Optimizing Cerebral Stroke Prediction: Combining SMOTE-ENN and XGBoost on Imbalanced Medical Data

Rohit Kumar Agrawal*, Vishal Meena*, Rishi Raj*, Mahendra Kumar Gaurisaria*, Manoj Sahni†, Walayat Hussain‡
*School of Computer Science and Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India
Email: {2206421@kiit.ac.in, vishal.meenafcs@kiit.ac.in, rishi.rajfcs@kiit.ac.in, mkgourisariafcs@kiit.ac.in}
†Department of Mathematics, School of Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India
Email: manojsahani117@gmail.com
‡Peter Faber Business School, Australian Catholic University, North Sydney, 2160, NSW, Australia
Email: walayat.hussain@acu.edu.au

***Abstract*—Cerebral stroke represents a major public health issue globally, contributing to elevated rates of morbidity and mortality. The ability to predict stroke risk with precision at an early stage is essential for effective preventive measures. Medical datasets for stroke prediction are often imbalanced, leading to challenges in model performance, particularly in minimizing false negatives. This paper presents a novel method that integrates The integration of SMOTE-ENN (Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors) and XGBoost is employed to enhance the accuracy and sensitivity of stroke prediction on imbalanced medical datasets. The proposed approach addresses the class imbalance by first applying SMOTE-ENN, which oversamples the minority class and eliminates noisy instances, creating a more balanced and cleaner dataset. This preprocessed data is then used to train the XGBoost classifier, which further enhances prediction performance. The experimental outcomes reveal that the proposed methodology markedly enhances sensitivity and accuracy, establishing it as a reliable tool for stroke risk evaluation in datasets with class imbalances.**

This hybrid method was evaluated using a real-world medical dataset of 43,400 samples, including 783 stroke occurrences. The optimized XGBoost model achieved a 98.14% accuracy, with a false negative rate of 12.61% and a sensitivity of 87.39%. These results demonstrate a significant improvement over traditional approaches, reducing the false negative rate by 34% compared to earlier methods, with a minimal increase in false positives.

***Index Terms*—Stroke prediction, SMOTE-ENN, XGBoost, Class imbalance, Machine learning, Medical data analysis**

I. Introduction

Stroke remains a major public health concern globally, contributing significantly to morbidity, mortality, and long-term disability. Stroke ranks as the second most common cause of death globally, accounting for approximately 11% of total fatalities based on recent data. Although medical advancements and preventive strategies have progressed, the burden of stroke remains significantly higher in low- and middle-income countries (LMICs).

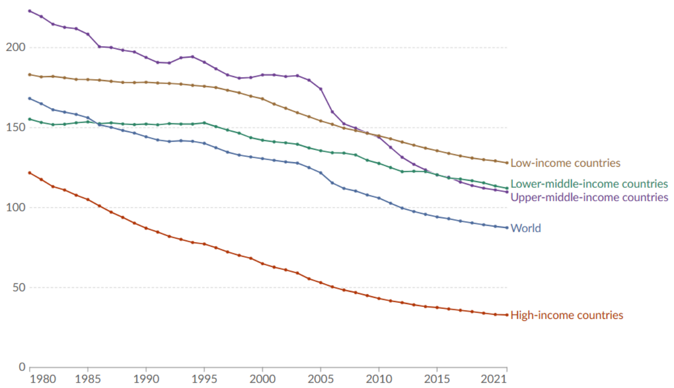


Fig. 1: Global stroke mortality trends (1980-2021) based on income level. Source: IHME, Global Burden of Disease (2024), Our World in Data

As shown in Figure 1, high-income nations have achieved notable reductions in stroke-related mortality over recent decades due to enhanced healthcare systems, early diagnostic tools, and improved risk factor management. However, in LMICs, the decline in stroke mortality has been slower, and the rates remain alarmingly high. This inequality in health outcomes emphasizes the need for advanced predictive models that can identify individuals at high risk of stroke, especially in resource-constrained settings.

A. Opportunities and Challenges

One of the significant challenges in stroke prediction lies in the imbalanced nature of medical datasets, where instances of stroke cases are often overshadowed by a larger number of non-stroke cases. Traditional machine learning models tend to perform poorly on such datasets, leading to a higher rate of false negatives, which can be particularly dangerous in medical diagnostics.

In this context, our study proposes a novel hybrid approach to stroke prediction by combining SMOTE-ENN for data balancing and XGBoost for classification. This approach aims to address the issue of data imbalance while improving the accuracy and sensitivity of stroke prediction models.

B. Related Works

Over the past two decades, machine learning applications in stroke prediction have witnessed significant advancements in scientific research [9], [10]. Further advancements include the development of spiking neural network reservoir systems, which significantly minimized prediction errors [12]. Additionally, in their comprehensive review, Arslan et al. [13] identified SVM and SGB as particularly effective models for stroke prediction.

Although most research emphasizes the use of complete and balanced datasets, such conditions are uncommon in medical data. Iterative mechanism-based imputation techniques have emerged as promising solutions, delivering acceptable levels of predictive accuracy [14], [15]. Zhang et al. [16] advanced the field further by employing expectation maximization algorithms coupled with Bayesian classification for imputing missing values; however, class imbalance continued to pose a significant challenge. Class imbalance is inherently present in stroke datasets derived from clinical settings, necessitating specific considerations when utilizing learning-based prediction approaches.

Class imbalance arises from an unequal distribution of samples between minority and majority classes [17]. Data-level strategies, including undersampling and oversampling, are commonly employed, with the method chosen based on the size of the minority class in the original dataset [18], [19]Advanced techniques, such as distance-based sampling [20], SMOTE [21], and cluster sampling [22], have been developed to address the limitations of random sampling, such as information loss and class overlap.

