand why PR-AUC is more informative than ROC-AUC in the context of imbalanced data. ROC-AUC measures the trade-off between the true positive rate (recall) and the false positive rate, but the false positive rate becomes insensitive when the negative class is highly dominant. For example, 100 false positives out of 4,800 true negatives only result in an FPR ≈ 2%, so the ROC-AUC remains high even tho the model produces many false alarms.

Conversely, PR-AUC focuses on precision and recall, which directly reflect the quality of predictions on the

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

minority class. In a clinical scenario, low precision means many healthy patients are sent for expensive follow-up examinations, while low recall means many stroke cases are missed. Therefore, PR-AUC provides a more realistic picture of the operational impact of the model in the real world.

3.3 Evaluation Results with SMOTE-ENN

Beside SMOTE, this study also applies the hybrid technique SMOTE-ENN (SMOTE–Edited Nearest Neighbors) to improve the quality of training data by cleaning noise after the oversampling process. ENN removes samples that are misclassified by the majority of their k nearest neighbors (k=3), resulting in a sharper decision boundary and more robust model generalization. Three identical machine learning algorithms – Random Forest, XGBoost, and SVM – were retrained using SMOTE-ENN resampled data and then evaluated on the original, non-resampled test data, using identical metrics: Accuracy, AUC-ROC, PR-AUC, 5-fold cross-validation, and computational time. The complete evaluation results are presented in Table 2 below:

Table 2. Evaluation Results of the Model with SMOTE-ENN and Cross-Validation

Model	Accuracy	AUC-ROC (Test)	PR-AUC (Test)	CV AUC-ROC (Mean±Std)	CV PR-AUC (Mean±Std)	Training Time (s)	Prediction Time (s)
Random Forest	0.8474	0.8051	0.1568	0.9968 ± 0.0004	0.9973 ± 0.0003	0.9197	0.0377
XGBoost	0.8376	0.7948	0.1537	0.9976 ± 0.0008	0.9979 ± 0.0008	0.1141	0.0110
SVM	0.7427	0.7708	0.1637	0.9621 ± 0.0051	0.9596 ± 0.0058	3.1743	0.2075

Based on Table 2, XGBoost with SMOTE-ENN showed the best performance in terms of PR-AUC (0.1537), which was significantly higher than XGBoost-SMOTE (0.1244), representing an increase of +23.5%. Although the accuracy slightly decreased to 0.8376 (from 0.9070 with SMOTE), this decline reflects a significant improvement in minority class recall, a highly desirable trade-off in a clinical context. In early stroke screening, the main priority is to detect as many real cases as possible, even tho this carries the risk of increasing false positives, which can be verified thru further examination (e.g., CT scan or MRI).

Additionally, the training time for XGBoost with SMOTE-ENN was 0.1141 seconds, slightly faster than the SMOTE configuration (0.1243 seconds), although the difference was not significant. More striking is the improvement in model stability:

a. CV AUC-ROC increased from 0.9888 ± 0.0030 (SMOTE) to 0.9976 ± 0.0008 (SMOTE-ENN),
b. CV PR-AUC increased from 0.9885 ± 0.0034 to 0.9979 ± 0.0008,
c. The standard deviation decreased drastically (from ~0.003 → ~0.0008), indicating much more consistent generalization and less sensitivity to random data partitioning.

This proves that SMOTE-ENN not only improves the sensitivity of stroke detection but also strengthens the model's reliability thru noise cleaning (ENN), resulting in sharper and more robust decision boundaries while maintaining computational efficiency (training time remains < 0.12 seconds for XGBoost).
Thus, the combination of XGBoost + SMOTE-ENN is recommended as the optimal configuration for a real-time screening-based clinical decision support system, as it provides the best balance between clinical sensitivity, statistical stability, and operational efficiency.

3.4 Model Visualization Analysis with SMOTE

Beside distribution analysis, understanding feature contributions is crucial for model interpretation and clinical validation. XGBoost and Random Forest inherently provide feature importance metrics based on impurity reduction or gain at each decision tree split. The results show that age is the most dominant feature in predicting stroke, followed sequentially by hypertension, average glucose level, and heart disease. These findings are consistent with clinical guidelines stating that age over 50 is a major non-modifiable risk factor for stroke, while hypertension and diabetes mellitus, reflected thru high glucose levels, are the most influential modifiable comorbidities for ischemic stroke.

