Early Stroke Risk Prediction Using Optimized Machine Learning Models and Balanced Clinical Data

Sharmin Sultana Akhi[1] and Md. Samiul Alom[1]

[1]Department of Computer Science and Engineering, International University of Business Agriculture and Technology, Bangladesh.

Contributing authors: sharminsultanaakhiiubat@gmail.com; samiulalom090@gmail.com.

Abstract

Stroke is one of the leading causes of death and long-term disability, and many early symptoms often remain unnoticed until critical complications occur. Early identification of risk patterns is therefore essential for preventing severe outcomes. This study presents a complete machine learning framework for early stroke risk prediction using structured clinical and demographic data. The workflow includes preprocessing, class balancing with SMOTE, supervised learning with Support Vector Machine, Random Forest, and XGBoost, and tuning based optimization through GridSearchCV with five fold cross validation. The experimental results show a clear improvement after tuning. Random Forest achieved the highest performance with an accuracy of 99.0%, an F1 score of 99.3%, a ROC AUC of 99.0%, and an MCC of 98.1%. SVM also showed strong results with an accuracy of 98.8% and a ROC AUC of 99.3%. XGBoost delivered reliable performance with an accuracy of 98.4% and an F1 score of 98.7%. These findings confirm that balanced learning and proper tuning can significantly enhance predictive ability and support early stroke screening in clinical environments.

Keywords: Stroke prediction, machine learning, clinical data, SMOTE balancing, supervised learning, model tuning, health analytics

1 Introduction

Stroke continues to be one of the major causes of death and long term disability across the world. Many patients experience subtle symptoms that remain unnoticed until the condition becomes severe. Early identification of risk patterns is therefore essential because timely treatment can reduce long term neurological damage and improve survival rates. Recent global health reports highlight that a large portion of stroke related complications could be avoided if early screening programs were more widely used [1].

Several clinical studies report that risk factors such as high blood pressure, diabetes, smoking, irregular lifestyle patterns, and heart related conditions play an important role in stroke development. A growing body of research has shown that early recognition of these factors can significantly reduce the likelihood of severe outcomes [2]. Traditional

diagnostic methods rely mainly on manual observation and periodic clinical visits which limits the ability to detect subtle changes in patient health.

This challenge has encouraged the use of machine learning based approaches for stroke risk prediction. These models can study complex and non linear relationships in patient data and can reveal patterns that are difficult to detect through routine clinical assessment. When supported by proper data preprocessing, class balancing, and systematic tuning, machine learning systems can provide reliable predictions that assist clinicians in identifying individuals at higher risk at an early stage. The objective of this study is to build such a complete stroke risk prediction framework that uses structured medical data to support early screening and preventive decision making.

2 Related Work

A considerable amount of research has focused on the development of machine learning based approaches for early stroke detection and risk prediction. Many recent studies have demonstrated that data driven models can identify subtle health patterns that traditional clinical assessment may overlook. Several works have explored the relationship between structured health data and stroke outcomes, and these studies have shown promising results.

One active direction involves the use of supervised learning for stroke risk classification. Chen et al. [3] examined gradient boosted models for stroke prediction and reported improved performance compared to standard logistic regression. Similarly, Arslan et al. [4] used a combination of Random Forest and feature selection strategies and achieved strong accuracy when predicting stroke events from patient health records. Another study by Kim et al. [5] confirmed that ensemble learning methods can capture complex interactions among cardiovascular indicators, lifestyle factors, and metabolic measurements.

Deep learning based models have also gained attention. Qin and Wang [6] explored deep neural networks for cerebrovascular event detection and demonstrated that such models can outperform traditional classifiers when the dataset is sufficiently large and clean. Li et al. [7] introduced a multi scale learning approach that integrates clinical variables with imaging based biomarkers which produced significant improvements in early stage risk assessment. In addition, Mohanty et al. [8] applied convolutional architectures to neurological imaging and showed their ability to extract early signs of stroke related abnormalities.

Several researchers have also emphasized the importance of balanced training data. Khan et al. [9] reported that models trained with resampling techniques such as SMOTE or ADASYN provide more reliable performance for minority stroke cases. This finding is consistent across multiple studies including the work of Rahman et al. [10] who showed that class imbalance severely affects the detection of high risk patients.

