International Research Journal of Medical Sciences and Health Care (IRJMSHC)
eISSN: 3087-405X
Volume03 Issue08 | Page No: - 23-33

Received: 30 June 2026, Revised: 15 July 2026, Accepted: 28 July 2026, Published: 15 August 2026

DOI: https://doi.org/10.55640/irjmshc-v03i08-05

Comparative Study of Machine Learning Models for Stroke Risk Prediction

Dr. Pradeep Laxkar [1]

[1] Associate Professor, Department of Computer Science and Engineering, ITM (SLS) University, Vadodara, Gujarat, India

ABSTRACT: A stroke is a chronic condition that occurs more and more every year and can be caused by the sudden stop of blood from reaching a part of the brain, causing serious neurological damage and a higher death rate if not caught on time. Early symptom detection offers insights that are important for stroke risk prediction and prompt clinical response for better stroke outcomes. This research proposes an effective machine learning-based framework for stroke risk prediction using the Kaggle Stroke Prediction Dataset. The approach involves thorough data preprocessing, such as filling missing values, label encoding, Min–Max normalization, and SMOTE–ENN data balancing, followed by the application of Random Forest (RF) and Extreme Gradient Boosting (XGBoost) classifiers. The experimental results demonstrate exceptional predictive performance, where the proposed RF model achieved 99.99% accuracy, 99.97% precision, 99.98% recall, and 99.96% F1-score, while the proposed XGBoost model attained 99.92% accuracy, 99.91% precision, 99.90% recall, and 99.93% F1-score. These findings indicate that ensemble machine learning techniques significantly improve stroke risk prediction.

KEYWORDS: Stroke risk prediction, Machine learning, Healthcare decision support, SMOTE, Preventive healthcare.

1. Introduction

The integration of modern technologies in healthcare and medicine has significantly transformed disease prediction, monitoring, and personalized treatment strategies. Stroke is one of the most critical neurological disorders and remains a major global public health concern due to its high mortality and long-term disability rates [1][2]. It occurs when the blood supply to the brain is interrupted either by a blockage in blood vessels (ischemic stroke) or by bleeding within the brain (hemorrhagic stroke) [3], resulting in the sudden loss of brain function. According to the World Health Organization (WHO), stroke is among the leading causes of death and disability worldwide, creating a significant burden on individuals, families, and healthcare systems [4]. Recent global statistics indicate that millions of people experience stroke annually, with a considerable proportion of survivors suffering from permanent physical and cognitive impairments. The increasing prevalence of stroke highlights the urgent need for effective risk assessment, early diagnosis, and preventive strategies [5][6].

The occurrence of stroke is influenced by multiple interacting risk factors, including age, hypertension, diabetes, obesity, high blood glucose levels, abnormal cholesterol levels, smoking habits, cardiovascular diseases, and lifestyle-related factors [7]. Among these, hypertension is considered one of the most significant modifiable risk factors associated with stroke development [8][9]. The complex relationship between genetic, environmental, clinical, and behavioral factors makes stroke prediction a challenging task [10][11]. Conventional diagnostic approaches, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), are widely used for detecting brain abnormalities [12][13][14]; however, these techniques are generally applied after symptoms appear and may have limitations in identifying individuals at high risk before stroke occurrence. Therefore, accurate predictive approaches are required to support early intervention and reduce stroke-related complications [15][16][17].

Traditional statistical methods often struggle to capture complex nonlinear relationships among multiple stroke-related factors. With advances in healthcare data analytics, machine learning (ML) techniques have emerged as effective tools for analyzing large-scale clinical datasets and identifying hidden patterns associated with disease risk [18][19]. ML algorithms can integrate diverse patient characteristics, including

demographic information, medical history, laboratory results, and lifestyle indicators, to improve stroke risk prediction accuracy. Various machine learning models have demonstrated promising capabilities in handling multidimensional healthcare data [20]; however, challenges such as class imbalance, feature variability, model optimization, and generalization remain major concerns in developing reliable stroke prediction systems [21][22][23][24].

A. Motivation and Contribution

The motivation behind this study is to develop an accurate and reliable machine learning-based framework for early stroke risk prediction to reduce mortality and long-term disabilities. Existing prediction approaches face challenges in handling complex clinical patterns, imbalanced datasets, and improving classification reliability. Therefore, this research focuses on applying advanced preprocessing techniques and ensemble learning models to enhance predictive performance. The proposed approach aims to support timely decision-making and assist healthcare professionals in effective stroke prevention and management. This research offers several key contributions as listed below:

- Developed an effective machine learning-based framework for accurate stroke risk prediction using the Kaggle Stroke Prediction Dataset.
- Performed comprehensive data preprocessing, including missing value handling, feature encoding, normalization, and outlier treatment to improve data quality.
- Applied SMOTE-ENN data balancing technique to overcome severe class imbalance and enhance minority stroke class representation.
- Implemented and evaluated advanced ensemble learning models, including the proposed Random Forest (RF) and XGBoost classifiers, for stroke classification.
- Conducted a comparative analysis with existing machine learning models such as Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting Machine (GBM), and K-Nearest Neighbor (KNN).
- Evaluated model effectiveness using multiple performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC analysis.
- Proposed a reliable decision-support approach that can assist healthcare professionals in early stroke risk identification and preventive clinical decision-making.