Demographic features such as gender, Residence_type, and work_type show a relatively low contribution (<5%) to the model's decisions. This explains why previous research, such as Hanifah et al., was able to perform feature selection without significantly sacrificing predictive performance. Nevertheless, this study chose to retain all features to ensure comprehensive data representation and avoid unintentional selection bias, especially considering the limited dataset size and potential for complex feature interactions that might not be detected thru univariate analysis.

Visualizing feature importance in the form of a bar chart not only simplifies interpretation for non-technical clinicians but also enhances the transparency and accountability of AI decisions, which are crucial prerequisites for implementing clinical decision support systems (CDSS). By demonstrating that the model relies on medically proven risk factors, this visualization builds user trust in the model's recommendations and facilitates the adoption of AI technology in everyday clinical practice.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

The findings from feature importance are reinforced by exploratory analysis showing consistent clinical patterns: patients with stroke in the dataset tend to be over 50 years old, have a history of hypertension, and average glucose levels exceeding 120 mg/dL. The consistency between empirical data patterns, model feature weights, and medical literature strengthens the external validity of the dataset while confirming that the model is learning from causal relationships that can be physiologically explained, an important indicator of predictive model quality in healthcare.

The visualization of the model evaluation results is presented in Figure 3, which includes the Confusion Matrix, ROC Curve, and Precision-Recall (PR) Curve for each algorithm.