Algorithmically, researchers have introduced extensive modifications to loss and objective functions in conventional prediction models [23]–[25]. A notable strategy involves redefining classification boundaries specifically for linearly separable datasets [26].Wang et al. [27] significantly contributed to the field by developing a robust loss function capable of minimizing misclassification costs across diverse noisy environments.Lin et al. [28] utilized focal loss in deep neural networks, emphasizing challenging examples to effectively manage class imbalance.Hybrid methodologies, including optimized k-NN [29] and fuzzy-based information decomposition techniques [30], have proven effective in addressing the complexities of imbalanced and incomplete datasets.

C. Our Approach

As highlighted, class imbalance and missing data are pivotal challenges in stroke prediction. Achieving high sensitivity (reducing false negatives) and maintaining acceptable specificity (reducing false positives) is a critical necessity in the medical field. To tackle these challenges, the proposed methodology employs a structured two-step process to construct a robust stroke prediction model for imbalanced and incomplete medical datasets.

1) Handling Missing Data: This approach involves analyzing the dataset's characteristics to implement targeted imputation techniques. For instance, missing BMI values are replaced with the median, while missing ”smoking status” entries are imputed as ’unknown.’ These strategies ensure that critical information is preserved for effective modeling.
2) We use a hybrid machine learning approach that combines SMOTE-ENN and XGBoost for class-imbalanced stroke prediction.

The contributions of this work are threefold:

- Hybrid Learning Framework: We here present a hybrid stroke prediction model combining SMOTE-ENN for data resampling and XGBoost for classification, effectively managing the class imbalance issue. Data-level resampling reduces overfitting, while algorithm-level optimization ensures model robustness without manual hyperparameter tuning.
- Hyperparameter Tuning with RandomizedSearchCV We include a hyperparameter tuning step to automatically choose the best-performing model configuration according to the validation set. This way, we are guaranteed that the model is well-tuned to accuracy and sensitivity, particularly in the case of class imbalance.
- Evaluation Metrics: A significant focus is placed on using class imbalance-sensitive metrics like false negative rate (FNR), false positive rate (FPR), sensitivity, and accuracy to evaluate the model. The goal of optimization is to lower FNR while keeping an acceptable FPR, hence making it suitable for use in real-world stroke prediction tasks.

II. Problem Description

Stroke is a life-taking medical condition, and sooner prediction is important to reduce its severity. However, stroke prediction faces many challenges, like missing data and class imbalance in medical datasets. Such obstacles generally skew the predictions towards the majority class while reducing accuracy for underrepresented cases, leading often to high rates of false negatives. Specifically, stroke datasets predominantly consist of ”non-stroke” (majority class) instances, whereas positive cases of stroke remain comparatively infrequent. Such an imbalance makes model training difficult, since conventional algorithms may fail to capture the minority-class patterns. Further to these, medical records usually carry incomplete data caused by missing entries in patient records, further complicating model development. Either such missing data can be handled via deletion, resulting in a loss of information or through imputation strategies, with the aim of estimating the missing values.

Let D = $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ denote the dataset, where $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{id}]$ is a vector of d features for patient i, and $y_i \in \{0, 1\}$ is the binary outcome. The number of stroke cases (N_1) is much smaller than the number of non-stroke cases (N_0), i.e., $N_1 \ll N_0$.

x_1	x_2	$\cdots$	x_{d-1}	x_d	y'
1	1	$\cdots$	1	1	Yes
1	1	$\cdots$	1	1	Yes
1	1	$\cdots$	1	1	Yes
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	1	No

(a) Complete and balanced

x_1	x_2	$\cdots$	x_{d-1}	x_d	y'
1	1	$\cdots$	1	0	Yes
1	1	$\cdots$	1	1	Yes
1	1	$\cdots$	0	0	Yes
1	1	$\cdots$	0	1	No
1	1	$\cdots$	1	1	No
1	1	$\cdots$	0	1	No

(b) Incomplete and balanced

x_1	x_2	$\cdots$	x_{d-1}	x_d	y'
1	1	$\cdots$	1	1	Yes
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	1	No

(c) Complete and imbalanced

x_1	x_2	$\cdots$	x_{d-1}	x_d	y'
1	1	$\cdots$	1	1	Yes
1	1	$\cdots$	0	1	No
1	1	$\cdots$	0	0	No
1	1	$\cdots$	0	0	No
1	1	$\cdots$	1	1	No
1	1	$\cdots$	1	0	No

(d) Incomplete and imbalanced

Fig. 2: Different dataset scenarios

The goal is to learn a function $f(\mathbf{x})$ that maps the input features $\mathbf{x}_i$ to the outcome y_i which minimizes the following empirical risk function:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f(x_i;\theta), y_i) \tag{1}$$

where $\ell(\cdot)$ represents a loss function and θ are the model parameters. However, the dataset is imbalanced, and directly optimizing this loss function may lead to poor performance on the minority class. To address this, we introduce a weighted loss function:

$$L_{weighted}(\theta) = \frac{1}{N_0}\sum_{i:y_i=0} \ell(f(x_i;\theta), 0) + \frac{1}{N_1}\sum_{i:y_i=1} \ell(f(x_i;\theta), 1) \tag{2}$$

This ensures that the minority class contributes equally to the overall loss, reducing the bias towards the majority class.

Additionally, to deal with missing values, let x_i^{miss} denote the set of missing features for instance i. Missing data can be imputed using:

- **Mean/Median Imputation:** Impute missing values using either the mean or median of the observed data.
- **K-Nearest Neighbors (KNN)** Imputation: The missing values are estimated using the feature values of the nearest neighbors.
- **Multivariate Imputation:** This is using models like regression to predict missing values based on observed data. Finally, class imbalance is handled by combining the **Synthetic Minority Over-sampling Technique (SMOTE)** with **Edited Nearest Neighbors (ENN)**. This method produces synthetic instances for the minority class and cleans the majority class of noisy or redundant data. Formally, let the synthetic sample for the minority class be generated as:

$$x_{synthetic} = x_i + \lambda(x_j - x_i), \lambda \sim U(0,1) \tag{3}$$

, where x_i and x_j be two minority class examples, and λ a random scalar chosen uniformly from the set. After generation through synthesis, noisy examples belonging to the majority class are removed by the ENN algorithm, an approach that deals with missing data along with class imbalance for sensitive and accurate overall predictions in the model. As mentioned earlier, the task of missing or unbalanced dataset handling is very difficult in conventional approaches and it becomes complex when both the missing and balancing problems occur simultaneously, as shown in Figure 2(d). We have used “Yes” and “No”, to represent the stroke and non stroke events respectively. In Figure 2 ’0’ indicates missing data and ’1’ represents complete data.

