ISSN: 2473-4901
v11 n6330

Improving AI-Driven Stroke Prediction Models: A Comparative Evaluation of SMOTE and Undersampling Methods

Dara Tourt
dara.tourt@my.metrostate.edu

Queen Booker
queen.booker@metrostate.edu

Simon Jin
simon.jin@metrostate.edu

College of Business and Management
Metropolitan State University
Minneapolis, Minnesota

Abstract

Artificial intelligence (AI) is improving the field of predictive healthcare by enabling data-driven decision-making through advanced machine learning (ML) algorithms. Stroke prediction is challenging due to highly imbalanced clinical datasets, where positive cases are rare. This study investigates the impact of data-level resampling methods on the performance of AI-driven predictive models. Four widely used classifiers—Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gradient Boosting (GB)—were applied to a highly imbalanced stroke dataset. Models were evaluated across key AI performance metrics. Paired t-tests assessed the statistical significance of observed differences. This comparative analysis offers critical insights into how data balancing techniques impact the reliability of AI models. The findings support the development of more effective and ethically responsible AI systems for early stroke detection.

Keywords: Stroke Prediction, Class Imbalance, SMOTE, Undersampling, Machine Learning

ISSN: 2473-4901
v11 n6330

Improving AI-Driven Stroke Prediction Models: A Comparative Evaluation of SMOTE and Undersampling Methods

Dara Tourt, Queen Booker, and Simon Jin

1. INTRODUCTION

Artificial intelligence (AI) is increasingly important in healthcare, supporting diagnosis, prediction, and management of complex conditions. Stroke prediction is a particularly high-impact application given the sudden onset and severe consequences of stroke. Machine learning (ML) models show promise for identifying risk by analyzing large-scale electronic health records, but their effectiveness is often limited by classification imbalance: stroke-positive cases represent only a small fraction of the data, making accurate detection difficult. In such scenarios, models tend to favor the majority class and overlook minority cases, raising the risk of false negatives—an unacceptable outcome in clinical settings where early detection is critical.

To address this, researchers apply resampling techniques that adjust class distribution in training data. Oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) create synthetic minority samples, while undersampling approaches (e.g., Tomek Links, Edited Nearest Neighbors, and NearMiss) reduce majority samples to balance the data. Despite their widespread use, few studies provide systematic, side-by-side comparisons of these methods across multiple ML classifiers in stroke prediction.

This study fills that gap by evaluating the impact of SMOTE and several undersampling techniques on four common classifiers: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting (GB). We assess performance using multiple evaluation metrics and statistical testing to identify trade-offs and practical implications.
Our research is guided by two questions:

- **RQ1:** How do SMOTE and selected undersampling methods compare in improving model performance for stroke prediction with highly imbalanced datasets?
- **RQ2:** What trade-offs arise between predictive performance when using SMOTE versus undersampling techniques?

By addressing these questions, this study contributes to building more accurate and clinically relevant AI models for early stroke detection.

2. LITERATURE REVIEW

This section reviews key literature on stroke prediction by focusing on classification imbalance, oversampling methods such as SMOTE, undersampling techniques, and the use of machine learning models. It concludes by identifying current research gaps that this study aims to address.

Class Imbalance in Stroke Prediction

Class imbalance is a well-documented challenge in healthcare datasets, where stroke-positive cases are far fewer than non-stroke cases. This imbalance biases models toward majority classifications, leading to poor sensitivity in detecting actual stroke cases and an elevated risk of false negatives (Salmi et al., 2024; Chen et al., 2024; Lin et al., 2024). In clinical contexts, missed diagnoses have serious consequences, underscoring the importance of addressing imbalance in predictive modeling (Aish et al., 2024).

Over-sampling Techniques: SMOTE

Over-sampling increases the representation of minority cases in training data. The Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely adopted methods, generating synthetic samples by interpolating between existing minority cases (Chawla et al., 2002). Studies show that SMOTE improves sensitivity and F1-scores in medical predictions (Salmi et al., 2024), though it can also create overlapping regions or introduce noise, increasing overfitting risk (Elreedy et al., 2024; Fernández et al., 2018a, 2018b).

Undersampling Techniques

Undersampling reduces imbalance by removing majority-class cases. While effective in improving minority detection, it risks discarding valuable information in smaller datasets.

- **Random Undersampling (RU):** Efficient and often improves recall but

ISSN: 2473-4901
v11 n6330

may remove informative samples (He & Garcia, 2009).

- **Tomek Links:** Identifies neighboring pairs from different classes and removes majority instances to clean decision boundaries. It can enhance separability but may discard useful borderline cases (Tomek, 1976; Batista et al., 2004).
- **Edited Nearest Neighbors (ENN):** Removes samples that disagree with most neighbors, reducing noise but sometimes overly eliminating data and increasing computational cost (Wilson, 1972; Laurikkala, 2001).
- **NearMiss:** Retains majority cases based on distance to minority samples, emphasizing boundary representation (Mani & Zhang, 2003). Variants differ in focus: NearMiss-1 enhances sensitivity but raises false positives; NearMiss-2 reduces overlap but may miss borderline patterns (Yen & Lee, 2009); NearMiss-3 sharpens boundaries but can preserve noisy examples.
-

Machine Learning and Sampling Methods

Machine learning classifiers such as Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB) are widely used in healthcare prediction. Logistic Regression remains a strong baseline due to its interpretability, particularly when imbalance is corrected through resampling (Sáez et al., 2015). Ensemble models like RF and GB capture nonlinear patterns and perform well with SMOTE or hybrid approaches (Chawla et al., 2002; Fernández et al., 2018). XGB, leveraging gradient boosting, demonstrates strong predictive power and is enhanced by resampling techniques (Chen & Guestrin, 2016; Haixiang et al., 2017). Combining these classifiers with oversampling or undersampling consistently improves minority-class detection in medical datasets (Douzas & Bacao, 2018).

3. METHODOLOGY

This study follows the research methodology used by Kamiri and Mariga (2021), which includes Data Collection, Data Pre-processing, Model Training, Model Testing, and Model Evaluation. A process model outlining the methodology's steps is shown in Appendix F.

We aim to look at how different resampling techniques can help improve stroke prediction when the data is heavily imbalanced. We compare SMOTE, a popular oversampling method, with several undersampling approaches like Random Undersampling, Tomek Links, Edited Nearest Neighbors (ENN), and the three versions of NearMiss. To see how these methods impact results, we test them across four commonly used machine learning models: Logistic Regression, Random Forest, XGBoost, and Gradient Boosting, all using a real-world stroke dataset where positive cases are rare.

Gaps in Existing Research

While SMOTE and undersampling techniques have been studied independently, there is limited research that compares a broad range of these methods within the context of stroke prediction across multiple classifiers. Many existing studies also lack statistical validation of their findings, which limits the reproducibility and reliability of their conclusions. Furthermore, few works evaluate performance using a comprehensive set of metrics, such as precision, recall, F1-score, ROC-AUC, and PR-AUC, which are essential for assessing models trained on imbalanced data.

This study addresses these limitations by performing a side-by-side comparison of several resampling techniques across four classifiers, applying consistent evaluation criteria and statistical significance testing. The findings offer practical insights into how different resampling methods affect stroke prediction performance and provide guidance for building more reliable machine learning models in healthcare.