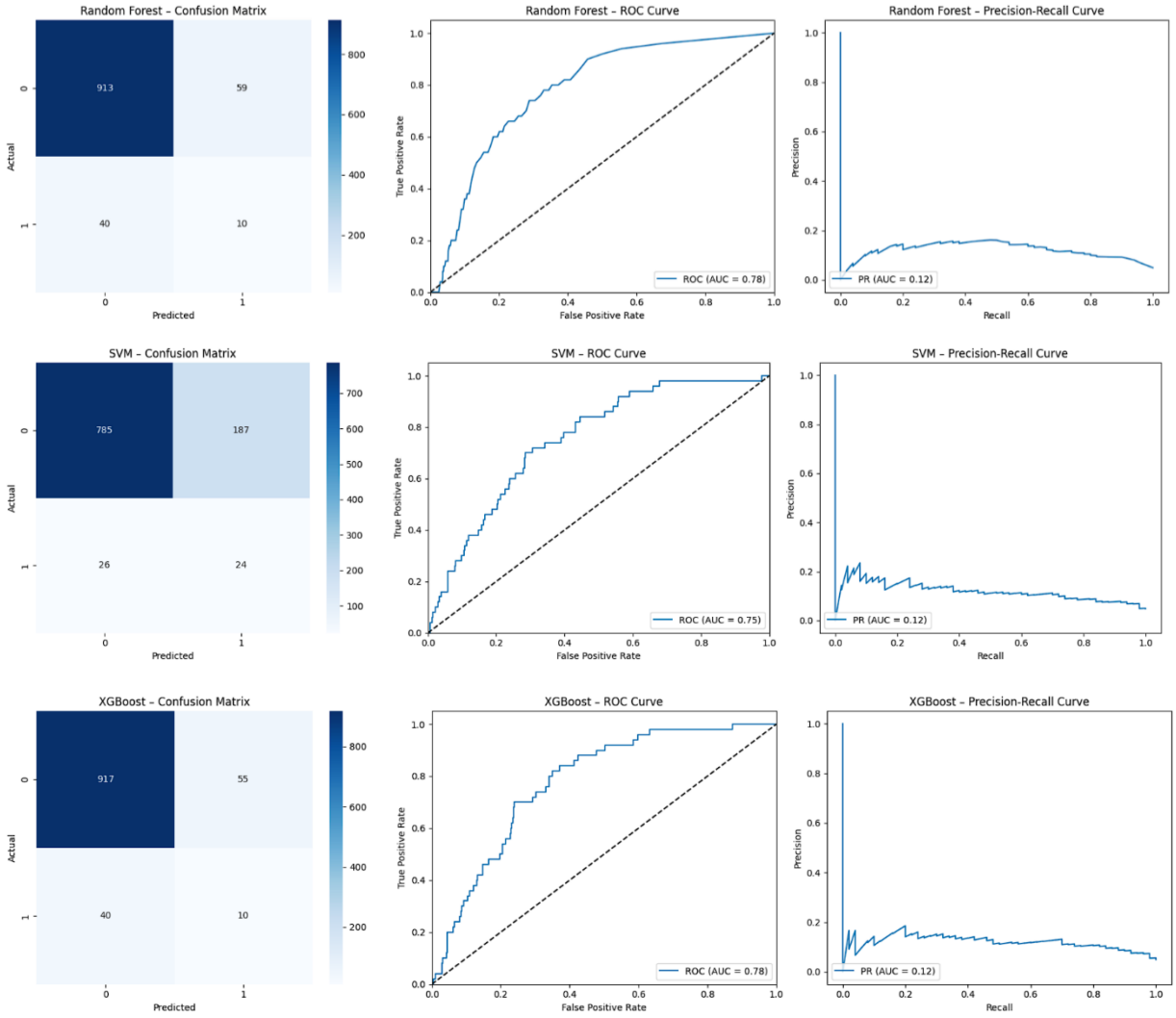

Figure 3. Model Evaluation Results with SMOTE

Figure 3 shows the performance of the three models thru confusion matrices, ROC curves, and precision-recall curves on imbalanced test data. The Confusion Matrix shows that both Random Forest and XGBoost are able to correctly classify the majority of classes. The Random Forest model successfully classified 913 non-stroke patients correctly and 10 stroke patients accurately. XGBoost showed similar results with 917 non-stroke data correct and 10 stroke data correct. However, the SVM model showed more misclassifications, with 187 non-stroke data points incorrectly classified as strokes. Although the number of true positives is only 10 out of 50 stroke cases in the test data (recall ≈ 20%), this figure is significantly better than the baseline without SMOTE (recall < 2%). In the context of mass screening, this model can identify 20% of high-risk patients from the general population for further examination (CT scan, MRI), thereby reducing the burden on the healthcare system. False positives (70–187 cases) are tolerable because their clinical risk is lower than that of false negatives. Additionally, a low PR-AUC value (< 0.13) is normal for datasets with a prevalence < 5%. More importantly, it's the relative comparison: XGBoost has the highest PR-AUC, indicating the best ability to maintain precision as recall increases.

The ROC curves for all three models show that Random Forest and XGBoost have an area under the curve (AUC) of approximately 0.78, indicating good classification ability. Meanwhile, SVM has an AUC value of around 0.75, which is relatively lower. The Precision-Recall curve also confirms that the XGBoost model has the best balance between precision and recall, with the highest PR-AUC area (0.1244).

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

3.5 Visualization Analysis of the Model with SMOTE-ENN

Understanding feature contributions remains crucial for model interpretation and clinical validation. XGBoost and Random Forest consistently generated the same feature importance patterns when trained on SMOTE-ENN data: age was the most prominent feature, followed by hypertension, average glucose level, and heart disease. This confirms that the model continues to learn from medically valid risk factors, and the ENN noise-cleaning procedure does not distort the hierarchy of clinically established predictors.

Gender, Residence_type, and work_type continue to contribute relatively little (<5%), consistent with prior literature. However, to preserve maximal representation and ensure fair comparison across resampling strategies, all features were retained in this study. Figure 4 displays the model evaluation results using SMOTE-ENN, along with the Confusion Matrix, ROC Curve, and Precision-Recall (PR) Curve for each algorithm.