In such conditions, the main problem that arises is to predict the stroke events moderately well and dependably. We therefore present a solution to this, with the full process shown in Figure 3. In this context, the primary challenge is accurately and reliably predicting strokes. To address this issue, we present a solution, as detailed in Figure 3.

III. Materials and Methods

A. Data Preprocessing

Classifier model performance can be negatively impacted by outliers and noise in datasets. Two widely used approaches exist for filtering out such anomalies: statistical and non-parametric methods [32]. Due to its straightforward implementation and reliance on established knowledge, we chose the statistical method to eliminate clear noise and outliers that contradict fundamental medical principles.

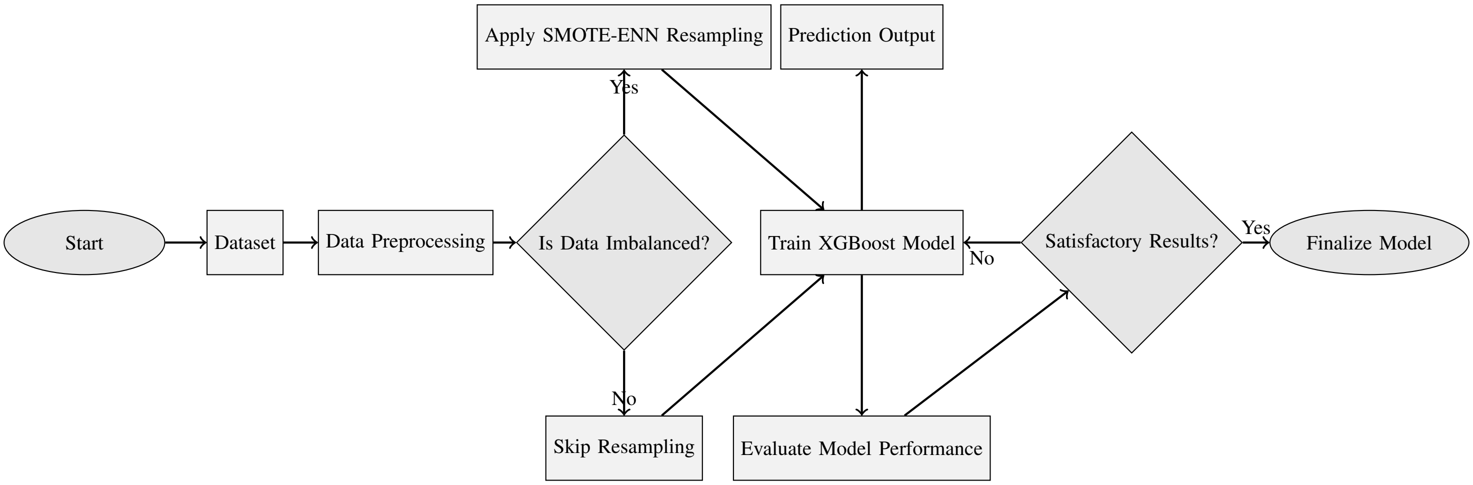


Fig. 3: Workflow diagram for the machine learning model training process

1) Standard Scaling and Splitting Features and Target: To normalize feature values and ensure that each feature contributes equally to the model, we applied the Standard Scaler during data preprocessing. This scaling method adjusts the mean of each feature to zero and scales it by the standard deviation, transforming the data into a standardized form. Standardizing features is particularly beneficial for algorithms like XGBoost, which perform better with normalized data. The dataset was divided into features (independent variables) and the target variable (stroke occurrence). This separation facilitates model training by enabling the model to learn patterns in the features that correlate with the target, ultimately improving prediction accuracy.

B. Imputation Methods for Missing Data

We evaluated several imputation techniques: Handling missing data effectively is critical in ensuring robust predictions. In this study, we evaluated several imputation techniques, considering the nature and distribution of missing values such as:

- **Mean Imputation:** Impute missing values using the mean of the observed data. Although straightforward, this method may introduce bias, particularly in imbalanced datasets.
- **K-Nearest Neighbors (KNN) Imputation:** Employ the K-nearest neighbors algorithm to impute missing values based on feature similarity, thereby preserving inter-feature relationships.
- **Multivariate Imputation by Chained Equations (MICE):** Iteratively impute missing values across all features, enabling advanced handling of missing data, albeit with increased computational demands.

In our model, we selected KNN imputation due to its balance between simplicity and effectiveness in preserving data relationships. KNN imputation leverages similarities within the dataset, which is beneficial for medical datasets where related health features often exhibit correlated patterns.

C. Assessment Metrics

1) Assessment Metrics for Class Imbalance: To manage class imbalance, the XGBModel leverages critical metrics derived from the confusion matrix, including true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). Key metrics such as sensitivity, specificity, and G-Mean are employed, with G-Mean being particularly vital as it reflects model performance on imbalanced datasets. A higher G-Mean denotes superior performance. The model further utilizes the ROC curve for visual evaluation of these metrics.

2) Assessment Metrics for Stroke Prediction: For stroke prediction, emphasis is placed on metrics crucial for medical diagnostics, specifically minimizing the false negative rate (FNrate) and false positive rate (FPrate), which are key to achieving accurate predictions. Although overall accuracy is considered, FNrate and FPrate are prioritized due to their significant influence on clinical outcomes.

D. Model Architecture: The Hybrid Prediction Model

The proposed model architecture incorporates SMOTE-ENN and XGBoost, with each component addressing distinct challenges inherent in stroke prediction on imbalanced datasets.