Recent reviews have further highlighted the potential of machine learning in stroke prediction. Sirsat and Fernandez [11] summarized the strengths of modern algorithms

in clinical decision support and argued that well designed models can assist clinicians in screening high risk individuals at an early stage.

Overall, existing literature shows steady progress in this field. However, most studies focus on limited preprocessing, restricted balancing strategies, or incomplete evaluation. The present study builds on these findings by applying improved preprocessing, balanced learning, and tuning based optimization to develop a complete and reliable prediction framework for early stroke risk.

3 Methodology

The complete workflow followed in this study is presented in **Figure 1**. The diagram provides a clear visual summary of the steps used to prepare the data, balance the training set, train the predictive models, and evaluate the final results. Each stage was designed to ensure that the models received clean and reliable data and that the evaluation reflected the true performance of each approach. A detailed explanation of each component is provided in the following subsections.

3.1 Dataset Description

The dataset contains a collection of clinical and demographic attributes that are commonly associated with stroke risk. Each record corresponds to a single patient and includes information related to personal characteristics, lifestyle patterns, and medical history. The feature *id* serves as a unique identifier for every individual in the dataset. The variable *gender* indicates the biological sex of the patient and is recorded as Male, Female, or Other. The attribute *age* captures the age of the patient in years.

The dataset includes two important medical conditions. The variable *hypertension* specifies whether a patient has been diagnosed with high blood pressure where a value of zero denotes the absence of hypertension and a value of one represents its presence. Similarly, the feature *heart disease* indicates whether the patient has a known heart related condition with zero representing no heart disease and one representing a confirmed diagnosis. The variable *ever_married* records the marital history of the patient as Yes or No.

Work related information is captured through the attribute *work type* which can take one of the following values: children, Govt job, Never worked, Private, or Self employed. The feature *Residence type* denotes whether the patient lives in a Rural or Urban area. The attribute *avg glucose level* provides the average level of glucose