Feature	Feature description
id	Unique identifier
gender	Male, Female, or Other
age	Age of the patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	No or Yes
work_type	Children, Govt_job, Never_worked, Private, or Self-employed
Residence_type	Rural or Urban
avg_glucose_level	Average glucose level in the blood
bmi	Body mass index
smoking_status	Formerly smoked, Never smoked, Smokes, or Unknown
stroke	1 if the patient had a stroke, 0 if not

Table 1: Description of Stroke Dataset

ISSN: 2473-4901
v11 n6330

The following section provides a detailed explanation of the research methodology adopted in this study.

Dataset

The dataset used in this study contains 5,110 instances, each representing a patient, and is publicly available on Kaggle (Fedesoriano, n.d.). The dataset includes a variety of attributes relevant to predicting the occurrence of a stroke, with detailed descriptions provided in Table 1 below.

Among the patients, 2,994 were female, 2,115 were male, and 1 was categorized as other. The average age was 43 years, with a range from 18 to 82 years. Additionally, 498 patients had hypertension, and 276 were diagnosed with heart disease. The variable smoking_status represents the patient's self-reported smoking behavior. It is a categorical feature with four possible values:

- formerly smoked – the individual has smoked in the past but is no longer a smoker.
- never smoked – the individual has never smoked.
- smokes – the individual is a current smoker.
- Unknown – the smoking history of the individual is not recorded (i.e., missing or unavailable information).

The variable work_type describes a patient's type of employment or occupational status. It is a categorical variable with five distinct values:

- Private – employed in the private sector.
- Self-employed – working independently or running their own business.
- Govt_job – employed in government service.
- Never_worked – individuals who have never been employed.
- Children – individuals classified as a dependent and had not entered into the workforce.

The dataset exhibits a significant classification imbalance, with the majority of cases being non-stroke. Specifically, 95.1% (4,861 cases) are non-stroke, while 4.9% (249) represent stroke cases, as shown in Figure 1. This imbalance mirrors real-world scenarios, where stroke events are less frequent but have significant clinical implications.

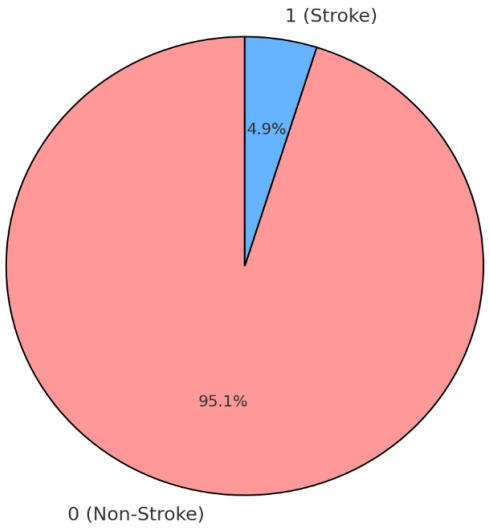


Figure 1: Classification Distribution in Stroke Dataset

Data Preprocessing

To prepare the dataset for model development, several preprocessing steps were implemented. Missing BMI values, which accounted for approximately 4% of the data, were imputed using the mean to maintain consistency. Irrelevant features, such as patient ID, were excluded, while categorical features (e.g., gender, ever_married, work_type, Residence_type, and smoking_status) were label-encoded into binary values for compatibility with the model.

Numerical features were standardized to address discrepancies in magnitude and units, ensuring fair evaluation and preventing data leakage. For example, average glucose levels are in the hundreds, while BMI values are typically in the tens.

Selected Features

Significant independent variables and independent variables deemed significant to stroke prediction but determined not highly correlated to stroke were included in our analysis. Based on the correlation between the target variable (stroke) and the independent variables, as shown in Figure 2 and Figure 4 in Appendix D, we excluded variables such as *work_type*, *Residence_type*, and *gender* from model development because their correlations with the target variable were negligible ($|r| \leq 0.03$; Cohen, 1988) and were not determined to not be significant to the model.

ISSN: 2473-4901
v11 n6330

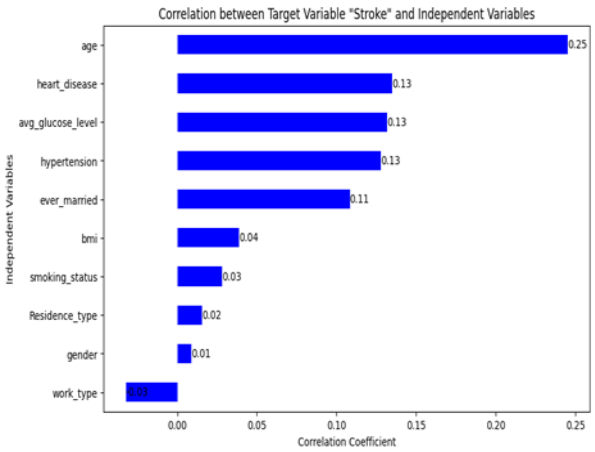


Figure 2: Correlation between Target Variable (Stroke) and Independent Variables

Although the simple correlations between BMI, smoking status, and stroke were negligible in our dataset, we retained these variables in model development for both theoretical and methodological reasons. First, BMI and smoking are widely recognized in the epidemiological literature as important risk factors for stroke (Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration, 2014; Pan et al., 2019), and excluding them could undermine the clinical relevance of our findings. Second, correlation with the outcome alone does not capture the potential contribution of these variables in a multivariate framework, where nonlinear associations or interactions with other predictors may enhance predictive performance (Molnar, 2022). Finally, including BMI and smoking status supports comparability with prior stroke prediction studies, ensuring that our results can be interpreted within the broader body of research.

Model Selection and Development

To evaluate the impact of different resampling strategies on stroke prediction, we implemented a pipeline-based approach that combined resampling, standardization, and classification. Four machine learning classifiers were selected based on their widespread use and demonstrated effectiveness in binary classification tasks. These included Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting Classifier (GB). Together, these models represent a balanced mix of linear and ensemble-based learning algorithms commonly used in healthcare data analysis.

The dataset was split into 70 percent for training and 30 percent for testing. All resampling techniques and machine learning models were implemented using their default settings, with the random state parameter set to 42 to ensure reproducibility. To ensure reliable performance estimates, we used 5-fold stratified cross-validation during model training. In this approach, the dataset was divided into five equal parts (folds) while preserving the proportion of stroke and non-stroke cases in each fold. For each iteration, four folds were used to train the model and the remaining fold was used for testing. This process was repeated five times, with each fold serving once as the test set. The results from all five iterations were then averaged to produce a more stable and generalizable estimate of model performance.

Cross-validation is particularly important in imbalanced datasets such as stroke prediction, because it prevents performance results from being overly influenced by a single train–test split. Without this procedure, the distribution of minority cases (stroke events) in the test set could vary widely, leading to poor sensitivity and a higher risk of false negatives. By averaging across folds, cross-validation provides a more accurate picture of how often the model is likely to produce false negatives (missed stroke cases, which carry high clinical cost) and false positives (incorrectly flagged non-stroke cases, which increase system burden and unnecessary interventions). This ensures that the evaluation reflects not only statistical performance but also the potential clinical and operational costs associated with deploying the models in practice

Each classifier was evaluated using a range of data balancing techniques. These included one oversampling method, Synthetic Minority Over-sampling Technique (SMOTE), and several undersampling methods, such as Random Undersampling, Tomek Links, Edited Nearest Neighbors (ENN), and the three NearMiss variants (NearMiss-1, NearMiss-2, and NearMiss-3). Model performance was also evaluated using the original imbalanced dataset, referred to as the "None" configuration, to serve as a baseline.