1) Hybrid Sampling with SMOTE-ENN: The Synthetic Minority Over-sampling Technique (SMOTE) combined with Edited Nearest Neighbors (ENN) offers a powerful approach to tackle data imbalance. SMOTE creates synthetic instances for the minority class through interpolation between existing samples, thereby enhancing the representation of stroke cases in the training dataset. Subsequently, ENN eliminates noisy or misclassified samples from the majority class, improving the quality of data used for training.

Mathematically, given a minority class sample $\mathbf{x_i}$, SMOTE generates synthetic instances by selecting a neighbor $\mathbf{x_{i_{\text{neighbor}}}}$ and creating a new point:

$$x_{\phi_i} = x_i + \delta \times (x_{i_{\text{neighbor}}} - x_i) \tag{4}$$

, where δ is a random number between 0 and 1. ENN subsequently refines the dataset by removing samples with differing class labels among their k-nearest neighbors.

2) XGBoost Classifier: The XGBoost classifier is an ensemble learning technique based on gradient-boosted decision trees. XGBoost builds trees sequentially, with each subsequent tree attempting to correct errors from the previous ones. Given its robustness and efficiency, XGBoost is well-suited for handling structured data and exhibits strong performance on imbalanced datasets.

For each observation i, XGBoost minimizes a regularized loss function L that includes a differentiable convex loss function l and a regularization term ω to penalize model complexity:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{k} \omega(f_k) \tag{5}$$

, where $\hat{y}_i$ is the predicted value, $y_i$is the actual value, and f_k represents the regularization of each function in the ensemble.

Our model further optimized the loss function by introducing a weight α for the minority class, accounting for the imbalance ratio in the dataset. This adjustment enhances the sensitivity of the model towards stroke cases by penalizing false negatives more heavily, ensuring that the model accurately identifies positive cases.

3) Cross-Validation and Model Evaluation: To enhance robustness and mitigate overfitting, a 10-fold cross-validation approach was employed. The dataset was divided into ten subsets, with nine subsets used for training and the remaining one for testing during each iteration. This process was repeated, and the results were averaged to yield stable and reliable performance metrics. The evaluation metrics employed include accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC-ROC). Sensitivity, in particular, is a key focus of the model, as it represents its ability to accurately detect stroke cases—a critical objective due to the severe implications of false negatives in medical settings.

IV. Experiment and Results

A. Dataset

The dataset used in this research originates from HealthData.gov and serves as a benchmark dataset for a Kaggle competition (as detailed in Table I). It presents a significant class imbalance, containing 43,400 samples with only 783 instances of stroke, representing 1.18% of the total. Additionally, the dataset includes missing values: 30% of entries lack smoking status information, and 3% have missing body mass index (BMI) data. Various preprocessing methods are applied to handle these missing values.

B. Experimental Setup

1) Prediction with XGBModel based on SMOTE-ENN Strategy: For an optimal learning scenario, the process of class-imbalance addressing under SMOTE-ENN encompasses both oversampling the minority class and undersampling the majority class, which indeed makes it a reduction in the number of the majority class data instances while concurrently increasing the instances in the minority class data instances for training.

Algorithm 1 Stroke Prediction Model Using SMOTE-ENN and XGBoost

```
Input: $X, y, k_s, T, \Theta_{XGB}$
Output: $M_{opt}, CM$
Data Preprocessing:
for each dataset $D$ do
    Fill missing values for BMI and smoking status
    Perform one-hot encoding for categorical variables
    Split into $X$ (features) and $y$ (target)
end for
Apply SMOTE-ENN:
for each data batch $(X, y)$ do
    $D_{new} \leftarrow$ Apply SMOTE with $k_s$ nearest neighbors
    Undersample majority class using ENN to create $D_{resampled}$
end for
Split $D_{resampled}$ into $(X_{train}, y_{train})$ and $(X_{test}, y_{test})$
for each hyperparameter configuration $\theta \in \Theta_{XGB}$ do
    Optimize $\theta$ using RandomizedSearchCV to find $\theta_{opt}$
end for
for $t = 1$ to $T$ do
    Train XGBoost on $X_{train}, y_{train}$ using $\theta_{opt}$
end for
$y_{pred} \leftarrow$ Use $M_{opt}$ to predict labels on $X_{test}$
$CM \leftarrow$ Calculate confusion matrix and metrics
Return $M_{opt}, CM$
```

Features	Values
Patient ID	1-43400
Gender(gen)	Male/Female
Residence type	Urban/Rural
Avg-glucose(glu)	55-291
Work type(work)	Private/Employed
Smoking status	Smoked/Formerly/Never
Hypertension(hyp)	Yes/No
Married(mar)	Yes/No
Heart disease(hd)	Yes/No
BMI	10.1-97.6
Age	0.08-82

TABLE I: Dataset description

Selecting instance within the model is another way of controlling the balancing ratio. This is basically done by sampling instances through a strategic selection; the model will not overselect and distort the original source distribution. The hyperparameter for this selection is trained based on the loss functions of the prediction, for which the sampling is obtained dynamically from a set that is defined, ensuring that the false negative should be at its minimum in such a way that generally the accuracy is above 70

Our model combines online learning with a modified approach of XGBoost to adjust weight coefficients using automatic differentiation, optimizing for both sensitivity and