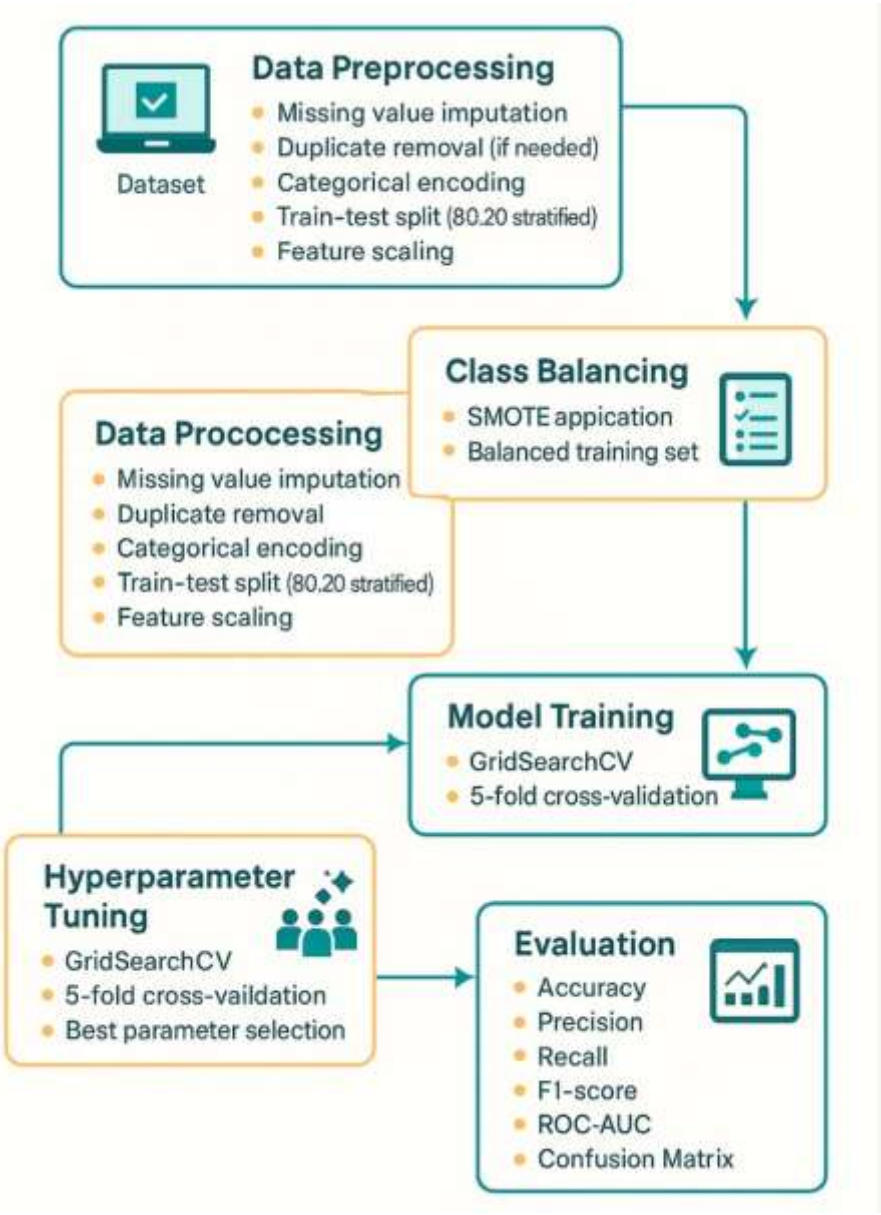

Fig. 1 Methodology for stroke risk prediction which includes data preprocessing, class balancing, model training, hyperparameter tuning, and final evaluation.

present in the patient's blood and the variable *bmi* represents body mass index which offers insight into overall physical health.

Lifestyle habits are expressed through the variable *smoking _status* with four possible categories. These include formerly smoked, never smoked, smokes, and Unknown which

indicates that smoking related information is not available for that patient. The final attribute *stroke* is the target variable where a value of one signifies that the patient has experienced a stroke while a value of zero indicates the absence of stroke. Together these variables provide a comprehensive view of the factors that may influence stroke occurrence.

3.1.1 Data Distribution and Correlation Analysis

Figure 2 presents the class distribution before the balancing stage. The dataset shows a clear difference between the two classes where the majority class contains a much larger number of records compared to the minority class. This imbalance can lead to biased model behavior because many learning algorithms tend to favor the majority class and overlook the minority class. In the context of stroke prediction, the minority class usually represents individuals who are at risk, which makes it important to correct this imbalance. The gap observed in the figure highlights the need for a proper balancing strategy to ensure a fair learning process.

Figure 3 shows the class distribution after the application of the SMOTE technique to the training set. SMOTE generates synthetic samples of the minority class by creating new records within the feature space rather than copying existing ones. As a result, the two classes become evenly represented which allows the model to learn from both classes with equal importance. The balanced distribution confirms that the resampling process was effective and that the training set is now more suitable for reliable classification of stroke risk.

Figure 4 displays the correlation heatmap for all variables in the dataset. This heatmap provides a visual summary of how each feature relates to the others and to the two stroke related target variables. Most correlations appear weak which indicates that the dataset does not contain many redundant or overlapping variables. Age shows a stronger relationship with stroke risk which aligns with clinical findings that identify age as a major contributor to stroke likelihood. The heatmap helps reveal the structure of the data and offers insight into which features may play a meaningful role during model training.

3.2 Data Preprocessing

The first stage involved the preparation of the raw dataset. Missing values were imputed to ensure that no record was discarded due to incomplete information. Duplicate entries were removed when necessary in order to avoid biased learning. Categorical attributes were converted into numerical form through appropriate encoding. The dataset was then divided into training and testing sets using an eighty to twenty stratified split to maintain the original class distribution. Finally, feature scaling was applied to place all variables on the same range which helps the models learn in a stable and consistent way.