For every combination of classifier and resampling method, a machine learning pipeline was constructed. The pipeline began with a resampling step (where applicable), followed by feature standardization using the StandardScaler, and concluded with the selected classifier. To ensure a fair and robust assessment of model performance, five-fold stratified cross-validation was employed. This approach maintained the original classification distribution in each fold, which is particularly important when dealing with

ISSN: 2473-4901
v11 n6330

imbalanced datasets.

Model performance was evaluated using six metrics: accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. These metrics provided a well-rounded evaluation framework, especially suitable for assessing models trained on imbalanced data. To avoid data leakage, only the test scores from each fold were collected during cross-validation.

After cross-validation, each model pipeline was trained on the full training set and then evaluated on the held-out test set. A confusion matrix was generated for each configuration to capture the counts of true positives, false positives, false negatives, and true negatives. These results were recorded for detailed error analysis.

Finally, predictions from each model-resampling configuration were saved for further statistical analysis and visualizations. Confusion matrices were also plotted to offer a visual understanding of how well each model performed under different resampling scenarios.

This experimental setup provided a consistent and reproducible framework for evaluating how various resampling methods influenced the classification performance of different machine learning models.

Evaluation Metrics

To assess model performance under different resampling strategies, several well-established evaluation metrics were used. These metrics help capture both overall accuracy and the model's ability to correctly identify minority classification instances in an imbalanced dataset (He & Garcia, 2009).

Accuracy

Accuracy measures the proportion of correctly classified instances. Although commonly used, it can be misleading in imbalanced datasets since a model may achieve high accuracy by always predicting the majority classification (Jeni et al., 2013).

Precision (Positive Predictive Value)

Precision calculates the proportion of true stroke cases among all cases predicted as stroke. High precision reflects fewer false positives, which is essential to avoid unnecessary medical interventions (Sokolova & Lapalme, 2009).

Recall (Sensitivity or True Positive Rate)

Also known as sensitivity, recall measures the ability to correctly identify all actual stroke cases. It is critical in healthcare to minimize false negatives, which may result in missed diagnoses (Davis & Goadrich, 2006).

A low recall score indicates that many stroke patients are incorrectly classified as non-stroke, increasing the risk of undiagnosed cases.

F1-Score

The F1-score is the harmonic mean of precision and recall, offering a balanced view when both false positives and false negatives matter. It is particularly useful in imbalanced classification settings (Fernández et al., 2018a).

A high F1-score means that the model effectively balances precision and recall, making it a more informative metric for evaluating stroke prediction performance.

Confusion Matrix

The confusion matrix summarizes predictions into four categories: true positives, true negatives, false positives, and false negatives. It provides a clear view of how the model performs on each classification (Tharwat, 2020; Swaminathan & Tantri, 2024), especially in minimizing false negatives.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) indicates how well the model distinguishes between stroke and non-stroke cases across various threshold settings. It evaluates the trade-off between true positive rate and false positive rate and is widely used in binary classification tasks (Fawcett, 2006; Choi et al., 2024).

Area Under the Precision-Recall Curve (PR-AUC)

The Area Under the Precision-Recall Curve (PR-AUC) focuses on the model's ability to correctly identify stroke cases among all positive predictions. It is particularly useful when dealing with highly imbalanced datasets, where the number of actual positive cases is small (Saito & Rehmsmeier, 2015; Sofaer et al., 2019).

Statistical Significance: t-test

To determine whether the differences in model performance was statistically significant, a paired t-test was conducted to confirm or reject our hypotheses. The t-test evaluates whether the observed performance variations are attributable to the resampling techniques or occur by chance (Demšar, 2006).

ISSN: 2473-4901
v11 n6330

After obtaining the prediction results from each model across different sampling methods, we manually performed a paired t-test in Excel, using the two-sample test assuming equal variances.

The statistical significance level was set at a p-value of less than 0.05, indicating that differences between models are considered statistically significant if the p-value falls below this threshold.

The next section details the results and discussion of this study.

4. RESULTS AND DISCUSSION

This section presents the results of the model evaluations using several key performance metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. The findings, summarized in Table 2 in Appendix A, highlight how each resampling method affected the performance of the machine learning models on the imbalanced stroke dataset.

Accuracy
Models trained on the original imbalanced dataset (None) with Logistic Regression (0.951), the dataset processed with TomekLinks using Logistic Regression (0.951), and the original imbalanced dataset with Gradient Boosting Classifier (0.951) all achieved the highest accuracy. However, this metric was misleading, since models achieved near-perfect accuracy by predicting the majority classification (non-stroke) while entirely failing to detect stroke cases (Recall = 0.000). This highlights the critical limitation of accuracy as a performance measure in highly imbalanced datasets.

Precision
Precision was highest at 0.413 with Gradient Boosting and no resampling. However, once resampling techniques were introduced, precision dropped significantly. For example, when using NearMiss2 with Random Forest, precision fell to just 0.049. This reflects a common trade-off in resampling: as recall improves, precision tends to decline (Saito & Rehmsmeier, 2015). No single method achieved strong results for both.

Recall
As shown in Table 2, without resampling, recall scores were nearly zero across all models, which means they failed to identify most stroke cases. The highest recall, 0.905, was achieved by NearMiss2 combined with Random Forest. While this result shows that almost all stroke cases were caught, it came at the cost of many false alarms. RandomUnder combined with Logistic Regression offered a more balanced approach, reaching a recall of 0.784 and a more moderate precision of 0.131.

F1-Score
F1-scores were generally low across all combinations. The best result came from RandomUnder with Logistic Regression, reaching 0.224. Although SMOTE increased recall for Gradient Boosting up to 0.447, its F1-score remained lower at 0.201. This suggests that RandomUnder offered a better balance between recall and precision in this context.

ROC-AUC
Several models without resampling showed high ROC-AUC values. For example, Gradient Boosting reached 0.839. However, this did not reflect meaningful performance, since the models failed to detect stroke cases. Among the resampled methods, ENN combined with Gradient Boosting achieved the highest ROC-AUC at 0.846. Still, its recall remained low, which reinforces the idea that ROC-AUC can be misleading when working with imbalanced datasets.

PR-AUC
Precision-recall area under the curve scores were low across the board. The highest score was 0.213, recorded by TomekLinks with Gradient Boosting. These low values show how difficult it is to achieve both strong precision and recall when stroke cases are rare.

Precision-Recall Trade-off
Each resampling method showed a clear trade-off between precision and recall. NearMiss2 reached very high recall, such as 0.905 with Random Forest, but suffered a major loss in precision. Non-resampled models had high precision but almost no recall.

RandomUnder combined with Logistic Regression stood out as a reasonable compromise. It offered a recall of 0.784 and a precision of 0.131. SMOTE with Gradient Boosting also performed well in recall at 0.447, though its precision was lower.