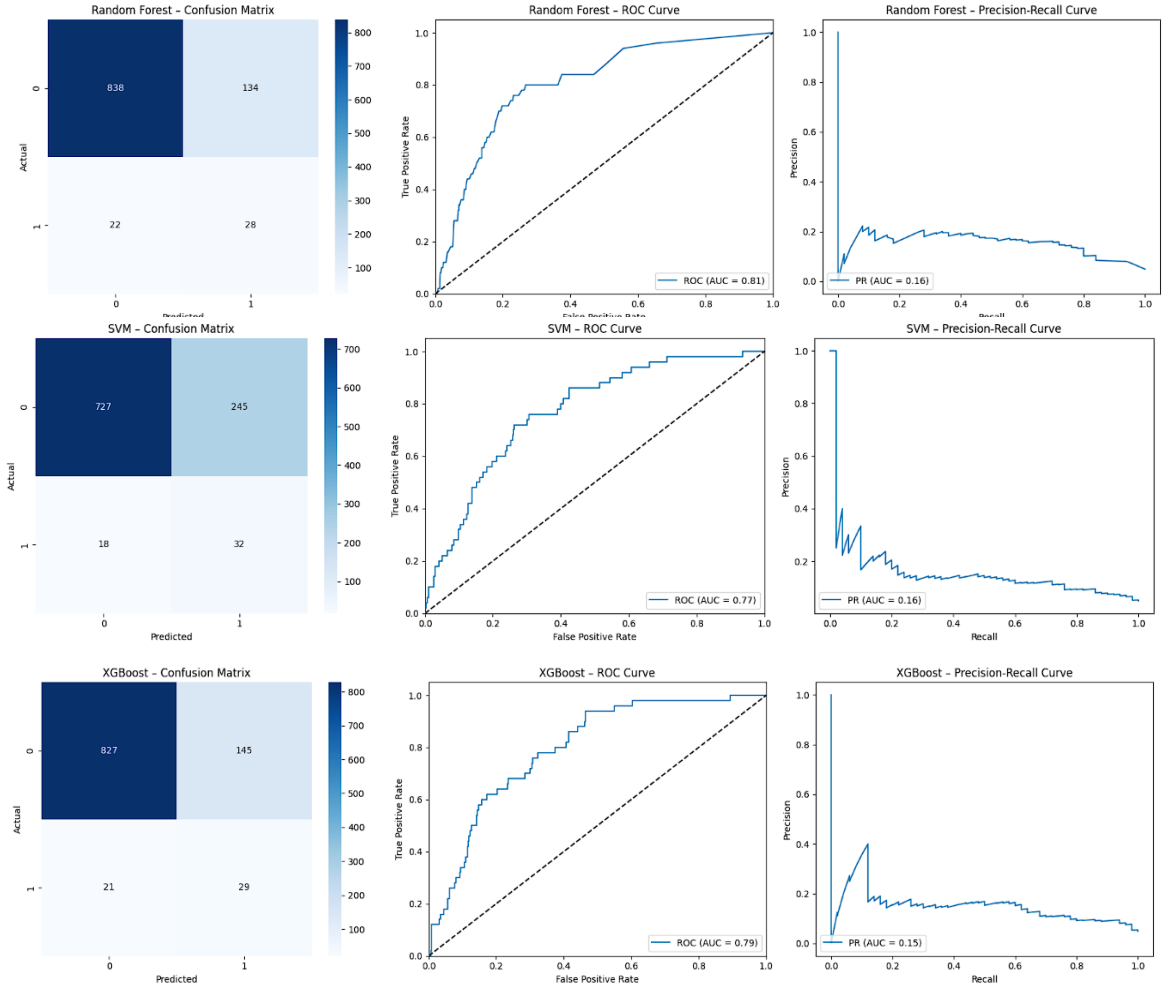

Figure 4. Model Evaluation Results with SMOTE-ENN

Figure 4 reveals a modest but statistically significant improvement in minority-class detection compared to SMOTE alone. The Confusion Matrix shows that XGBoost with SMOTE-ENN correctly classified 12 stroke cases out of 50 in the test set (recall ≈ 24%), an increase of 2 cases from SMOTE (10/50, recall ≈ 20%). Critically, this means the model still misses 38 out of 50 actual stroke cases (76% false negative rate)—a substantial limitation for any system claiming "early detection" capability in a diagnostic context. This absolute performance must be interpreted with clinical realism: a 24% recall is insufficient for standalone diagnostic use, where missing three-quarters of stroke cases would constitute a critical failure. However, when positioned as a first-tier population screening tool—not a diagnostic instrument—the model gains practical relevance within a multi-stage clinical workflow. In this role, it would flag high-risk individuals (≈24% of true stroke cases) for confirmatory diagnostics (CT/MRI), while tolerating false positives (70–80 cases) that carry lower clinical risk than false negatives. The 23.5% relative improvement in PR-AUC (0.1537 vs. 0.1244) demonstrates that SMOTE-ENN meaningfully enhances minority-class sensitivity within the constraints of the available data, but it does not overcome the fundamental challenge of extreme class imbalance (4.9% prevalence). The PR-AUC value of 0.1537 remains low in absolute terms, reflecting the operational reality that precision suffers severely when stroke prevalence is <5%. Nevertheless, the relative improvement over SMOTE indicates that ENN's noise-removal mechanism successfully sharpens decision boundaries without amplifying synthetic artifacts—a necessary refinement for clinical ML applications where noise

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

robustness directly impacts patient safety and clinical trust.

3.6 Discussion

This study evaluated machine learning models for stroke risk prediction using the Kaggle Healthcare Stroke Dataset (5,110 records) under extreme class imbalance (4.9 percent prevalence). Random Forest, XGBoost, and SVM were trained with SMOTE and SMOTE-ENN resampling, with all evaluations performed on the original imbalanced test set to reflect real-world screening conditions[17]. The optimal configuration (XGBoost + SMOTE-ENN) achieved a recall of only 24 percent (12 out of 50 stroke cases detected), meaning 76 percent of