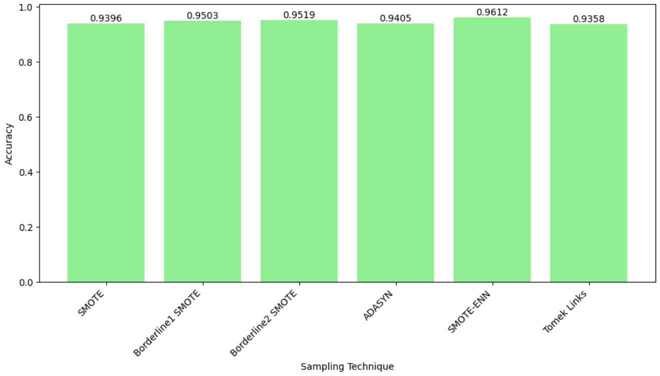


Fig. 4: Accuracy comparison for different Sampling Techniques

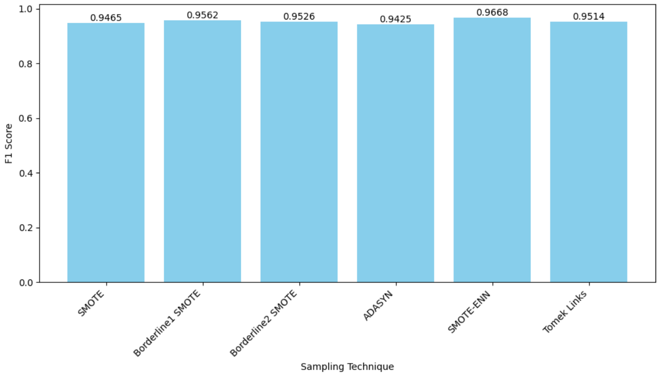


Fig. 5: F1 Score comparison for different Sampling Techniques

specificity. This is combined with dynamic weight calibration for the minority class to ensure that the ratio of post-sampling imbalances matches the target closely so that there will be minimal misclassification regarding stroke cases. With high stakes regarding medical misdiagnosis, the loss function focuses on sensitivity as its primary objective, along with specificity. We use 10-fold cross-validation to avoid overfitting, processing the full model prediction multiple times and averaging the results for robust, generalized outcomes.

Standard evaluation metrics as well as stroke-specific assessment metrics are used as a form of validation of the model, with key performance indicators of the prediction success being overall accuracy, sensitivity, and decrease in false negatives.

2) SMOTE-ENN Reason for Choice: Our comparative analysis of various sampling techniques revealed SMOTE-ENN as the optimal choice for addressing class imbalance in our dataset. As shown in Figure 5 and Figure 4, SMOTE-ENN achieved an F1-score of 0.9465 and accuracy of 0.9398, demonstrating robust performance across both metrics. While other techniques like SMOTE and Random SMOTE showed comparable results, SMOTE-ENN's balanced performance in both F1-score and accuracy suggests it effectively handles both majority and minority classes without overfitting. The two-step nature of the technique, with first SMOTE oversampling the minority class and then noisy samples being cleaned using ENN Edited Nearest Neighbors, creates cleaner decision boundaries but maintains the integrity of the original data distribution. Such a balanced approach is more critical in medical applications, where false positives and false negatives have major implications.

C. Computational Results and Analysis

The analysis of the correlation coefficient table (Table II) provides additional insights:

BMI has the strongest correlations with age (0.110), hypertension (0.130), marital status (0.150), and glucose levels (0.180). Smoking has the strongest correlation with gender (0.091) and a weaker correlation with other variables.

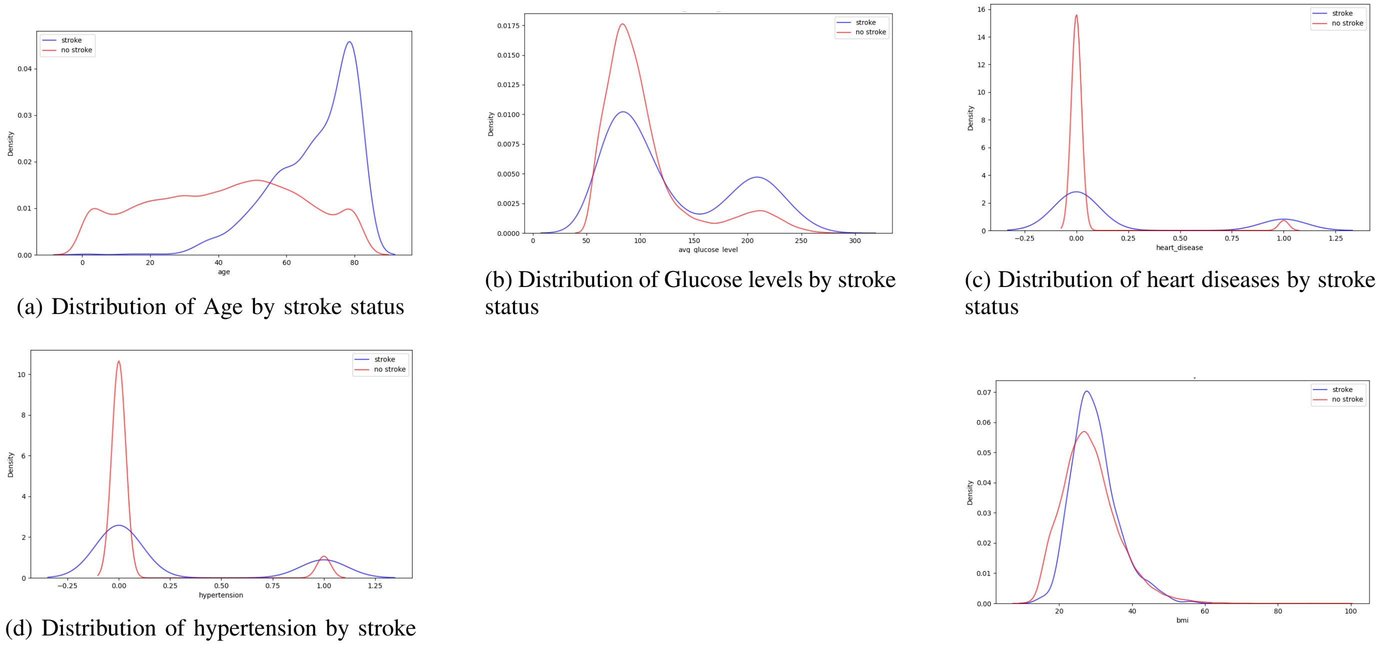
The comprehensive distribution analysis and correlation coefficients provide valuable insights for developing targeted screening protocols and personalized preventive strategies, particularly for aging populations with concurrent cardiovascular and metabolic disorders.