Confusion Matrix of a Well-Balanced Model: RandomUnder with Logistic Regression
Table 3 of Appendix B shows the confusion matrix for the Random Undersampling + Logistic Regression configuration, which demonstrated one of the best trade-offs between precision and recall.

ISSN: 2473-4901
v11 n6330

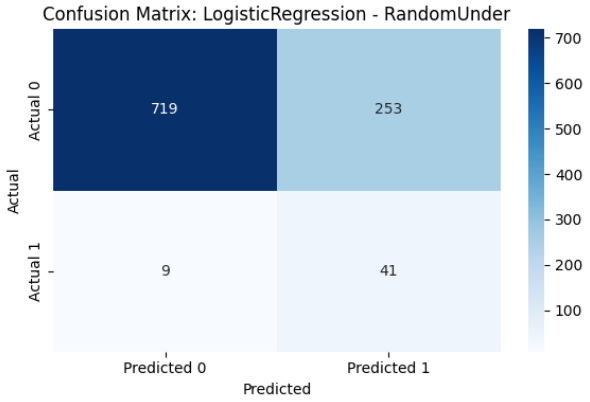


Figure 3: Confusion Matrix of Logistic Regression with Random Undersampling Method

Figure 3 above shows the model correctly identified 41 stroke cases (true positives) while minimizing false negatives (only 9 missed cases). Although 253 non-stroke cases were incorrectly classified as strokes (false positives), the model achieved a strong recall of 0.784, making it a practical choice for screening scenarios where detecting true stroke cases is critical.

Notable Exceptions (p-value > 0.05):

Logistic Regression (LR):

Model + Method	Model + Method	p-value
LR + None	LR + TL	1.00
	LR + ENN	0.76
LR + TL	LR + ENN	0.76
LR + NM2	LR + NM3	0.78

Random Forest (RF):

Model + Method	Model + Method	p-value
RF + None	RF + TL	0.84
	RF + ENN	0.77
	LR + None	0.84
	LR + TL	0.84
	LR + ENN	0.92
RF + RU	LR + RU	0.21
RF + TL	LR + None	0.69
	LR + TL	0.69
	RF + ENN	0.92
RF + ENN	LR + None	0.62
	LR + TL	0.62
	RF + ENN	0.84
RF + NM3	LR + NM3	0.08

Statistical Significance: t-test

Tables 4(a) and 4(b) of Appendix C present the p-values obtained from paired t-tests comparing the accuracy of various machine learning models under different resampling methods for stroke prediction. The majority of the p-values are approximately 0.00, suggesting that differences in model performance are statistically significant across most resampling techniques. However, several exceptions with higher p-values were observed, indicating no statistically significant difference in those specific comparisons.

Extreme Gradient Boosting (XGB):

Model + Method	Model + Method	p-value
XGB + None	XGB + TL	0.84
	XGB + ENN	0.12
	LR + None	0.55
	LR + TL	0.55
	LR + ENN	0.77
XGB + RU	LR + RU	0.05
	LR + NM2	0.15
	LR + NM3	0.09
XGB + TL	LR + None	0.69
	LR + TL	0.69
	LR + ENN	0.92
	XGB + ENN	0.08

Gradient Boosting (GB):

Model + Method	Model + Method	p-value
GB + None	GB + TL	1.00
	GB + ENN	0.39
	LR + None	0.76
	LR + TL	0.76
	LR + ENN	1.00
GB + RU	GB + NM3	0.92
	LR + NM2	0.81
	LR + NM3	0.60
GB + TL	LR + None	0.76
	LR + TL	0.76
	LR + ENN	1.00
	GB + ENN	0.39
GB + ENN	LR + None	0.24
	LR + TL	0.24
	LR + ENN	0.39
GB + NM3	LR + NM2	0.74
	LR + NM3	0.54

These findings highlight that some resampling strategies yield similar classification performance, especially when applied to models that share similar decision boundary behavior or sensitivity to classification imbalance (e.g., Tomek Links and ENN). These statistically non-significant results offer insight into which combinations may provide equivalent predictive performance, allowing for flexibility in method selection.

Answering Research Questions

After evaluating individual performance metrics, this study aimed to answer two guiding research questions related to the effectiveness of resampling strategies for stroke prediction.

RQ1: How do SMOTE and selected undersampling techniques compare in improving the performance of machine learning models for stroke prediction using highly imbalanced datasets?

The results across all metrics and models indicate that both SMOTE and undersampling techniques significantly outperformed models trained without resampling. SMOTE was especially effective in boosting recall for ensemble models like Gradient Boosting, while RandomUnder with Logistic Regression achieved the best balance between recall (0.784) and F1-score (0.224). NearMiss2 achieved the highest recall overall (0.905 with Random Forest), but at the cost of extremely low precision. These findings confirm that resampling methods are essential for improving minority-class detection and overall model effectiveness.

RQ2: What are the trade-offs between predictive performance when using SMOTE versus undersampling techniques in highly imbalanced stroke prediction models?

The analysis showed a consistent trade-off between recall and precision across all resampling methods. Techniques like SMOTE and NearMiss2 greatly improved recall but significantly reduced precision, leading to more false positives. Among the resampling techniques, RandomUndersampling offered the best compromise, demonstrating that carefully chosen undersampling methods can improve recall without overwhelming the system with false positives.

The next section provides concluding insights based on the findings and details of the limitations of this study.

5. CONCLUSION AND LIMITATIONS

This study conducted a comprehensive comparative analysis of SMOTE and various undersampling techniques for addressing classification imbalance in stroke prediction using four artificial intelligence (AI)-driven machine learning models: Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Gradient Boosting (GB). By evaluating model performance across key AI evaluation metrics—including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC—this study provides actionable insights into how data-level resampling methods influence the reliability and fairness of AI systems in healthcare.

The results demonstrate that all resampling strategies significantly improved the ability of AI models to detect minority stroke cases compared to models trained on imbalanced data. Notably, SMOTE and NearMiss2 achieved high recall rates, enhancing the AI system's sensitivity to stroke events, while Random Undersampling combined with Logistic Regression achieved the most balanced performance. These findings affirm the critical role of resampling in enhancing the trustworthiness and effectiveness of AI-based stroke prediction tools. The statistical significance of most performance differences, confirmed through paired t-tests, further validates the robustness of these observations.

The trade-off between recall and precision observed in the study has important cost implications for stroke prediction. High recall models, such as NearMiss2 with Random Forest, successfully identified nearly all stroke cases but did so at the expense of extremely low precision, resulting in a large number of false positives. In clinical practice, false positives generate unnecessary diagnostic procedures, increase healthcare expenditures, and contribute to patient anxiety. They may also create "alert fatigue" for clinicians, reducing trust in AI systems and limiting adoption. Conversely, models with high precision but low recall risk producing false negatives, where actual stroke cases are missed. This outcome carries an even higher patient cost, as it delays treatment and increases the likelihood of long-term disability or mortality.