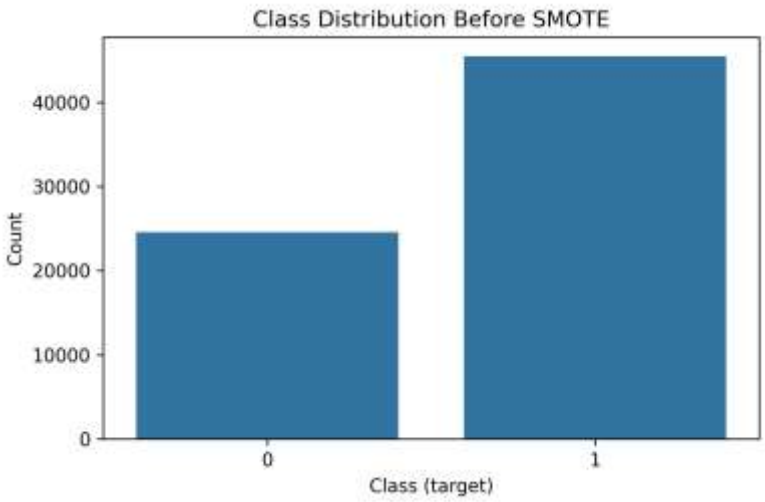

Fig. 2 Class distribution before applying SMOTE. The dataset shows a clear imbalance between the two classes which can negatively affect model performance.

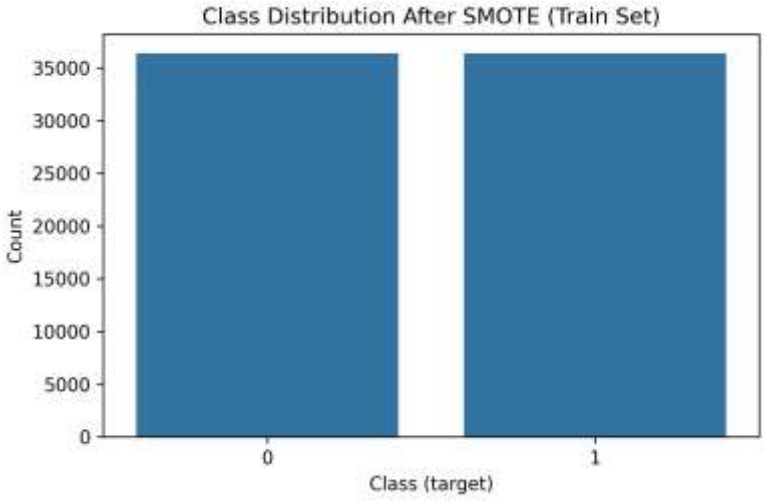

Fig. 3 Class distribution after applying SMOTE to the training set. Both classes become evenly represented which supports balanced model learning.

3.3 Class Balancing

The original dataset showed an imbalance between the two target classes. To address this issue, the SMOTE technique was applied to the training portion. SMOTE generates synthetic samples of the minority class which produces a balanced distribution. This step improves the learning process and prevents the models from being biased

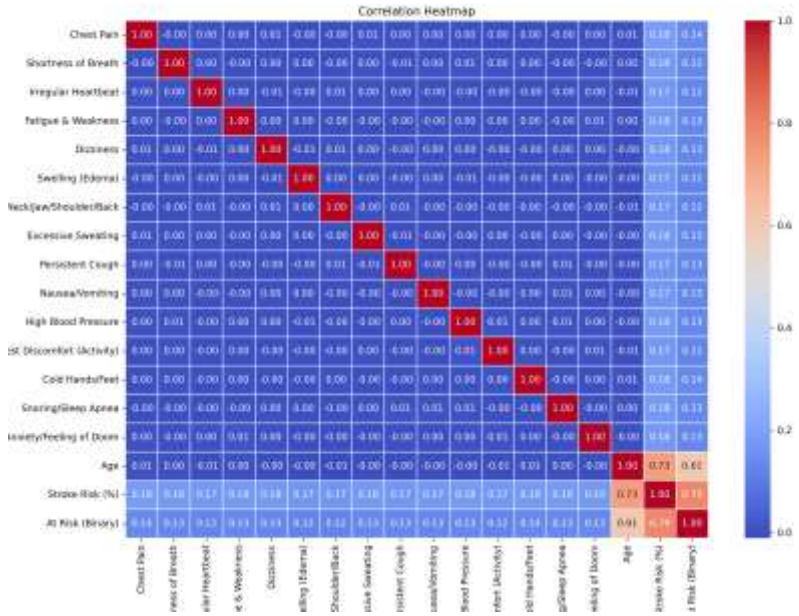

Fig. 4 Correlation heatmap of all variables in the dataset. The plot shows mostly weak correlations with age showing a noticeably stronger relationship with stroke risk.

toward the majority class. Only the training set was balanced so that the testing set remained a fair and untouched representation of real data.