This study is justified by the need for an accurate and reliable stroke risk prediction system capable of addressing class imbalance and improving early detection. The novelty lies in integrating SMOTE-ENN-based data balancing with advanced ensemble learning models, namely Random Forest and XGBoost, to enhance predictive performance. The proposed framework effectively captures complex clinical patterns and achieves superior classification results compared with conventional ML approaches. This approach provides a robust decision-support solution for timely stroke prediction and preventive healthcare management.

Organization of the Paper

The paper is organized as follows: Section II reviews existing stroke risk prediction studies, Section III presents the dataset, preprocessing, and proposed models, Section IV discusses results and comparisons, Section V concludes the study and outlines future research directions.

2. Literature Review

The literature review examines recent machine learning and deep learning approaches for stroke risk prediction, emphasizing class imbalance, preprocessing, model performance, and explainability. Table I summarizes the reviewed studies, datasets, key findings, limitations, and potential directions for future research.

A. G (2026) Machine learning is increasingly used for automated stroke risk estimation using electronic health records, but severe class imbalance can cause misleadingly high accuracy and poor stroke detection. The study proposes an imbalance-aware framework using normalization, SMOTE, and cost-sensitive learning, with precision, recall, and F1-score for evaluation. Results show that imbalance control improves stroke detection, with Logistic Regression achieving the highest recall (0.80) and F1-score (0.2469), making it the most effective for identifying high-risk patients despite moderate accuracy [25].

S. Lila et al. (2026) compared DL and traditional ML models after imputation, encoding, z-score normalization, stratified 5-fold cross-validation, and SMOTE on the training set. Five ML models (RF, LR, KNN, SVM, DT) and three DL models (MLP, CNN, ANN) were evaluated using F1-score due to class imbalance. LR achieved the best performance (30% F1-score, 88% accuracy), followed by MLP (28% F1-score, 84% accuracy), indicating that classical ML models can be competitive on structured tabular data [26].

S. N. Pasha, R. Dadi, and K. Vennela (2025) discuss in detail some of the various machine learning and neural

network models applied in predicting the risk of brain stroke. Brain stroke remains one of the major global health concerns because of its ranking among various causes of mortality and disability. The ability to predict the risk of stroke before symptoms occur would save lives, reduce complications during treatment, and enhance the quality of life for patients if interventions could be made earlier. They explore the predictive capability of four models—Random Forest, Support Vector Machine (SVM), XGBoost, and Artificial Neural Networks (ANN)—trained on a dataset of demographic, clinical, and lifestyle variables. Results show that ensemble methods, Random Forest and XGBoost, yield the best accuracy scores of 94% and 93%, respectively [27].

M. Z. Hossain Zamil,et al. (2025) an efficient machine learning system for stroke risk prediction using clinical and demographic information. The dataset had extreme class imbalance, with strokes occurring in only 4.87% of patients. SMOTE was used to achieve a 50%-50% distribution. Four algorithms—Logistic Regression, Random Forest, XGBoost, and Cat Boost—were compared. CatBoost achieved the best performance, with 95.58% accuracy, 96.82% precision, 94.23% recall, and 95.51% F1-score. Explainable machine learning was employed to improve model interpretation and clinical decision support. The proposed model outperformed baseline methods and demonstrated potential for assisting clinicians in stroke risk assessment [28].

M. Sharma, et al. (2024) present a predictive modeling approach for assessing stroke risk using machine learning algorithms. The study compares several models using a large dataset of patient attributes, including data exploration, preprocessing, model creation, and performance evaluation. Logistic regression achieved the highest accuracy of 94.6%. The study also discusses model interpretability for clinical decision-making and highlights how machine learning can support early intervention, reduce healthcare burdens, and improve patient outcomes [29].

M. C. Das et al. (2023) The majority of strokes are triggered by blocked pathways between the heart and brain and remain a major cause of death worldwide. Several risk factors associated with stroke have been identified, leading to numerous studies on its prediction and classification. Early diagnosis can help prevent such occurrences and support a healthy life. This study used nine machine learning algorithms, including SVM, KNN, XGBoost, AdaBoost, Random Forest, Decision Tree, Light GBM, and Logistic Regression, to evaluate accuracy, recall, ROC, precision, and F1-score. The results showed that Random Forest outperformed the other methods with an accuracy of 98.4% [30].

R. M. D. Saputra et al. (2023) Absence of blood supply prevents the brain from receiving oxygen and nutrients, causing brain cells to perish and impairing bodily functions regulated by the affected regions. It is a leading cause of death and disability globally. While most research focuses on predicting heart attacks, relatively little addresses brain stroke risk. Therefore, this study developed machine learning models using physiological factors and algorithms including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Naive Bayes, and Support Vector Machine. The study shows that Random Forest achieved the best performance, with an accuracy of 82.07% [31].

TABLE I. A Review of Recent Machine Learning Approaches for Stroke Risk Prediction

Author	Proposed Work	Dataset	Findings	Limitations & Future Work
A. G, (2026)	Proposed an imbalance-aware machine learning framework for stroke risk prediction using structured clinical data with preprocessing, normalization, SMOTE, and cost-sensitive learning.	Structured clinical stroke dataset	Logistic Regression achieved the highest recall (0.80) and F1-score (0.2469), demonstrating improved minority stroke detection.	Limited by class imbalance challenges; future work includes improved imbalance handling and integration of advanced prediction techniques.
S. Lila et al., (2026)	Compared traditional ML and deep learning models for stroke prediction using preprocessing, stratified cross-validation, and SMOTE-based balancing.	Kaggle Stroke Prediction Dataset	Logistic Regression achieved 88% accuracy with 30% F1-score, while MLP achieved 84% accuracy with 28% F1-score.	Severe class imbalance affected performance; future research requires improved feature engineering and hybrid models.
S. N. Pasha, R. Dadi, and K. Vennela, (2025)	Developed and compared RF, SVM, XGBoost, and ANN models for stroke risk prediction using demographic, clinical, and lifestyle features.	Clinical and lifestyle stroke dataset	Random Forest achieved 94% accuracy and XGBoost achieved 93% accuracy.	Requires larger datasets including genetic and imaging information for improved prediction capability.
M. Z. Hossain	Developed an efficient ML-based stroke prediction system using	Clinical and demographic	CatBoost achieved 95.58% accuracy,	Future work includes real-time deployment and