From a methodological perspective, this trade-off demonstrates that optimizing for recall alone is insufficient. Instead, model evaluation must consider both types of errors and their asymmetric consequences. Future research should incorporate cost-sensitive learning frameworks that explicitly weight false negatives more heavily, while controlling the operational burden of false positives (Khan et al., 2017). Such approaches would allow stroke prediction models to better align with the realities of clinical decision-making, where both medical safety and resource efficiency are critical

Despite its contributions, this study has several limitations. First, the stroke dataset provided by Fedesoriano on Kaggle is a publicly available collection of patient-level records designed for predictive modeling of stroke risk. While the

ISSN: 2473-4901
v11 n6330

dataset offers a useful benchmark for developing and testing machine learning models, it has notable data integrity considerations. For example, there are missing values in the BMI attribute that require imputation before analysis. Furthermore, since the dataset is aggregated and anonymized, there is limited information on its clinical provenance, meaning that while it is suitable for methodological exploration and comparative studies, caution should be exercised in generalizing findings to real-world clinical populations.

Next, the analysis is confined to a single publicly available stroke dataset, which may affect the generalizability of results to other medical conditions or populations. All models and resampling methods were applied using default hyperparameters, suggesting that further tuning could yield even stronger results. Third, while traditional classification metrics were used, the study did not explicitly incorporate fairness, interpretability, or cost-sensitive evaluation—critical considerations for the responsible deployment of AI in clinical settings. Finally, the use of t-tests assuming equal variances may not fully account for dependencies introduced through resampling.

The following section outlines the future research directions based on the findings of this study.

6. FUTURE RESEARCH DIRECTIONS

Future AI-driven research should expand this work by exploring hybrid resampling strategies, integrating fairness-aware and interpretable AI models, tuning hyperparameters, and validating findings across multiple datasets. These enhancements will support the development of more accurate, equitable, and clinically actionable AI systems for early stroke detection and other high-impact healthcare applications.

Specifically, the following avenues should be considered and implemented below:

- **Exploring Hybrid and Ensemble Resampling Methods**: Combining oversampling and undersampling strategies (e.g., SMOTE-ENN, SMOTE-Tomek) or integrating resampling within ensemble frameworks (e.g., BalancedBagging) may further improve performance.

- **Model Tuning and Optimization**: Future studies should investigate the impact of hyperparameter tuning on both classifiers and resampling methods to optimize performance.

- **Fairness and Interpretability**: Incorporating fairness-aware algorithms and interpretable models is critical, especially when deploying in high-stakes domains like healthcare.

- **Cross-Dataset Evaluation**: To ensure generalizability, testing on multiple stroke or related healthcare datasets from diverse populations would strengthen the findings.

- **Cost-Sensitive Learning**: Integrating cost-sensitive learning approaches could help reduce false negatives while accounting for the asymmetric costs of misclassification in clinical decision-making.

7. REFERENCES

Aish, M. A., Ghafoor, A. A., Nasim, F., Ali, K. I., Akhter, S., & Azeem, S. (2024). Improving stroke prediction accuracy through machine learning and synthetic minority over-sampling. *Journal of Computing & Biomedical Informatics, 7*(2), 566-0702. https://doi.org/10.56979/702/2024

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR), 49*(2), Article 31, 1-50. https://doi.org/10.1145/2907070

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

ISSN: 2473-4901
v11 n6330

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review, 57*, 137. https://doi.org/10.1007/s10462-024-10759-6

Choi, Y. J., Kim, J. H., Lee, S. H., & Park, M. J. (2024). Explainable artificial intelligence for stroke prediction through deep learning and machine learning models. *Scientific Reports, 14*, Article 82931. https://doi.org/10.1038/s41598-024-82931-5

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery, 24*, 136-158. https://doi.org/10.1007/s10618-011-0222-1

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, 233-240. https://doi.org/10.1145/1143844.1143874

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1-30. https://dl.acm.org/doi/10.5555/1248547.1248548

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91, 464–471. https://doi.org/10.1016/j.eswa.2017.09.030

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning, 113*(12), 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fedesoriano. (n.d.). *Stroke Prediction Dataset*. Kaggle. Retrieved April 21, 2025, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Fernández, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018a). *Learning from Imbalanced Data Sets*. Springer.

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018b). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration. (2014). Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: A pooled analysis of 97 prospective cohorts with 1.8 million participants. *The Lancet, 383*(9921), 970–983. https://doi.org/10.1016/S0140-6736(13)61836-X

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: review of methods and applications. *Expert Systems with Applications*, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *In Proceedings of the International Conference on Intelligent Computing, 3644,* 878-887. Springer. https://doi.org/10.1007/11538059_91

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hoens, T. R., Qian, Q., Chawla, N. V., & Zhou, Z. H. (2012). Building decision trees for the multi-class imbalance problem. *In Pacific-Asia conference on knowledge discovery and data mining*, 122-134. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30217-6_11

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced datasets: from sampling to classifiers. Imbalanced learning: foundations, algorithms, and applications, 43-59. https://doi.org/10.1002/9781118646106.ch3

Japkowicz, N. (2000a). The classification imbalance problem: significance and strategies. *In Proceedings of the 2000*

ISSN: 2473-4901
v11 n6330

International Conference on Artificial Intelligence, 56, 111-117.

Japkowicz, N. (2000b). Learning from imbalanced data sets: a comparison of various strategies. In *Proceedings of the AAAI 2000 Workshop on Learning from Imbalanced Data Sets*. AAAI Press.

Japkowicz, N., & Stephen, S. (2002). The classification imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429-449. https://doi.org/10.3233/IDA-2002-6504

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. *In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 245-251. https://doi.org/10.1109/ACII.2013.47

Kamiri, J., & Mariga, G. W. (2021). Research methods in machine learning: a content analysis. *International Journal of Computer and Information Technology, 10*(2), 78-84. https://doi.org/10.24203/ijcit.v10i2.79

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems, 29*(8), 3573-3587. https://doi.org/10.1109/TNNLS.2017.2732482

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing classification distribution. In *Proceedings of the 8th Conference on AI in Medicine in Europe, 2101,* 63-66. Springer. https://doi.org/10.1007/3-540-48229-6_9

LemaÃŽtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research, 18*(17), 1-5. http://jmlr.org/papers/v18/16-365.html.

Lin, CH., Chen, YA., Jeng, JS., Sun, Y., Wei, CY., Yeh, PY., Chang, WL., Fann, YC., Hsu, KC., Lee, JT., & Taiwan Stroke Registry Investigators. (2024). Predicting ischemic stroke patients' prognosis changes using machine learning in a nationwide stroke registry. *Medical & Biological Engineering & Computing, 62*, 2343-2354. https://doi.org/10.1007/s11517-024-03073-4

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141. https://doi.org/10.1016/j.ins.2013.07.007

Mani, I., & Zhang, I. (2003). kNN approach to unbalanced data distributions: A case study involving information extraction. *In Proceedings of Workshop on Learning from Imbalanced Datasets, 126,* 1-7. ICML.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Pan, B., Jin, X., Jun, L., Qiu, S., Zheng, Q., & Pan, M. (2019). The relationship between smoking and stroke: A meta-analysis. *Medicine, 98*(12), e14872. https://doi.org/10.1097/MD.0000000000014872

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, 291, 184–203. https://doi.org/10.1016/j.ins.2014.08.051

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review, 57*, 273. https://doi.org/10.1007/s10462-024-10884-2

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution, 10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