	gen	age	hyp	mar	glu
bmi	0.023	0.110	0.130	0.150	0.180
smoke	0.091	0.061	0.008	0.070	0.026

TABLE II: Correlation coefficients

1) Results of Data Preprocessing: This section demonstrates the application of our stroke prediction model to the dataset and validates the missing data handling techniques described in Section 3. The dataset includes numerous outliers and noisy entries, particularly in the age and BMI features. As per established guidelines, including those from the MONICA project [**?**], stroke monitoring generally targets individuals aged 25 and older. However, the dataset contains samples with ages below this threshold, including infants as young as 0.08 years.Moreover, BMI values typically fall between 10% and 50%, yet the dataset includes instances with BMI values ranging from 60% to 97.6%.To ensure dataset quality, outliers were removed, specifically those with ages below 25 and BMI values exceeding 60%. Filter-based methods were applied for feature selection, enabling the removal of irrelevant and redundant attributes to enhance model performance. For instance, features such as Residence type, which display comparable distributions across stroke and non-stroke cases, and Patient ID, which serves only as an identifier, were excluded from the analysis.This preprocessing step ensures a cleaner dataset that aligns more closely with the model's emphasis on key predictive attributes.

In the preprocessing stage of our model, we handled missing values in key features like smoking-status and BMI to ensure the completeness of data before model training. Since missing data can introduce bias and reduce model accuracy, particularly for imbalanced datasets, it is crucial to implement robust imputation methods tailored to the characteristics of each feature.



(a) Distribution of Age by stroke status

(b) Distribution of Glucose levels by stroke status

(c) Distribution of heart diseases by stroke status

(d) Distribution of hypertension by stroke status

(e) Distribution of BMI by stroke status

Fig. 6: Comparative Analysis of Clinical Variables

The correlation matrix reveals notable differences in relationship strengths across variables. BMI demonstrates relatively stronger correlations with other features, particularly with glucose levels (0.18) and marital status (0.15). In contrast, smoking status exhibits consistently weaker correlations across all features, with correlation coefficients mostly below 0.10. This suggests that while BMI has meaningful associations with several variables in the dataset, smoking status appears to be largely independent of the other measured characteristics.

- For BMI, missing values were imputed using the median value, as it is less sensitive to extreme values compared to the mean. This choice also aligns well with the dataset's structure, where outliers in BMI may distort the average but have less impact on the median.
- For smoking-status, which is a categorical variable with around 30% missing data, we employed a more targeted approach using a combination of the mode imputation for cases with frequent categories and predictive modeling for infrequent categories. The model utilizes surrounding demographic features, such as age, work-type, and Residence-type, to estimate plausible values for smoking-status. This imputation strategy ensures that missing data is handled in a way that maintains the dataset's original distribution and reduces the impact of imputation on model performance.
- This targeted approach to imputation not only preserves the distribution of these variables but also helps maintain the dataset's integrity, providing a more reliable basis for prediction and helping to reduce the effects of class imbalance during model training.

2) Comparative Analysis of Clinical Variables in Stroke Risk Assessment: The comprehensive analysis of health parameters between stroke and non-stroke populations reveals distinct patterns of risk factors. Age is the most significant predictor, with stroke incidence peaking around age 80 (Fig. 6(a)). Cardiovascular indicators, particularly heart disease (Fig. 6(c)) and hypertension (Fig. 6(d)), demonstrate a strong association with stroke occurrence. Blood glucose levels (Fig. 6(b)) show a bimodal distribution pattern, with stroke patients exhibiting a more pronounced peak in the diabetic range. BMI distributions (Fig. 6(e)) show the least distinctive pattern, suggesting BMI alone may not be a reliable independent predictor of stroke risk.

3) Model and Sampling Technique Comparison: To identify the most effective combination of model and sampling technique for predicting cerebral stroke, we evaluated several machine learning algorithms, each paired with various resampling methods to address class imbalance. Table III presents the performance metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, false negative rate (FNR), and false positive rate (FPR) for each combination.

From these results, the XGBoost model coupled with SMOTE-ENN sampling yielded the highest balance between sensitivity (86.02%) and specificity (99.05%) and achieved a strong overall accuracy of 98.12%. This model also demonstrated a low FNR of 1.08%, indicating it missed fewer actual stroke cases compared to other configurations. High sensitivity is crucial in medical predictions where the cost of a false negative—failing to identify a stroke patient—could have severe implications. SMOTE-ENN, as a hybrid resampling technique, was particularly effective in refining class bound-

Sampling Technique	Model	Accuracy (%)	Precision	Recall (Sensitivity)	Specificity	F1-Score	FNR (%)	FPR (%)
SMOTE	KNN	95.86	0.79	74.33	98.01	0.77	2.55	21.10
	Naive Bayes	35.57	0.12	99.77	29.15	0.22	0.08	87.66
	SVM	86.39	0.39	85.05	86.52	0.53	1.70	61.32
	XGBoost	97.26	0.88	81.38	98.85	0.84	1.85	12.38
Borderline1 SMOTE	KNN	96.23	0.78	80.99	97.76	0.80	1.91	21.71
	Naive Bayes	39.78	0.13	99.61	33.80	0.23	0.12	86.93
	SVM	89.46	0.46	88.34	89.57	0.60	1.28	54.16
	XGBoost	97.61	0.88	85.29	98.84	0.87	1.47	11.95
Borderline2 SMOTE	KNN	96.24	0.78	82.16	97.65	0.80	1.79	22.28
	Naive Bayes	38.03	0.13	99.84	31.85	0.23	0.05	87.23
	SVM	87.78	0.42	87.17	87.85	0.56	1.44	58.25
	XGBoost	97.60	0.86	87.95	98.56	0.87	1.21	14.07
ADASYN	KNN	95.73	0.79	74.31	97.97	0.77	2.67	20.77
	Naive Bayes	37.04	0.13	99.25	30.55	0.23	0.26	87.02
	SVM	85.99	0.39	83.82	86.22	0.53	1.92	61.16
	XGBoost	97.32	0.89	82.17	98.91	0.85	1.85	11.32
SMOTE-ENN	KNN	97.10	0.82	76.59	98.68	0.79	1.80	18.20
	Naive Bayes	41.30	0.11	99.89	36.77	0.20	0.02	89.12
	SVM	89.40	0.39	86.14	89.65	0.54	1.18	60.85
	XGBoost	**98.12**	**0.88**	**86.02**	**99.05**	**0.87**	**1.08**	**12.49**
Tomek Links	KNN	98.00	0.05	1.70	99.44	0.03	2.07	22.95
	Naive Bayes	37.12	0.11	96.37	28.14	0.18	1.89	88.12
	SVM	88.32	0.41	87.39	89.13	0.56	1.56	56.12
	XGBoost	98.45	0.90	87.62	99.32	0.88	1.23	11.17