Zamil,et al., (2025)	SMOTE balancing and explainable machine learning approaches.	stroke dataset	96.82% precision, 94.23% recall, and 95.51% F1-score.	advanced explainability methods for clinical applications.
M. Sharma, et al., (2024)	Proposed predictive modeling for stroke risk assessment using multiple machine learning algorithms with data exploration and evaluation.	Large-scale patient attribute dataset	Logistic Regression achieved maximum accuracy of 94.6%.	Limited model complexity; future studies should focus on advanced ML algorithms and clinical integration.
M. C. Das et al., (2023)	Evaluated nine ML algorithms including SVM, KNN, XGBoost, AdaBoost, RF, DT, LightGBM, and LR for stroke classification.	Stroke risk dataset	Random Forest outperformed other models with 98.4% accuracy.	Requires improvement in generalization and validation using larger healthcare datasets.
R. M. D. Saputra et al., (2023)	Developed multiple ML classifiers including LR, DT, RF, KNN, NB, and SVM for brain stroke prediction.	Physiological and clinical factor dataset	Random Forest achieved the highest accuracy of 82.07%.	Limited predictive performance; future work should explore optimization and advanced ensemble models.

Research gaps: Previous research in the field of stroke risk prediction has shown the potential of machine learning models, yet issues of class imbalance, minority class detection and model generalization remain unresolved. Existing approaches are mostly based on accuracy-based evaluation that might not fully represent the actual predictive power for stroke patients. In addition, the balancing of effective models in ensemble learning has yet to be studied in detail. This study addresses these gaps by integrating SMOTE-ENN balancing with advanced Random Forest and XGBoost models to achieve more reliable and accurate stroke risk prediction.

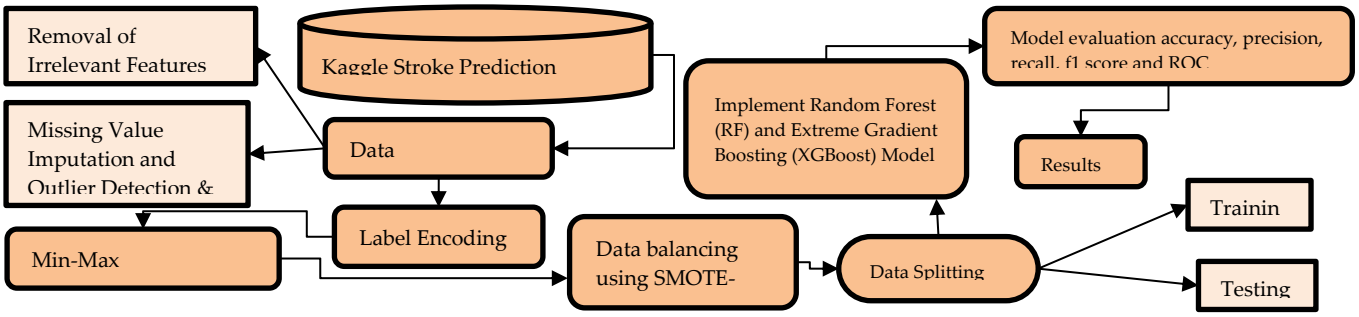

Fig.1. Flow Diagram of the Proposed Machine Learning Approach for Stroke Risk Assessment Models for Stroke Risk Prediction.

3. Methodology

The proposed methodology for stroke risk prediction utilizes the Kaggle Stroke Prediction Dataset, followed by comprehensive data preprocessing, including missing value treatment, label encoding, Min–Max normalization, and SMOTE-ENN balancing to improve data quality and address class imbalance. The preprocessed dataset is divided into training and testing sets using an 80:20 stratified split, and Random Forest (RF) and XGBoost models are trained to classify stroke risk effectively. Finally, the models are evaluated using a confusion matrix, accuracy, precision, recall, F1-score, and ROC curve to assess their classification performance and predictive reliability. Figure 1 illustrates the proposed flowchart for Machine Learning

The following section presents a detailed description of the individual stages involved in the proposed methodology:

A. Dataset Collection and Exploratory Analysis

This study employs the Kaggle Stroke Prediction Dataset to develop and evaluate a machine learning model for stroke risk prediction. The dataset comprises 43,400 instances with 12 features, providing comprehensive demographic and clinical information for analysis. Exploratory data analysis was performed using visualizations such as bar plots and correlation heatmaps to examine data distribution, feature relationships, and underlying patterns, as presented below:

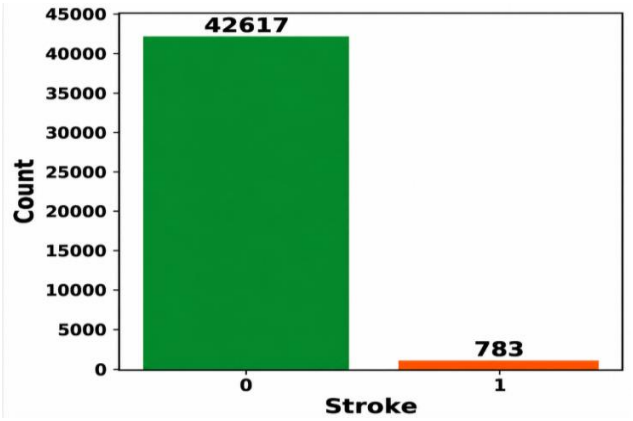

Fig. 2. Bar graph of class distribution

Figure 2 shows the bar chart of the original class distribution of the stroke dataset with class 1 (stroke) having 783 and class 0 (non-stroke) having 42,617. This is a very imbalanced dataset. This significant class imbalance highlights the need for data balancing techniques to improve model training and ensure accurate prediction of the minority stroke cases.

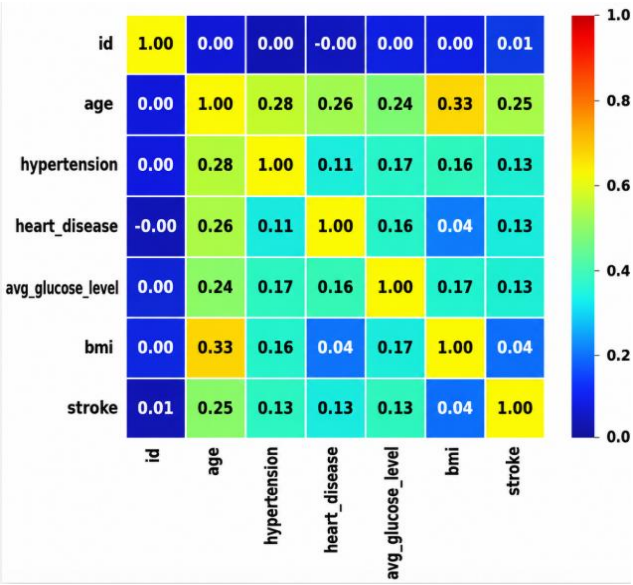

Fig. 3. Correlation Matrix Heatmap

Figure 3 presents the correlation matrix heatmap illustrates the relationships among the selected stroke prediction features, where lighter colors represent stronger positive correlations and darker colors indicate weaker correlations. The results demonstrate that age has the highest positive correlation with BMI (0.33) and moderate correlation with hypertension (0.28), heart disease (0.26) and stroke (0.25); the other feature pairs have weak correlations, suggesting low multicollinearity in the dataset.

B. Data Pre-Processing

The Kaggle Stroke Prediction Dataset was utilized for data preparation, involving data integration, cleaning, and feature engineering processes. The preprocessing phase included the removal of irrelevant attributes, handling of missing values, detection and treatment of outliers, followed by label encoding and data normalization. The key preprocessing steps applied to enhance data quality and model performance are summarized as follows:

Removal of Irrelevant Features: The id attribute is removed from the dataset as it does not contribute significantly to stroke prediction outcomes.

Missing Value Treatment: Missing values present in the bmi and smoking status attributes are identified and replaced using appropriate imputation techniques.

Outlier Detection and Handling: Outlier Detection and Handling: Outliers in numerical features are analyzed and treated to reduce their impact on model performance.

Label Encoding: Label encoding is applied to transform categorical attributes into numerical representations, making them suitable for machine learning algorithms. Features such as gender, ever married, work type, residence type, and smoking status are encoded into numerical values for efficient model processing.

C. Min-Max Normalization

The normalization of records was carried out using the min–max technique to confine values within a range between 0 and 1. This was performed with the aim of optimizing the performance of the classifiers used and mitigating the effect of outliers. Normalization was conducted according to the following mathematical formula (1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X represents the original value of the feature, X' is the normalized value, X_min is the minimum value of the feature, and X_max is the maximum value of the same.

D. Data Balancing Using SMOTE-ENN

Data balancing is an essential preprocessing step used to overcome the problem of unequal class distribution, where one class contains fewer samples compared to another. In this study, the imbalance in the stroke prediction dataset is addressed using Random Oversampling/SMOTE-ENN techniques to increase minority class representation and achieve a more balanced class distribution. This process reduces model bias toward the majority class, improves learning efficiency, and enhances the overall predictive performance of machine learning models.

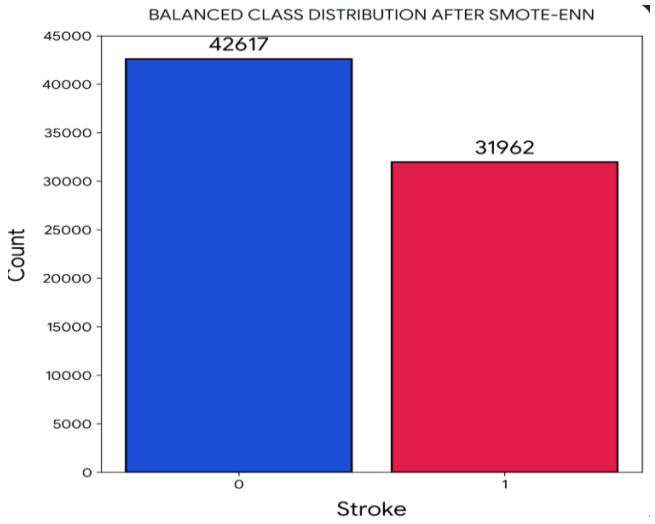

Fig.4 Class Distribution after SMOTE-ENN Balancing

Figure 4 illustrates the SMOTE-ENN-based Class Distribution Balancing Diagram illustrates the effectiveness of the sampling technique in addressing dataset imbalance. SMOTE-ENN generated synthetic minority class samples and eliminated noisy samples near class boundaries, producing a more balanced target distribution. After balancing, the dataset contained 42,617 non-stroke instances (Class 0) and 31,962 stroke instances (Class 1), ensuring improved class representation and enhancing the reliability of subsequent machine learning model training.