ISSN: 2473-4901
v11 n6330

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: a review. *International journal of pattern recognition and artificial intelligence*, *23*(04), 687-719. https://doi.org/10.1142/S0218001409007326

Swaminathan, S., & Tantri, B. R. (2024). Confusion matrix-based performance evaluation metrics. *African Journal of Biomedical Research, 27*(4s), 4023-4031. https://doi.org/10.53555/AJBR.v27i4S.4345

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics, 6*(11), 769-772. http://dx.doi.org/10.1109/TSMC.1976.4309452

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-2*(3), 408-421. https://doi.org/10.1109/TSMC.1972.4309137

Yen, S. J., & Lee, Y. S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications, 36*(3), Part 1, 5718-5727. https://doi.org/10.1016/j.eswa.2008.06.108

APPENDIX A

Method	Model	*Accuracy	*Precision	*Recall	*F1	*ROC-AUC	*PR-AUC
No Resampling	LR	**0.951**	0.000	0.000	0.000	0.838	0.187
	RF	0.950	0.040	0.005	0.009	0.813	0.171
	XGBoost	0.941	0.217	**0.075**	**0.111**	0.808	0.173
	GB	**0.951**	**0.413**	0.025	0.046	**0.839**	**0.201**
SMOTE	LR	0.779	0.122	**0.568**	0.200	0.777	0.138
	RF	**0.899**	**0.147**	0.221	0.176	**0.791**	0.146
	XGBoost	**0.899**	0.134	0.196	0.159	0.771	0.135
	GB	0.828	0.130	0.447	**0.201**	0.778	**0.149**
RandomUnder	LR	**0.735**	**0.131**	0.784	**0.224**	**0.833**	**0.187**
	RF	0.705	0.119	**0.804**	0.209	0.826	0.159
	XGBoost	0.715	0.121	0.774	0.209	0.804	0.147
	GB	0.704	0.120	0.794	0.208	0.820	0.183
TomekLinks	LR	**0.951**	0.000	0.000	0.000	0.838	0.186
	RF	0.950	0.067	0.005	0.009	0.808	0.176
	XGBoost	0.941	0.246	**0.090**	**0.130**	0.815	0.173
	GB	0.950	**0.390**	0.040	0.072	**0.840**	**0.213**
ENN	LR	**0.946**	0.192	0.030	0.051	0.840	0.192
	RF	0.938	0.158	0.050	0.074	0.830	0.182
	XGBoost	0.920	0.204	**0.206**	**0.204**	0.822	0.171
	GB	0.934	**0.257**	0.136	0.172	**0.846**	**0.202**
NearMiss1	LR	**0.437**	**0.051**	0.553	**0.092**	**0.503**	**0.066**
	RF	0.196	0.041	0.688	0.077	0.432	0.054
	XGBoost	0.206	0.042	0.699	0.079	0.440	0.050
	GB	0.183	0.043	**0.739**	0.081	0.384	0.039
NearMiss2	LR	**0.663**	**0.103**	0.768	**0.182**	**0.775**	**0.139**
	RF	0.143	0.049	**0.905**	0.093	0.664	0.098
	XGBoost	0.104	0.047	**0.905**	0.089	0.685	0.131
	GB	0.112	0.047	0.900	0.090	0.498	0.073
NearMiss3	LR	**0.707**	**0.101**	**0.633**	**0.174**	**0.750**	**0.142**
	RF	0.655	0.084	0.617	0.147	0.670	0.114
	XGBoost	0.654	0.084	0.613	0.148	0.676	0.120
	GB	0.673	0.088	0.608	0.153	0.710	0.128

Table 2: Results of Machine Learning Models under Different Resampling Methods for Stroke Prediction

ISSN: 2473-4901
v11 n6330

APPENDIX B

Method	Model	True-Negative	False-Positive	False-Negative	True-Positive
None	LR	**972**	**0**	50	0
	RF	970	2	50	0
	XGBoost	960	12	**44**	**6**
	GB	968	4	49	1
SMOTE	LR	767	205	**14**	**36**
	RF	906	66	41	9
	XGBoost	**912**	**60**	43	7
	GB	816	156	27	23
RandomUnder	LR	719	253	**9**	**41**
	RF	695	277	10	40
	XGBoost	685	287	14	36
	GB	657	315	11	39
TomekLinks	LR	**971**	**1**	49	1
	RF	968	4	50	0
	XGBoost	961	11	**43**	**7**
	GB	967	5	48	2
ENN	LR	**964**	**8**	45	5
	RF	955	17	38	12
	XGBoost	935	37	**36**	**14**
	GB	952	20	42	8
NearMiss1	LR	**340**	**632**	22	28
	RF	131	841	11	39
	XGBoost	124	848	**5**	**45**
	GB	130	842	10	40
NearMiss2	LR	**652**	**320**	11	39
	RF	91	881	3	47
	XGBoost	61	911	**1**	**49**
	GB	64	908	4	46
NearMiss3	LR	656	316	**21**	**29**
	RF	619	353	22	28
	XGBoost	648	324	25	25
	GB	**669**	**303**	**21**	**29**

Table 3: Confusion Matrix Results of Machine Learning Models under Different Resampling Methods for Stroke Prediction

APPENDIX C

Model + Method	**LR + SMOTE**	**LR + RU**	**LR + TL**	**LR + ENN**	**LR + NM1**	**LR + NM2**	**LR + NM3**
LR + None	**0.00***	**0.00***	1.00	0.76	**0.00***	**0.00***	**0.00***
LR + SMOTE		**0.02***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
LR + RU			**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
LR + TL				0.76	**0.00***	**0.00***	**0.00***
LR + ENN					**0.00***	**0.00***	**0.00***
LR + NM1						**0.00***	**0.00***
LR + NM2							0.78
Model + Method	**RF + SMOTE**	**RF + RU**	**RF + TL**	**RF + ENN**	**RF + NM1**	**RF + NM2**	**RF + NM3**
RF + None	**0.00***	**0.00***	0.84	0.77	**0.00***	**0.00***	**0.00***
RF+ SMOTE		**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
RF + RU			**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
RF + TL				0.92	**0.00***	**0.00***	**0.00***
RF + ENN					**0.00***	**0.00***	**0.00***
RF + NM1						**0.048***	**0.00***
RF + NM2							**0.00***
Model + Method	**XGB + SMOTE**	**XGB + RU**	**XGB + TL**	**XGB + ENN**	**XGB + NM1**	**XGB + NM2**	**XGB + NM3**
XGB + None	**0.00***	**0.00***	0.84	0.12	**0.00***	**0.00***	**0.00***
XGB + SMOTE		**0.00***	**0.00***	**0.02***	**0.00***	**0.00***	**0.00***
XGB + RU			**0.00***	**0.00***	**0.00***	**0.00***	**0.02***
XGB + TL				0.08	**0.00***	**0.00***	**0.00***
XGB + ENN					**0.00***	**0.00***	**0.00***
XGB + NM1						**0.00***	**0.00***
XGB + NM2							**0.00***
Model + Method	**GB + SMOTE**	**GB + RU**	**GB + TL**	**GB + ENN**	**GB + NM1**	**GB + NM2**	**GB + NM3**
GB + None	**0.00***	**0.00***	1.00	0.39	**0.00***	**0.00***	**0.00***
GB + SMOTE		**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
GB + RU			**0.00***	**0.00***	**0.00***	**0.00***	0.92
GB + TL				0.39	**0.00***	**0.00***	**0.00***
GB + ENN					**0.00***	**0.00***	**0.00***
GB + NM1						**0.00***	**0.00***
GB + NM2							**0.00***