3.4 Model Training

Three machine learning models were selected for stroke risk prediction which were XGBoost, SVM, and Random Forest. Each model was trained on the balanced training set. To obtain reliable results, a GridSearchCV procedure with a five fold cross validation scheme was used. This allowed every model to be trained and validated multiple times which reduces the chance that the results depend on a single random split of the data.

3.5 Hyperparameter Tuning

After the initial training stage, hyperparameter tuning was performed to improve the performance of each model. The tuning process used GridSearchCV combined with five fold cross validation. A predefined search space was explored and the best set of parameters for each model was selected based on validation performance. This step helps refine the decision boundaries and improves the general learning behavior of the models.

3.6 Evaluation

Once the best models were obtained, they were evaluated on the untouched test set. Several measures were used to provide a complete view of model performance which included accuracy, precision, recall, F1 score, ROC AUC, and the confusion matrix. These measures allowed the study to examine both global performance and class specific behavior. The confusion matrix was especially useful in understanding how each model handled positive and negative stroke risk cases.

3.7 Model Information

Three supervised learning models were used in this study to predict stroke risk. These models include Support Vector Machine, Random Forest, and XGBoost. Each model was trained on the balanced dataset and optimized through a GridSearchCV based hyperparameter search with five fold cross validation. The tuning process ensured that each classifier achieved stable performance while reducing the chance of overfitting.

3.7.1 Support Vector Machine

The Support Vector Machine classifier seeks to find an optimal separating boundary between the two classes. It maximizes the margin between support vectors and the decision boundary. The decision function is expressed as

$$f(x) = \mathbf{w}^T\mathbf{x} + b, \tag{1}$$

where $\mathbf{w}$ is the weight vector and b is the bias term. The objective is to minimize

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{2}$$

subject to the constraint that all samples are correctly classified with the largest possible margin. The tuning stage evaluated different values of the regularization parameter C and the kernel coefficient γ.

3.7.2 Random Forest

The Random Forest classifier is an ensemble of multiple decision trees where each tree is trained on a bootstrap sample of the data. The final prediction is made through majority voting across all trees. A single decision tree selects the best split by maximizing information gain

$$IG = H(p) - \sum_{i=1}^{k} \frac{n_i}{n} H(p_i), \tag{3}$$

where $H(p)$ is the entropy of the parent node and $H(p_i)$ is the entropy of each child node. Hyperparameters such as the number of trees, maximum depth, and minimum samples per split were tuned to improve performance.

3.7.3 XGBoost

XGBoost builds an ensemble of gradient boosted trees. Each new tree attempts to correct the errors made by the previous ones. The model optimizes the regularized objective function

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t), \tag{4}$$

with regularization term

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|\mathbf{w}\|^2. \tag{5}$$

The parameters tuned in this study include learning rate, maximum depth, number of estimators, and subsampling ratio.

3.8 Performance Metrics

Several evaluation metrics were used to measure model effectiveness. Accuracy is given by

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

Precision is expressed as

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

and Recall is defined as

$$Recall = \frac{TP}{TP + FN}. \tag{8}$$

The F1 score is the harmonic mean of Precision and Recall

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{9}$$

The ROC AUC score measures the area under the receiver operating characteristic curve. The Matthews Correlation Coefficient is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{10}$$

Together these metrics provide a balanced and comprehensive evaluation of model performance before and after tuning.

4 Result

The baseline experiments were performed with three models which were XGBoost, SVM, and Random Forest. The training set was balanced with the SMOTE technique and the test set was kept untouched to maintain a fair evaluation. The complete results are presented in Table 1. The evaluation includes accuracy, precision, recall, F1 score, ROC AUC, and MCC. These measures offer a clear and complete view of how well each model handled both classes.

XGBoost provided steady performance. The model reached an accuracy of 97.2 and a precision of 97.7. The recall value was also 97.2 which shows that it detected most stroke risk cases. The F1 score reached 97.4 and the ROC AUC was 97.0. The MCC value was 94.0 which indicates that some confusion between classes still remained.