E. Data Splitting

The dataset was divided into training and testing subsets using an 80:20 split ratio. A stratified sampling strategy was employed to ensure that both subsets maintained the same class distribution as the original dataset, thereby preserving class balance and enabling reliable model training and evaluation.

F. Implementation of Models

Impact of Models in Machine Learning-Based Stroke Risk Prediction: Random Forest (RF) and XGBoost models enhance stroke risk prediction by effectively capturing complex clinical patterns, improving classification accuracy, and providing reliable risk assessment for early stroke detection.

1) Proposed Random Forest (RF) Model

Random Forest (RF) is an ensemble machine learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting. The model generates several independent decision trees using randomly selected subsets of training samples and features. Each tree provides an individual prediction, and the final classification outcome is determined through majority voting among all generated trees. The RF prediction function can be represented as eq (2).

$$\hat{Y} = mode\{T_1(X), T_2(X), T_3(X), \dots\dots, T_N(X)\} \quad (2)$$

where Y ˆ represents the final predicted class, Ti(X) denotes the prediction of the ith decision tree, X represents the input feature vector, and n indicates the total number of trees in the forest. The ensemble learning mechanism of RF enhances model stability, improves generalization capability, and provides effective classification performance for stroke risk prediction.

The probability estimation in the Random Forest classifier is calculated by aggregating the outputs of all individual trees, which can be expressed as eq (3):

$$P(Y = c|X) = \frac{1}{N}\sum_{i=1}^{N} I(T_i(X) = c) \quad (3)$$

where P(Y=CHX) represents the probability of class c, N is the number of decision trees, and I (·) is an indicator function that returns 1 when the condition is satisfied. This probability-based decision mechanism enables RF to accurately distinguish between stroke and non-stroke cases. The proposed Random Forest model was implemented with optimized hyperparameters including n_estimators = 200, max_depth = 20, min_samples_split = 2, min_samples_leaf = 1, max_features = "sqrt", and bootstrap = True to improve classification accuracy and reduce overfitting during stroke risk prediction.

2) Proposed Extreme Gradient Boosting (Xgboost) Model

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning technique based on gradient-boosted decision trees. It builds a sequence of weak learners, where each new tree improves the prediction errors produced by previous trees. XGBoost optimizes the objective function by minimizing prediction loss while controlling model complexity through regularization. The overall prediction function of XGBoost is formulated as eq (4):

$$\hat{Y}_i = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in F \quad (4)$$

where Y ^_i represents the predicted output for the i^{th}sample, K denotes the total number of trees, $f_k$represents the kth decision tree, X_i is the input feature vector, and F represents the set of all possible regression trees. This boosting strategy allows XGBoost to progressively improve classification performance by reducing prediction errors.

The objective function of XGBoost consists of the training loss and regularization term, which is expressed as:

$$obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (5)$$

where $L(y_i, \hat{y}_i)$ represents the loss function between actual and predicted values, y_i denotes the actual class label, $\hat{y}_i$ iindicates the predicted output, and $\Omega(f_k)$ controls the complexity of each tree. By incorporating regularization and sequential boosting, XGBoost provides high predictive accuracy, improved robustness, and efficient classification capability for stroke risk prediction. The proposed XGBoost model was configured with optimized parameters including n_estimators = 300, learning_rate = 0.05, max_depth = 6, min_child_weight = 1, subsample = 0.8, colsample_bytree = 0.8, and gamma = 0.1 to enhance model stability, learning efficiency, and predictive performance.

G. Evaluation Metrics

The performance of the proposed model was evaluated using various classification metrics. A confusion matrix was generated to analyze the prediction outcomes by representing the number of correctly and incorrectly classified instances for each class. The matrix provided four key parameters: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These values were subsequently used to compute important evaluation measures, including accuracy, precision, recall, and F1-score, as presented below:

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \quad (6)$$

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

$$Recall = \frac{\text{TP}}{TP+FN} \quad (8)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

The accuracy of a model is the percentage of samples that are correctly predicted by the model. Precision is the proportion of positive predictions that are correct, measuring the accuracy of positive predictions. Recall measures the proportion of true positive cases detected by the model out of all the cases. The F1-score is the harmonic mean of precision and recall, offering a decent balance of the model's performance. The Receiver Operating Characteristic (ROC) curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) over a range of classification thresholds to present the performance of classification. TPR is also known as sensitivity or recall, and FPR is also known as 1-specificity.

4. Results and Discussion

This section presents the experimental setup and evaluates the performance of the proposed models. The experiments were conducted on Intel Core i9-13900K (3.0 GHz) CPU, 64 GB DDR5 RAM, NVIDIA RTX 4090 GPU (24 GB VRAM), and Windows 11 Pro. Python libraries, including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, were used for preprocessing, visualization, model implementation, and evaluation. Performance was assessed using accuracy, precision, recall, and F1-score, as presented in Table II. Random Forest (RF) achieved 99.99% accuracy, 99.97% precision, 99.98% recall, and 99.96% F1-score, while XGBoost achieved 99.92% accuracy, 99.91% precision, 99.90% recall, and 99.93% F1-score. These results confirm that both models provide highly accurate stroke and non-stroke classification, with RF exhibiting slightly better overall predictive performance.