Table 4(a): P-Value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

t-test: Two-sample Assuming Equal Variances
p-value* < 0.05: Significant difference between pair of Model + Method

Model + Method	**LR + None**	**LR + SMOTE**	**LR + RU**	**LR + TL**	**LR + ENN**	**LR + NM1**	**LR + NM2**	**LR + NM3**
RF + None	0.84	**0.00***	**0.00***	0.84	0.92	**0.00***	**0.00***	**0.00***
RF+ SMOTE	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
RF + RU	**0.00***	**0.00***	0.21	**0.00***	**0.00***	**0.00***	**0.03***	**0.02***
RF + TL	0.69	**0.00***	**0.00***	0.69	0.92	**0.00***	**0.00***	**0.00***
RF + ENN	0.62	**0.00***	**0.00***	0.62	0.84	**0.00***	**0.00***	**0.00***
RF + NM1	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
RF + NM2	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
RF + NM3	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.04***	0.08
XGB + None	0.55	**0.00***	**0.00***	0.55	0.77	**0.00***	**0.00***	**0.00***
XGB + SMOTE	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
XGB + RU	**0.00***	**0.00***	0.05	**0.00***	**0.00***	**0.00***	0.15	0.09
XGB + TL	0.69	**0.00***	**0.00***	0.69	0.92	**0.00***	**0.00***	**0.00***
XGB + ENN	**0.03***	**0.00***	**0.00***	**0.03***	0.07	**0.00***	**0.00***	**0.00***
XGB + NM1	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
XGB + NM2	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
XGB + NM3	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	0.40	0.57
GB + None	0.76	**0.00***	**0.00***	0.76	1.00	**0.00***	**0.00***	**0.00***
GB + SMOTE	**0.00***	**0.045***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
GB + RU	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	0.81	0.60
GB + TL	0.76	**0.00***	**0.00***	0.76	1.00	**0.00***	**0.00***	**0.00***
GB + ENN	0.24	**0.00***	**0.00***	0.24	0.39	**0.00***	**0.00***	**0.00***
GB + NM1	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
GB + NM2	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***
GB + NM3	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	**0.00***	0.74	0.54

Table 4(b): P-Value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

t-test: Two-sample Assuming Equal Variances
p-value* < 0.05: Significant difference between pair of Model + Method