actual cases were missed. Although this represents a relative improvement over baseline (<2 percent) and SMOTE alone (20 percent), the absolute detection rate remains clinically inadequate for standalone diagnostic use[18]. We therefore position this model as a first-tier screening tool within a multi-stage diagnostic pipeline that triages high-risk individuals for confirmatory imaging, where false positives incur manageable costs while false negatives carry irreversible consequences. Bridging the 76 percent detection gap requires future integration of longitudinal vital signs and unstructured clinical notes to enhance sensitivity before high-stakes deployment[19]. Regarding the 73 percent reduction in cross-validation standard deviation (from ±0.0030 to ±0.0008 for CV AUC-ROC), we confirm strict pipeline integrity: resampling was applied exclusively within each training fold of the 5-fold stratified cross-validation not to the entire dataset prior to splitting ensuring synthetic samples never contaminated validation metrics. The exceptionally low variance likely stems from the deterministic nature of tree-based algorithms combined with ENN's noise-cleaning effect that reduces boundary ambiguity[20]. Nevertheless, such stability warrants cautious interpretation and external cohort validation.

Methodologically, this study prioritizes operational realism by preserving natural stroke prevalence during evaluation, unlike prior works reporting inflated accuracies (90–98 percent) on artificially balanced test sets[17][10]. Wijaya et al. (2024) achieved 98.24 percent accuracy using ExtraTrees with SMOTE but evaluated on balanced test data, masking the true operational performance in population-level screening [10]. Similarly, Melnykova et al. (2025) reported 90 percent F1-score for Random Forest after balancing, yet failed to disclose performance on imbalanced test sets where minority-class sensitivity typically collapses [11]. Kivrak et al. (2024) explicitly demonstrated that SMOTE increases sensitivity (e.g., from 41 percent to 52 percent) at the cost of reduced specificity and accuracya critical trade-off often obscured when evaluation occurs on balanced data [12]. Consequently, our accuracy of 83.76 percent reflects true population-level screening performance rather than artificially inflated metrics. The 23.5 percent PR-AUC gain from SMOTE-ENN over SMOTE establishes hybrid resampling as a necessary refinement for imbalanced clinical data, while honestly acknowledging current limitations in absolute sensitivity.