TABLE II. Classification Performance Results of the Proposed Model for Stroke Risk Prediction

Matrix	RF	XGBoost
Accuracy	99.99	99.92
Precision	99.97	99.91
Recall	99.98	99.90
F1-score	99.96	99.93

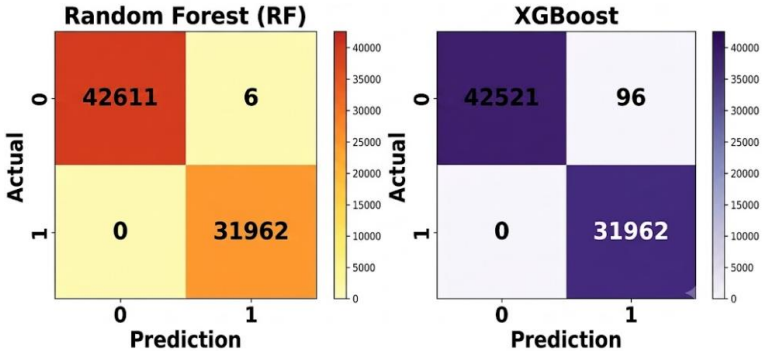

Fig.5. Confusion Matrix for the proposed Models

Figure 5 presents the confusion matrices of the proposed Random Forest (RF) and XGBoost models in a unified, high-contrast visualization with bold and clear labels for improved readability. The Random Forest matrix (left) records 42,611 True Negatives, 6 False Positives, 0 False Negatives, and 31,962 True Positives, while the XGBoost matrix (right) records 42,521 True Negatives, 96 False Positives, 0 False Negatives, and 31,962 True Positives. Both models correctly identify all positive stroke cases without any False Negatives, demonstrating excellent classification performance, with the Random Forest model producing fewer False Positives.

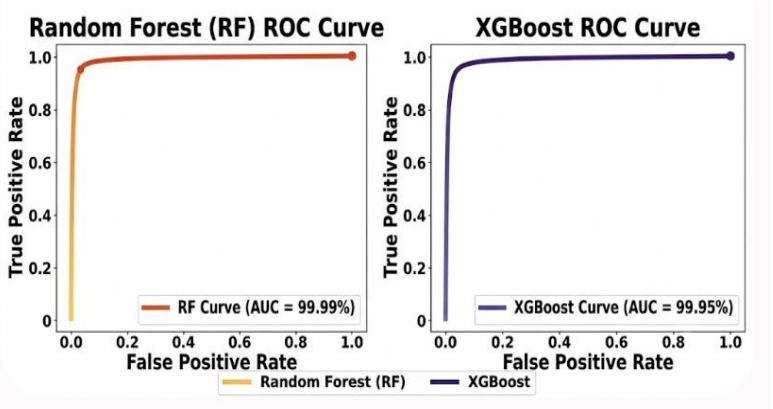

Fig.6. ROC Curve for Proposed Models

Figure 6 presents the ROC curve visualization of the proposed Random Forest (RF) and XGBoost models for binary classification. The Random Forest model achieves an AUC value of 99.99%, while the XGBoost model attains an AUC of 99.95%, indicating outstanding predictive performance. Both ROC curves approach the top-left region of the plot with a minimal False Positive Rate and a maximum True Positive Rate, demonstrating high classification reliability and effective discrimination ability of the proposed models.

A. Comparative Analysis

The effectiveness of the proposed models is analyzed by comparing their performance with the existing Machine Learning models presented in Table III. The comparison is evaluated using critical classification metrics such as accuracy, precision, recall, and F1 score. The existing models, such as Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Gradient Boosting Machine (GBM), and K-Nearest Neighbors (KNN), achieved comparatively lower performance, with accuracy values ranging from 63% to 94.45%. In contrast, the proposed Random Forest (RF) model demonstrated superior classification capability by achieving 99.99% accuracy, 99.97% precision, 99.98% recall, and 99.96% F1-score. Similarly, the proposed XGBoost model achieved outstanding results with 99.92% accuracy, 99.91% precision, 99.90% recall, and 99.93% F1-score. These results indicate that the proposed models significantly outperform conventional machine learning techniques and provide highly reliable predictions for stroke risk assessment.

TABLE III. Performance Comparison of Different Machine Learning Models for Stroke Risk Prediction

Model	Accuracy	Precision	Recall	F1-score
NB[32]	63	78	63	57
DT[33]	72	11	66	19
LR[34]	77.85	-	-	79.35
GBM[35]	89.63	91	90	90
KNN[35]	94.45	92.47	94.45	92.99
Proposed RF	99.99	99.97	99.98	99.96
Proposed XGBoost	99.92	99.91	99.90	99.93

The proposed Random Forest (RF) and XGBoost models demonstrated exceptional performance for stroke risk prediction, achieving accuracies of 99.99% and 99.92%, respectively. Their superior accuracy, precision, recall, and F1-score indicate highly reliable and robust classification of stroke and non-stroke cases. The ensemble learning strategies employed by both models effectively reduce prediction errors, improve generalization, and enhance resistance to overfitting. These advantages make the proposed models highly suitable for accurate, efficient, and dependable stroke risk prediction in clinical decision-support systems.