ISSN: 2473-4901
v11 n6330

APPENDIX D

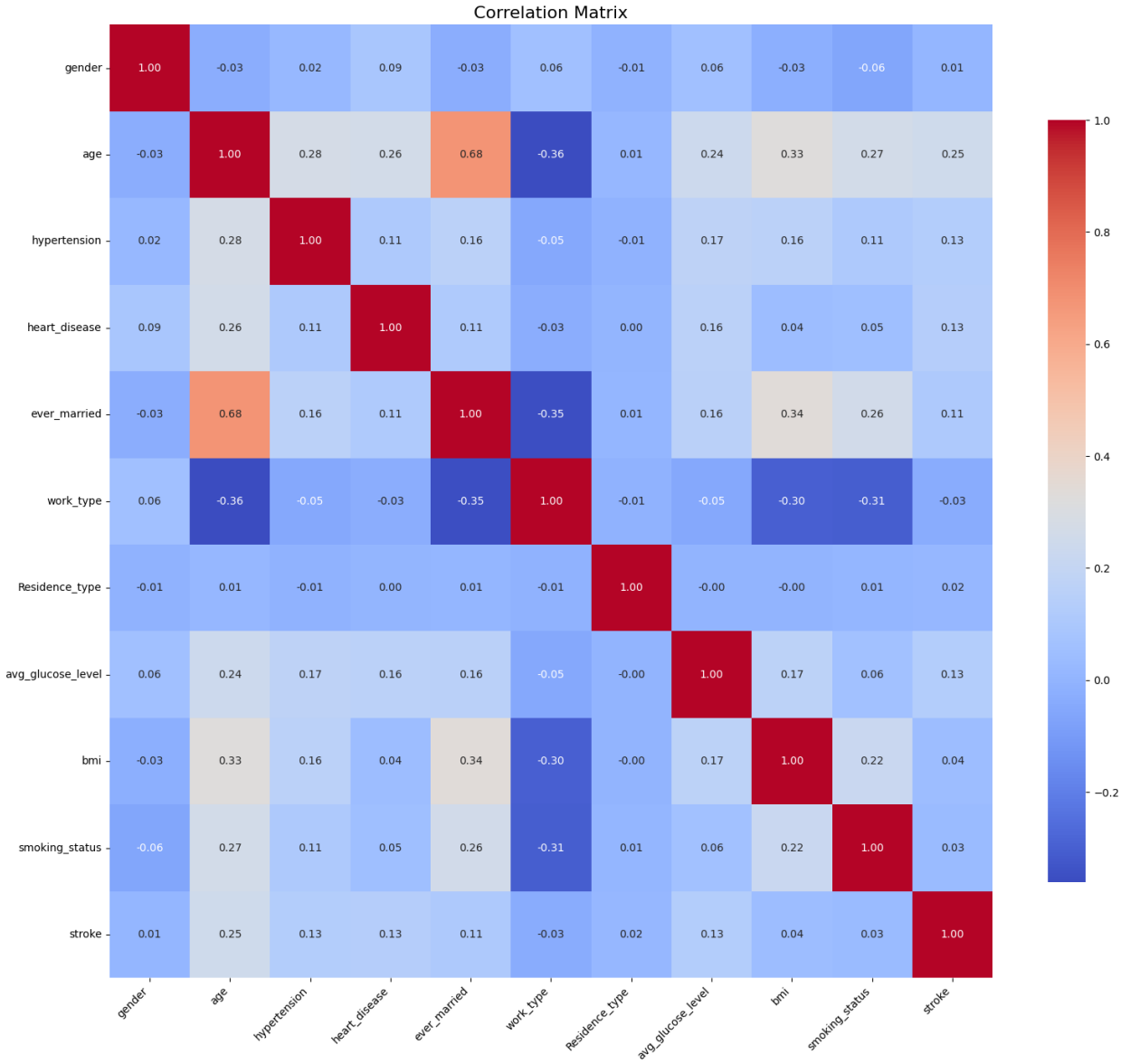


Figure 4: Correlation Matrix

APPENDIX E

Model + Method	LR + SMOTE	LR + RU	LR + TL	LR + ENN	LR + NM1	LR + NM2	LR + NM3
LR + None	11.40	13.61	0.00	0.30	35.88	17.05	17.35
LR + SMOTE		2.24	-11.40	-11.13	21.54	5.63	5.91
LR + RU			-13.61	-13.34	18.89	3.37	3.65
LR + TL				0.30	35.88	17.05	17.35
LR + ENN					35.54	16.78	17.08
LR + NM1						-15.06	-14.75
LR + NM2							0.28
Model + Method	**RF + SMOTE**	**RF + RU**	**RF + TL**	**RF + ENN**	**RF + NM1**	**RF + NM2**	**RF + NM3**
RF + None	4.56	14.69	0.20	0.30	57.85	64.02	19.06
RF+ SMOTE		10.35	-4.37	-4.27	48.32	52.94	14.67
RF + RU			-14.51	-14.42	30.27	33.06	4.18
RF + TL				0.10	57.43	63.53	18.89
RF + ENN					57.23	63.29	18.80
RF + NM1						1.98	-24.49
RF + NM2							-26.93
Model + Method	**XGB + SMOTE**	**XGB + RU**	**XGB + TL**	**XGB + ENN**	**XGB + NM1**	**XGB + NM2**	**XGB + NM3**
XGB + None	3.89	15.03	-0.20	1.55	57.20	69.61	17.42
XGB + SMOTE		11.33	-4.08	-2.37	49.04	58.54	13.69
XGB + RU			-15.21	-13.62	29.35	34.66	2.28
XGB + TL				1.74	57.06	70.18	17.59
XGB + ENN					53.95	65.10	15.99
XGB + NM1						3.81	-26.16
XGB + NM2							-31.07
Model + Method	**GBC + SMOTE**	**GBC + RU**	**GBC + TL**	**GBC + ENN**	**GBC + NM1**	**GBC + NM2**	**GBC + NM3**
GB + None	9.18	16.54	0.00	0.86	57.64	70.48	16.44
GB + SMOTE		7.41	-9.18	-8.38	39.13	46.23	7.31
GB + RU			-16.54	-15.76	27.57	32.73	-0.09
GB + TL				0.86	57.64	70.48	16.44
GB + ENN					55.84	67.93	15.66
GB + NM1						3.87	-27.70
GB + NM2							-32.88

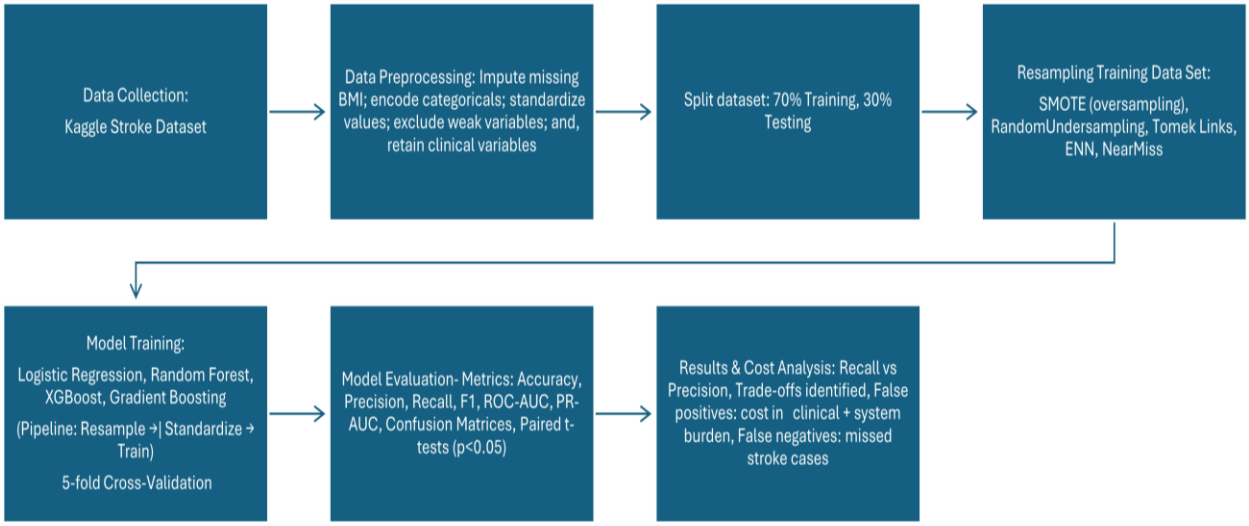
Table 5(a): t-stat value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

Model + Method	LR + None	LR + SMOTE	LR + RU	LR + TL	LR + ENN	LR + NM1	LR + NM2	LR + NM3
RF + None	-0.20	11.22	13.43	-0.20	0.10	35.65	16.87	17.17
RF+ SMOTE	-4.76	6.84	9.09	-4.76	-4.47	30.04	12.52	12.82
RF + RU	-14.86	-3.49	-1.25	-14.86	-14.60	17.45	2.12	2.40
RF + TL	-0.40	11.04	13.26	-0.40	-0.10	35.42	16.70	16.99
RF + ENN	-0.50	10.95	13.17	-0.50	-0.20	35.31	16.61	16.91
RF + NM1	-58.27	-35.72	-32.14	-58.27	-57.64	-10.19	-27.24	-26.85
RF + NM2	-64.52	-38.93	-35.07	-64.52	-63.78	-12.20	-29.84	-29.42
RF + NM3	-19.24	-7.71	-5.43	-19.24	-18.98	12.82	-2.05	-1.76
XGB + None	-0.60	10.86	13.08	-0.60	-0.30	35.19	16.52	16.82
XGB + SMOTE	-4.47	7.13	9.37	-4.47	-4.18	30.40	12.81	13.11
XGB + RU	-15.56	-4.18	-1.93	-15.56	-15.30	16.67	1.44	1.72
XGB + TL	-0.40	11.04	13.26	-0.40	-0.10	35.42	16.70	16.99
XGB + ENN	-2.14	9.42	11.66	-2.14	-1.84	33.35	15.10	15.40
XGB + NM1	-58.44	-35.81	-32.23	-58.44	-57.81	-10.25	-27.31	-26.92
XGB + NM2	-71.37	-42.14	-37.96	-71.37	-70.48	-14.12	-32.36	-31.93
XGB + NM3	-17.94	-6.48	-4.22	-17.94	-17.68	14.13	-0.84	-0.56
GB + None	-0.30	11.13	13.34	-0.30	0.00	35.54	16.78	17.08
GB + SMOTE	-9.45	2.00	4.25	-9.45	-9.18	23.97	7.65	7.94
GB + RU	-16.80	-5.39	-3.13	-16.80	-16.54	15.33	0.24	0.52
GB + TL	-0.30	11.13	13.34	-0.30	0.00	35.54	16.78	17.08
GB + ENN	-1.17	10.34	12.57	-1.17	-0.86	34.53	16.01	16.31
GB + NM1	-58.27	-35.72	-32.14	-58.27	-57.64	-10.19	-27.24	-26.85
GB + NM2	-71.37	-42.14	-37.96	-71.37	-70.48	-14.12	-32.36	-31.93
GB + NM3	-16.70	-5.29	-3.04	-16.70	-16.44	15.43	0.33	0.61

Table 5(b): T-stat value from t-test of Comparing Accuracy Between Machine Learning Models under Different Resampling Methods for Stroke Prediction

ISSN: 2473-4901
v11 n6330

APPENDIX F: Process Model for Stroke Prediction Analysis



The workflow includes data collection, preprocessing, resampling, model training with 5-fold stratified cross-validation, and multi-metric evaluation. Results are analyzed to highlight the trade-offs between precision and recall, including the practical costs of false positives and false negatives.