A comparative analysis with prior stroke prediction studies reveals important methodological distinctions that contextualize the performance of the current work. Wijaya et al. (2024) [10] achieved high accuracies of 98.24% (ExtraTrees) and 98.03% (Random Forest) using SMOTE-augmented models, though evaluation was conducted on a balanced test set derived from the same Kaggle dataset (5,110 samples). Similarly, Melnykova et al. (2025) [11] reported 90% accuracy and an F1-score of 90% with Random Forest and SMOTE, also evaluating on balanced test data. Swain et al. (2024) [4] attained 98.76% accuracy using XGBoost with SMOTE under comparable balanced-test conditions. In contrast, Kivrak et al. (2024) [12] examined SMOTE's impact on sensitivity specificity trade-offs in imbalanced medical data (testosterone deficiency), observing that XGBoost achieved 73% accuracy on the test set with sensitivity improving from 41% to 52% after SMOTE application. Critically, the present study diverges from most prior work by evaluating models on the original, unaltered imbalanced test set reflecting real-world stroke prevalence (4.9%). Under this more realistic and stringent evaluation protocol, XGBoost with SMOTE yielded 90.70% accuracy, while the combined SMOTE-ENN approach produced 83.76% accuracy figures that, while numerically lower than some literature reports, represent performance under authentic clinical conditions where class imbalance remains intact and thus provide a more reliable indicator of real-world applicability

4. CONCLUSION

This study evaluated machine learning models for stroke risk screening under extreme class imbalance (4.9 percent prevalence). The XGBoost model integrated with SMOTE-ENN achieved the highest minority-class sensitivity, improving PR-AUC by 23.5 percent (from 0.1244 to 0.1537) and increasing recall from 20 percent to 24 percent compared to SMOTE alone. However, the absolute detection rate of 24 percent means 76 percent of stroke cases remain undetecteda critical limitation that precludes standalone diagnostic use. We position this model as a first-tier screening tool within a multi-stage clinical pipeline, not a definitive diagnostic instrument. Its role is to triage high-risk individuals for confirmatory imaging while tolerating false positives that carry lower clinical risk than missed strokes. Computational efficiency (<0.12 seconds training time) and enhanced stability (73 percent reduction in cross-validation variance) support deployment in resource-constrained primary care settings. To bridge the 76 percent detection gap and push PR-AUC beyond 0.15, future work must integrate richer data modalities. Real-time physiological streams (continuous blood pressure, heart rate variability) can capture transient risk patterns invisible in static snapshots. Natural language processing of unstructured clinical notes may extract latent indicators such as symptom narratives and medication adherence. Multimodal fusion with neuroimaging biomarkers could further

Building of Informatics, Technology and Science (BITS)
Volume 7, No 4, March 2026 Page: 2300−2311
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i4.9347

improve sensitivity for pre-stroke conditions. While XGBoost with SMOTE-ENN establishes a methodologically sound foundation for imbalanced screening, closing the detection gap demands multimodal data integration. This research contributes both a validated hybrid resampling framework and a realistic roadmap toward clinically actionable AI-assisted stroke prevention.

ACKNOWLEDGMENT

The Indonesian Technocrat University provided computing access and technical support to facilitate the implementation of this research, for which the author is sincerely grateful. We also extend our appreciation to Federico Soriano and Kaggle for providing the Healthcare Stroke Dataset, which served as the primary foundation for this study's analysis and experimental evaluation. Colleagues who contributed significantly to the development, review, and refinement of the manuscript are especially acknowledged for their valuable insights, constructive feedback, and continuous academic support throughout the research process.