TABLE III: Model Performance with Different Sampling Techniques

aries by not only oversampling minority instances (through SMOTE) but also removing ambiguous samples that could add noise (using Edited Nearest Neighbors), which further enhanced the model's robustness against class imbalance.

Although models like XGBoost with Borderline1 and Tomek Links sampling techniques produced high accuracies of 97.61% and 98.45% respectively, they exhibited lower sensitivity and a slightly higher FNR than SMOTE-ENN. In our application, sensitivity takes precedence, and SMOTE-ENN better optimized the balance between capturing positive stroke cases while maintaining low error rates.

The superior performance of the XGBoost with SMOTE-ENN configuration led us to select this model for final implementation, as it effectively maximized sensitivity without sacrificing overall model stability or introducing excessive false positives. This makes it a reliable choice for cerebral stroke prediction on an imbalanced medical dataset.

4) Analysis of Stroke Prediction Results: In stroke prediction, minimizing false negatives is paramount, as failing to identify potential stroke cases can lead to delayed intervention and permanent patient harm. While false positives may cause unnecessary patient anxiety, the primary focus remains on accurate stroke risk identification. Our study aimed to optimize this critical balance. The XGBoost model enhanced with SMOTE-ENN demonstrated superior performance, achieving an AUC score of 0.99 (Figure 7) on the ROC curve. This exceptional score indicates the model's strong ability to distinguish between stroke and non-stroke cases, surpassing the performance of alternative approaches including KNN, Naive Bayes, SVM, and Neural Network models. Our approach addresses a significant limitation of conventional methods: their tendency toward higher false negative rates. The model's balanced performance in sensitivity and specificity suggests its potential value as a clinical decision support tool. Future work should explore deep neural network feature analysis to potentially uncover additional predictive patterns and enhance clinical applicability. Table III provides a performance comparison of various models in terms of False Negative (FN) Rate, False Positive (FP) Rate, Accuracy, Specificity, Sensitivity, and G-Mean. The XGBoost model outperforms others with high accuracy (98.12%), specificity (99.05%), and G-Mean (92.31%), making it effective in handling class imbalance while maintaining both high sensitivity and specificity. KNN with Tomek Links, while included, does not have available metrics for FN, FP, Sensitivity, or Specificity.

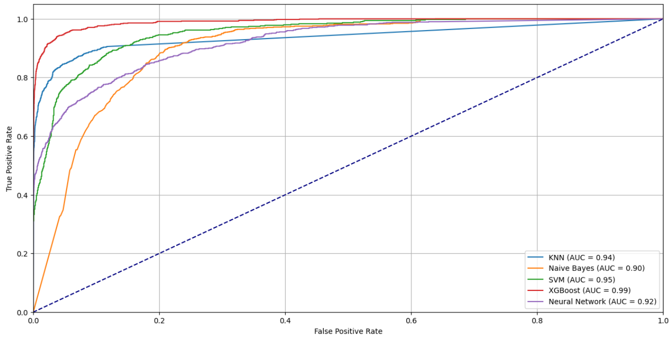


Fig. 7: Receiver Operating Characteristic (Curve)

V. Conclusion

This study presents a comprehensive approach to optimizing cerebral stroke prediction through a hybrid machine learning framework combining SMOTE-ENN with XGBoost. The proposed XGBModel demonstrated strong performance, achieving an accuracy of 98.14%, a false negative rate of 12.61%, and a sensitivity of 87.39%.

An extensive comparative analysis was conducted involving multiple machine learning models, such as K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), XGBoost, and Neural Networks. These models were evaluated under different sampling techniques, specifically employing SMOTE for oversampling and ENN for undersampling, to address the dataset's intrinsic class imbalance. The analysis underscored both the strengths and limitations of each algorithm while showcasing the effectiveness of the proposed approach.

The performance of the models was further assessed using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), a critical evaluation metric. The AUC-ROC analysis revealed that the XGBModel achieved a significantly higher AUC compared to other models, indicating its superior capability to distinguish between positive and negative classes. This distinction is particularly critical in medical predictions, given the high costs associated with false negatives.

Incorporating instance selection methods and a modified loss function designed for optimizing the minority class is vital in improving the generalization capability of the model. Our validation process involved robust verification through 10-fold cross-validation and repeated experiments to ensure solid stability and reliability of results. Our overall findings, therefore, are that advanced machine learning techniques can improve accuracy in stroke prediction, which must eventually result in timely medical interventions for patients. The future would focus on further refinement of the model, exploring applicability in real-world clinical settings, and investigating additional features that could contribute to even better predictive performance.

Declaration of Competing Interest

We declare that we have no commercial or associative interests that could present a conflict of interest regarding the submitted work.

References

[1] V. L. Feigin, B. Norrving, and G. A. Mensah, "Global burden of stroke," Circ Res, vol. 120, no. 3, pp. 439-448, 2017.

[2] M. Naghavi et al., "Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the global burden of disease study 2016," Lancet, vol. 390, no. 10100, pp. 1151-1210, 2017.

[3] A. Algra and M. J. Wermer, "Stroke in 2016: Stroke is treatable but prevention is the key," Nat Rev Neurol, vol. 13, no. 2, pp. 78, 2017.