5. Conclusion and Future Study

Stroke is a major cause of disability and mortality worldwide, requiring accurate and timely risk prediction for effective prevention and early intervention. This study presented a comparative machine learning framework using the Kaggle Stroke Prediction Dataset, Min–Max normalization, SMOTE-ENN class balancing, Random Forest (RF), and XGBoost algorithms. The proposed models achieved outstanding predictive performance, with RF obtaining 99.99% accuracy and XGBoost achieving 99.92% accuracy, outperforming existing methods, including Naïve Bayes (63%), Decision Tree (72%), Logistic Regression (77.85%), GBM (89.63%), and KNN (94.45%). The experimental results demonstrate that ensemble learning methods effectively capture complex clinical patterns, reduce classification errors, and improve the identification of stroke-risk patients. Therefore, the proposed framework provides a reliable, accurate, and computationally efficient approach for early stroke risk prediction, supporting timely clinical decision-making and potentially improving patient outcomes.

Future studies will focus on validating the proposed framework using larger, multi-center, and diverse healthcare datasets to improve generalizability and reliability across different clinical settings. Additional clinical and genetic features, along with real-time patient monitoring data, will be incorporated to strengthen predictive capability. Future research will also investigate advanced optimization, explainable AI, and real-time deployment to enhance clinical applicability and support personalized stroke risk assessment.

REFERENCES

1. C. W. Tsao et al., "Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association," Circulation, vol. 147, no. 8, Feb. 2023, doi: 10.1161/CIR.0000000000001123.
2. A. Warrier and K. Abhilash, "Autonomous Agentic AI for Clinical Workflow Orchestration: Self-Managing Healthcare Operations," in 2025 6th International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India: IEEE, 2025, pp. 1450–1457, December. doi: 10.1109/ICICNIS66685.2025.11315712.
3. R. V. S. S. Bharattej R, H. M. Al-Jawahry, S. Rafikiran, D. Beula, and E. T. V. R. Kanth, "Blood Pressure Monitoring and Hypertension Prediction using Dynamic Opposite Learning Based Prairie Dog Optimization Algorithm," in 2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC), IEEE, Dec. 2024, pp. 1–7. doi: 10.1109/ICMNWC63764.2024.10872294.
4. M. Mittal, "The Transformative Role of CRM Systems in Modern Healthcare: Bridging the Provider-Patient Gap," Int. J. Multidiscip. Res., vol. 6, no. 6, pp. 1–9, November-December, 2024.
5. G. Kayola et al., "Stroke Rehabilitation in Low- and Middle-Income Countries," Am. J. Phys. Med. Rehabil., vol. 102, no. 2S, pp. S24–S32, Feb. 2023, doi: 10.1097/PHM.0000000000002128.
6. G. Y. H. Lip et al., "Integrated care for optimizing the management of stroke and associated heart disease: a position paper of the European Society of Cardiology Council on Stroke," Eur. Heart J., vol. 43, no. 26, pp. 2442–2460, Jul. 2022, doi:

10.1093/eurheartj/ehac245.

7. V. S. E and R. D, "A Systematic Method of Stroke Prediction Model based on Big Data and Machine Learning," in 2022 Smart Technologies, Communication and Robotics (STCR), IEEE, Dec. 2022, pp. 1–5. doi: 10.1109/STCR55312.2022.10009283.

8. S. Tiwari, A. Joshi, N. Rai, and P. Satpathy, "doi.org/10.1055/s-0041-1735323," J. Neurosci. Rural Pract., vol. 12, p. 680, Sep. 2021, doi: 10.1055/s-0041-1735323.

9. V. L. Feigin et al., "Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization–Lancet Neurology Commission," Lancet Neurol., vol. 22, no. 12, pp. 1160–1206, Dec. 2023, doi: 10.1016/S1474-4422(23)00277-6.

10. R. Kiruba, S. S. Rithish, P. Jagatheeswaran, R. Gokul, and S. Santhiya, "Prediction of Heart Stroke using Machine Learning Algorithms," in 2024 5th International Conference on Smart Electronics and Communication (ICOSEC), IEEE, Sep. 2024, pp. 1798–1804. doi: 10.1109/ICOSEC61587.2024.10722063.

11. R. Lingam, "Agentic AI and Federated Learning for Privacy-Preserving Dental Imaging," in 2026 Innovations in Machine, Engineering, and Digital Conference (IMED), Kota Kinabalu, Malaysia: IEEE, Mar. 2026, pp. 1–6. doi: 10.1109/IMED68921.2026.11484887.

12. A. Gupta, Next-Generation Computing and Information Systems. London, India: CRC Press, 2026. doi: 10.1201/9781003648987.

13. R. V. S. S. Bharattej R, R. Riad Al-Fatlawy, S. Meenakshi Sundaram, E. A. Rathnakumari, and M. N. Sudha, "Quantum Variational Based Support Vector Machine for Early Detection of Sepsis," in 2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkuru, India: IEEE, Dec. 2024, pp. 1–5. doi: 10.1109/ICMNWC63764.2024.10872228.

14. S. A. D. Poovaiah, G. N. Mohan Sankaran, N. Jooluri, and M. Mukhlisa, "Alzheimer's Disease Detection from MRI Scans using Deep Hybrid Architecture with Metaheuristic Hyperparameter Tuning," in 2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3), Bhubaneswar, India: IEEE, 2025, pp. 1–7, July. doi: 10.1109/ISAC364032.2025.11156608.

15. N. Melnykova, R. Kulievych, Y. Vycluk, K. Melnykova, and V. Melnykov, "Anomalies Detecting in Medical Metrics Using Machine Learning Tools," Procedia Comput. Sci., vol. 198, pp. 718–723, 2022, doi: 10.1016/j.procs.2021.12.312.