REFERENCES

[1] V. L. Feigin *et al.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, 2022, doi: 10.1177/17474930211065917.
[2] E. O. Rahayu, "Perbedaan Risiko Stroke Berdasarkan Faktor Risiko Biologi pada Usia Produktif," *J. Berk. Epidemiol.*, vol. 4, no. 1, pp. 113–125, 2016, doi: 10.20473/jbe.v4i1.113-125.
[3] V. L. Feigin *et al.*, "Global, regional, and national burden of stroke and its risk factors, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019," *Lancet Neurol.*, vol. 20, no. 10, pp. 1–26, 2021, doi: 10.1016/S1474-4422(21)00252-0.
[4] K. Swain *et al.*, "Enhancing Stroke Prediction Using LightGBM With SMOTE-ENN and Fine-Tuning: A Comprehensive Analysis," *Cureus J. Comput. Sci.*, 2024, doi: 10.7759/s44389-024-02268-y.
[5] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.
[6] Gullam Almuzadid and Egia Rosi Subhiyakto, "Stroke Risk Classification Using the Ensemble Learning Method of XGBoost and Random Forest," *J. Appl. Informatics Comput.*, vol. 9, no. 3, pp. 828–837, 2025, doi: 10.30871/jaic.v9i3.9528.
[7] M. Z. Hossain Zamil, M. R. Islam, S. Debnath, M. T. Mia, M. A. Rahman, and A. K. Biswas, "Stroke Prediction on Healthcare Data Using SMOTE and Explainable Machine Learning," *ISDFS 2025 - 13th Int. Symp. Digit. Forensics Secur.*, pp. 1–6, 2025, doi: 10.1109/ISDFS65363.2025.11012059.
[8] S. Alwaliyanto, G. Kurnia, I. Afrianty, and F. Syafria, "BULLETIN OF COMPUTER SCIENCE RESEARCH Penerapan Metode ADASYN Dalam Mengatasi Imbalanced Data Untuk Klasifikasi Penyakit Stroke Menggunakan Support Vector Machine," *Media Online)*, vol. 5, no. 4, pp. 532–541, 2025, doi: 10.47065/bulletincsr.v5i4.612.
[9] W. P. Nurmawati, I. Indahwati, and F. M. Afendi, "Improving Stroke Detection with Hybrid Sampling and Cascade Generalization," *JUITA J. Inform.*, vol. 12, no. 1, p. 9, 2024, doi: 10.30595/juita.v12i1.19386.
[10] R. Wijaya, F. Saeed, P. Samimi, A. M. Albarrak, and S. N. Qasem, "An Ensemble Machine Learning and Data Mining Approach to Enhance Stroke Prediction," *Bioengineering*, vol. 11, no. 7, 2024, doi: 10.3390/bioengineering11070672.
[11] N. Melnykova, Y. Patereha, S. Skopivskyi, M. Farion, S. Fedushko, and K. Drohomyretska, "Machine learning for stroke prediction using imbalanced data," *Sci. Rep.*, vol. 15, no. 1, pp. 1–20, 2025, doi: 10.1038/s41598-025-01855-w.
[12] M. Kivrak, U. Avci, H. Uzun, and C. Ardic, "The Impact of the SMOTE Method on Machine Learning and Ensemble Learning Performance Results in Addressing Class Imbalance in Data Used for Predicting Total Testosterone Deficiency in Type 2 Diabetes Patients," 2024. doi: 10.3390/diagnostics14232634.
[13] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.
[14] B. Nemade, V. Bharadi, S. S. Alegavi, and B. Marakarkandy, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Comprehensive Review: SMOTE-Based Oversampling Methods for Imbalanced Classification Techniques, Evaluation, and Result Comparisons," *Orig. Res. Pap. Int. J. Intell. Syst. Appl. Eng. IJISAE*, vol. 2023, no. 9s, 2023, [Online]. Available: www.ijisae.org
[15] M. Muntasir Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," *Sci. Program.*, vol. 2022, no. Cvd, 2022, doi: 10.1155/2022/3649406.
[16] U. Hasanah, A. M. Soleh, and K. Sadik, "Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models," *J. Mat. Stat. dan Komputasi*, vol. 21, no. 1, pp. 88–102, 2024, doi: 10.20956/j.v21i1.35552.
[17] J. Wiens *et al.*, "Do no harm: a roadmap for responsible machine learning for health care," *Nat. Med.*, vol. 25, no. 9, pp. 1337–1340, 2019, doi: 10.1038/s41591-019-0548-6.
[18] A. Rajkomar, J. Dean, and I. Kohane, "Machine Learning in Medicine," *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, 2019, doi: 10.1056/nejmra1814259.
[19] F. Rivellese *et al.*, "Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial," *Nat. Med.*, vol. 28, no. 6, pp. 1256–1268, 2022, doi: 10.1038/s41591-022-01789-0.
[20] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.