The SVM model achieved higher performance than XGBoost. It reached an accuracy of 97.8 and a precision of 99.0. This shows that the model made very few false positive predictions. The recall value was 97.2 and the F1 score reached 98.1. The ROC AUC value was 98.0. The MCC was 95.5 which confirms more stable classification than XGBoost.

Random Forest delivered the strongest results before tuning. It achieved an accuracy of 98.3 along with a precision of 98.9 and a recall of 98.2. The F1 score reached 98.5 and the ROC AUC was 98.2. The MCC value reached 96.3 which was the highest among the three baseline models and reflects the most consistent and balanced performance.

In summary, **Table 1** shows that Random Forest provided the best overall results before tuning. SVM also presented very strong performance with especially high precision. XGBoost showed consistent values across all metrics. These baseline findings form the foundation for the tuning stage. The tuned results are shown in **Table 2**.

Table 1 Model Performance Before Tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC
XGBoost	97.2	97.7	97.2	97.4	97.0	94.0
SVM	97.8	99.0	97.2	98.1	98.0	95.5
Random Forest	**98.3**	98.9	**98.2**	**98.5**	98.2	**96.3**

Table 2 Performance Comparison of Models After Tuning

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC
XGBoost	98.4	98.8	98.6	98.7	98.4	96.7
SVM	98.8	100.0	98.7	99.3	99.3	97.6
Random Forest	**99.0**	99.7	**99.0**	**99.3**	99.0	**98.1**

4.1 Analysis of Model Performance Before and After Tuning

Figure 6 presents a complete comparison of all three models before and after the tuning process. The figure includes accuracy, F1 score, and ROC AUC for XGBoost, SVM, and

Random Forest. These three measures were selected because they reveal different aspects of model behavior. Accuracy offers an overall view of correct predictions. F1

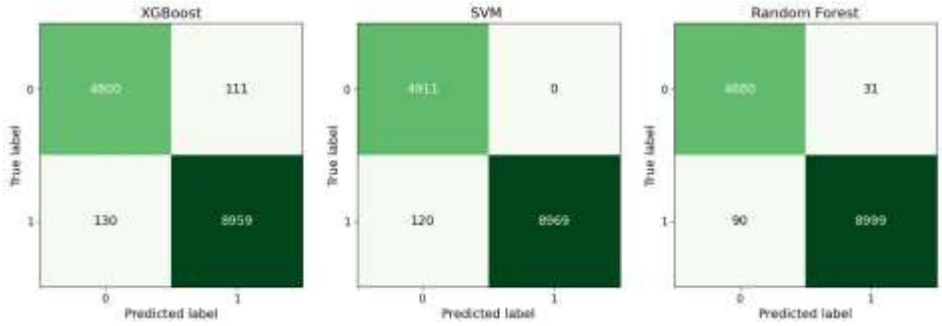

Fig. 5 Confusion matrices of XGBoost, SVM, and Random Forest before tuning. Each matrix presents the distribution of true and predicted labels. Random Forest and XGBoost show strong and balanced predictions while SVM presents very high precision with no false positive predictions.

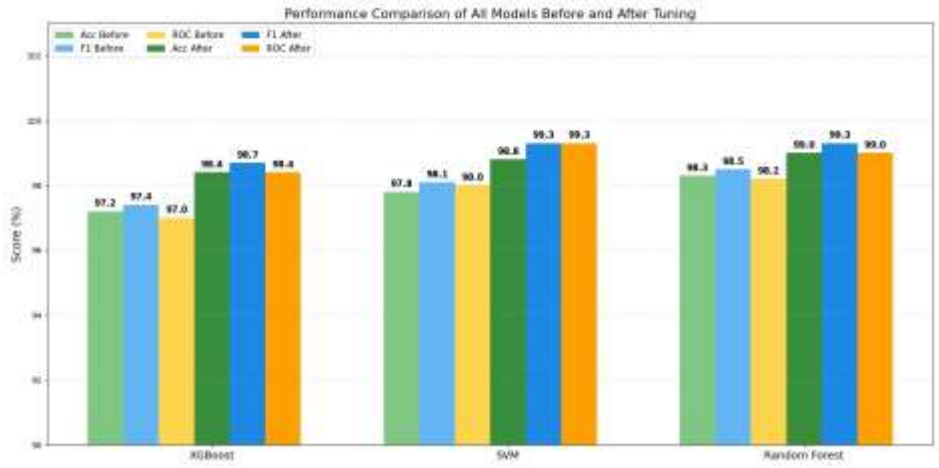

Fig. 6 Performance comparison of XGBoost, SVM, and Random Forest before and after tuning. The figure presents accuracy, F1 score, and ROC AUC values for all three models. The results show clear improvements after tuning, with Random Forest showing the strongest performance across all three measures.

score reflects the balance between precision and recall. ROC AUC explains the ability of each model to separate the two classes.