[4] M. J. O'Donnell et al., "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (interstroke): a case-control study," Lancet, vol. 388, no. 10046, pp. 761-775, 2016.

[5] I. Yoo et al., "Data mining in healthcare and biomedicine: a survey of the literature," J Med Syst, vol. 36, no. 4, pp. 2431-2448, 2012.

[6] A. N. Richter and T. M. Khoshgoftaar, "A review of statistical and machine learning methods for modeling cancer risk using structured clinical data," Artif Intell Med, vol. 90, pp. 1-14, 2018.

[7] C. R. Pereira et al., "A survey on computer-assisted parkinson's disease diagnosis," Artif Intell Med, vol. 95, pp. 48-63, 2018.

[8] A. Kaya, "Cascaded classifiers and stacking methods for classification of pulmonary nodule characteristics," Comput Methods Programs Biomed, vol. 166, pp. 77-89, 2018.

[9] O. R. Shishvan, D.-S. Zois, and T. Soyata, "Machine intelligence in healthcare and medical cyber physical systems: A survey," IEEE Access, vol. 6, pp. 46419-46494, 2018.

[10] C. Colak, E. Karaman, and M. G. Turtay, "Application of knowledge discovery process on the prediction of stroke," Comput Methods Programs Biomed, vol. 119, no. 3, pp. 181-185, 2015.

[11] N. Kasabov et al., "Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke," Neurocomputing, vol. 134, pp. 269-279, 2014.

[12] A. K. Arslan, C. Colak, and M. E. Sarihan, "Different medical data mining approaches based prediction of ischemic stroke," Comput Methods Programs Biomed, vol. 130, pp. 87-92, 2016.

[13] J. M. Jerez et al., "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," Artif Intell Med, vol. 50, no. 2, pp. 105-115, 2010.

[14] T.-P. Hong and C.-W. Wu, "Mining rules from an incomplete dataset with a high missing rate," Expert Syst Appl, vol. 38, no. 4, pp. 3931-3936, 2011

[15] X. Zhang, S. Song, and C. Wu, "Robust bayesian classification with incomplete data," Cogn Comput, vol. 5, no. 2, pp. 170-187, 2013

[16] G. Haixiang et al., "Learning from class-imbalanced data: Review of methods and applications," Expert Syst Appl, vol. 73, pp. 220-239, 2017

[17] N. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," SIGKDD Explor, vol. 6, pp. 1-6, 2004.

[18] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explor Newsletter, vol. 6, no. 1, pp. 20-29, 2004.

[19] Y.-M. Chyi, "Classification analysis techniques for skewed class distribution problems," Department of Information Management, National Sun Yat-Sen University.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J Artif Intell Res, vol. 16, pp. 321-357, 2002.

[21] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," Inf Sci, vol. 409, pp. 17-26, 2017.

[22] Y. Wang, L. Yang, and Q. Ren, "A robust classification framework with mixture correntropy," Inf Sci, vol. 491, pp. 306-318, 2019.

[23] L. Yang and H. Dong, "Robust support vector machine with generalized quantile loss for classification and regression," Appl Soft Comput, vol. 81, pp. 105483, 2019.

[24] A. Ghazikhani, R. Monsefi, and H. S. Yazdi, "Online cost-sensitive neural network classifiers for non-stationary and imbalanced data streams," Neural Comput Appl, vol. 23, no. 5, pp. 1283-1295, 2013.

[25] H. Yu et al., "Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data," Knowl-Based Syst, vol. 76, pp. 67-78, 2015.

[26] Y. Wang and L. Yang, "A robust loss function for classification with imbalanced datasets," Neurocomputing, vol. 331, pp. 40-49, 2019.

[27] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Comput. Soc. Conf. Comput., 2017, pp. 2980-2988.

[28] E. C. Ozan, E. Riabchenko, S. Kiranyaz, and M. Gabbouj, "An optimized k-nn approach for classification on imbalanced datasets with missing data," in Int. Symposium on Intelligent Data Analysis, Springer, 2016, pp. 387-392.

[29] S. Liu, J. Zhang, Y. Xiang, W. Zhou, "Fuzzy-based information decomposition for incomplete and imbalanced data learning," IEEE Trans Fuzzy Syst, vol. 25, no. 6, pp. 1476-1490, 2017.

[30] C. A. Leke and T. Marwala, "Introduction to missing data estimation," in Deep Learning and Missing Data in Engineering Systems, Springer, 2019, pp. 1-20.

[31] I. Ben-Gal, "Outlier detection," in Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 131-146.

[32] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: bagging and random forests for ecological prediction," Ecosystems, vol. 9, no. 2, pp. 181-199, 2006.

[33] M. Feurer et al., "Efficient and robust automated machine learning," Adv Neural Inf Process Syst, 2015, pp. 2962-2970.

[34] C. Ding and X. He, "K-means clustering via principal component analysis," in Proc. Twenty-First Int. Conf. Machine Learning, ACM, 2004, p. 29.

[35] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in k-means clustering," International J, vol. 1, no. 6, pp. 90-95, 2013.

[36] C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao, “Under-sampling class imbalanced datasets by combining clustering analysis and instance selection,” Inf Sci, vol. 477, pp. 47-54, 2019.

[37] R. Raj, P. Raut, M. K. Zope, J. Mathew, S. K. Kannath, and J. Rajan, “Resident Vision Transformer: Lightweight Deep Learning Model for Disease Diagnosis on Edge Devices,” in *2024 10th International Conference on Smart Computing and Communication (ICSCC 2024)*, pp. 349-355, 2024.

[38] R. Raj, J. Mathew, S. K. Kannath, and J. Rajan, “StrokeViT with AutoML for brain stroke classification,” Engineering Applications of Artificial Intelligence, vol. 119, p. 105772, 2023.

[39] M. Liu et al., “Stroke in china: epidemiology, prevention, and management strategies,” Lancet Neurol, vol. 6, no. 5, pp. 456-464, 2007.