16. N.-B. Le, T.-T.-H. Pham, S.-H. Nguyen, N.-M. Nguyen, and T.-N. Nguyen, "AI-powered Predictive Model for Stroke and Diabetes Diagnostic," Int. J. Intell. Syst. Appl., vol. 16, no. 1, pp. 24–40, Feb. 2024, doi: 10.5815/ijisa.2024.01.03.

17. T. Raj Ojha and A. Kumar Jha, "Analyzing the Performance of the Machine Learning Algorithms for Stroke Detection," Int. J. Educ. Manag. Eng., vol. 13, no. 2, pp. 27–35, Apr. 2023, doi: 10.5815/ijeme.2023.02.04.

18. E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," Sensors, vol. 22, no. 13, p. 4670, Jun. 2022, doi: 10.3390/s22134670.

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

20. A. Warrier, "Machine Learning-Enhanced Data Quality Validation for Electronic Health Record Integration," Int. J. Sci. Technol., vol. 11, no. 1, pp. 1–8, Feb. 2020, doi: 10.71097/IJSAT.v11.i1.8387.

21. M. Mittal and S. Balamurugan, "Ethical Challenges in AI-Based Healthcare: Balancing Innovation With Patient Privacy And Fairness," Int. J. Artif. Intell. Med., vol. 3, no. 2, pp. 1–20, Jul. 2025, doi: 10.34218/IJAIMED_03_02_001.

22. V. L. Feigin et al., "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," Int. J. Stroke, vol. 17, no. 1, pp. 18–29, Jan. 2022, doi: 10.1177/17474930211065917.

23. M. O. Owolabi et al., "Primary stroke prevention worldwide: translating evidence into action," Lancet Public Heal., vol. 7, no. 1, pp. e74–e85, Jan. 2022, doi: 10.1016/S2468-2667(21)00230-9.

24. J.-T. Kim et al., "Neural network-based clustering model of ischemic stroke patients with a maximally distinct distribution of 1-year vascular outcomes," Sci. Rep., vol. 12, no. 1, p. 9420, Jun. 2022, doi: 10.1038/s41598-022-13636-w.

25. A. G, "Stroke Risk Prediction Using Imbalance-Sensitive Machine Learning Models on Structured Clinical Data," in 2026 4th International Conference on Self-Sustainable Artificial Intelligence Systems (ICSSAS), IEEE, May 2026, pp. 715–720. doi: 10.1109/ICSSAS68835.2026.11559523.

26. S. Lila, Z. Altoka, S. Jukic, B. Karlik, and J. Kevric,

"Evaluating Machine and Deep Learning Models for Stroke Prediction with Class Imbalance Handling," in 2026 15th Mediterranean Conference on Embedded Computing (MECO), IEEE, Jun. 2026, pp. 1–4. doi: 10.1109/MECO70748.2026.11579030.

27. S. N. Pasha, R. Dadi, and K. Vennela, "Brain Stroke Prediction using Machine Learning and Neural Networks Models: A Comparative Study," in 2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), IEEE, Jun. 2025, pp. 1639–1646. doi: 10.1109/ICSSAS66150.2025.11080982.

28. M. Z. Hossain Zamil, M. R. Islam, S. Debnath, M. T. Mia, M. A. Rahman, and A. K. Biswas, "Stroke Prediction on Healthcare Data Using SMOTE and Explainable Machine Learning," in 2025 13th International Symposium on Digital Forensics and Security (ISDFS), IEEE, Apr. 2025, pp. 1–6. doi: 10.1109/ISDFS65363.2025.11012059.

29. M. Sharma, I. Kaur, G. Saini, V. Thakur, and A. Mishra, "Predictive Modeling for Stroke Risk Assessment Using Machine Learning," in 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS), IEEE, May 2024, pp. 1–6. doi: 10.1109/ISCS61804.2024.10581155.

30. M. C. Das et al., "A comparative study of machine learning approaches for heart stroke prediction," in 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), IEEE, Jul. 2023, pp. 1–6. doi: 10.1109/SmartNets58706.2023.10216049.

31. R. M. D. Saputra, Y. Chairul, D. Riana, A. S. Hewiz, and F. Aziz, "Stroke Prediction Based on Random Forest with SMOTE," in 2023 International Conference on Information Technology Research and Innovation (ICITRI), IEEE, Aug. 2023, pp. 17–21. doi: 10.1109/ICITRI59340.2023.10249261.

32. Y. Li, Z. Wang, T. Wu, and T. Zhou, "Comparison of Six Machine Learning Algorithms for Stroke Risk Estimation," Appl. Comput. Eng., vol. 8, no. 1, pp. 556–561, Aug. 2023, doi: 10.54254/2755-2721/8/20230274.

33. Z. He, "Machine Learning for Stroke Onset Risk Prediction and Model Comparison," ITM Web Conf., vol. 88, p. 01037, Jul. 2026, doi: 10.1051/itmconf/20268801037.

34. A. S. M. A. S. Akib, R. Khawla, and A. Hasib, "Optimizing Stroke Risk Prediction: A Machine Learning Pipeline Combining ROS-Balanced Ensembles and XAI," 2025.

35. Z. Gao, "Comparative Analysis of Machine Learning Models for Stroke Risk Prediction," in Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy, SCITEPRESS - Science and Technology Publications, 2024, pp. 93–101. doi: 10.5220/0012998400004601.