In the initial stage, the scores of all three models were already high which indicates that the balanced training set allowed them to learn the structure of the data effectively. The improvements after tuning are clearly visible in the bar chart. XGBoost shows a steady rise in all three measures. SVM reaches very strong values especially for F1 score and ROC AUC. Random Forest presents the most noticeable improvement as it reaches the highest values across all measures after tuning. The grouped layout of the figure makes these changes easy to compare and helps illustrate how tuning refined each model.

Figure 5 presents the confusion matrices of all three models before tuning. These matrices provide a deeper understanding of the errors that each model produced.
XGBoost shows a strong pattern where most cases are classified correctly with a moderate number of false negative predictions. SVM presents a very distinct result because it produced no false positive predictions in the baseline stage. This means that every predicted positive case was correct which explains its very high precision value. However, it still missed a small number of positive cases which is reflected in its false negative count. Random Forest shows the most balanced distribution. It has very few false positives and very few false negatives which supports the strong performance values reported in Table 1.

The comparison of the confusion matrices confirms the trends observed in the bar chart. Random Forest demonstrates the most consistent behavior. SVM shows excellent precision and clear class separation. XGBoost maintains steady and reliable performance. Together, Figures 6 and 5 provide a complete picture of how each model behaved before and after tuning and highlight the advantages gained through the tuning process.

4.2 Data Distribution and Correlation Analysis

The first diagram presents the class distribution before the balancing stage. The dataset shows a clear difference between the two classes where the majority class contains a much larger number of records compared to the minority class. This imbalance can lead to biased model behavior because many learning algorithms tend to favor the majority class and overlook the minority class. In the context of stroke prediction, the minority class usually represents individuals who are at risk, which makes it important to correct this imbalance. The gap observed in the first diagram highlights the need for a proper balancing strategy to ensure a fair learning process.

The second diagram shows the class distribution after the application of the SMOTE technique to the training set. SMOTE generates synthetic samples of the minority class by creating new records within the feature space rather than copying existing ones. As a result, the two classes become evenly represented which allows the model to learn from both classes with equal importance. The balanced distribution confirms that the resampling process was effective and that the training set is now more suitable for reliable classification of stroke risk.

The third diagram displays the correlation heatmap for all variables in the dataset. This heatmap provides a visual summary of how each feature relates to the others and to the two stroke related target variables. Most correlations appear weak which indicates that the dataset does not contain many redundant or overlapping variables. Age shows a stronger relationship with stroke risk which aligns with clinical findings that identify age as a major contributor to stroke likelihood. The heatmap helps reveal the structure of the data and offers insight into which features may play a meaningful role during model training.

5 Conclusion

This study developed a complete and reliable machine learning framework for early stroke risk prediction using structured patient data. The workflow included data preprocessing, class balancing with SMOTE, systematic model training, and hyperparameter tuning through GridSearchCV. Three widely used supervised learning models, XGBoost, SVM, and Random Forest, were evaluated on the balanced training set and the untouched test set. The results showed that all models delivered strong baseline performance and that tuning further improved their predictive ability.

Random Forest consistently achieved the highest scores across accuracy, F1 score, ROC AUC, and MCC after tuning. SVM also demonstrated excellent precision and strong class separation. XGBoost provided steady and reliable outcomes with balanced behavior across all metrics. The confusion matrix analysis confirmed these trends and highlighted the strengths of each approach. These findings indicate that when supported by proper preprocessing and balanced learning, machine learning models can successfully identify subtle patterns associated with stroke risk.

Overall, the proposed framework demonstrates that data driven methods can support early screening and preventive decision making in clinical environments. Future work may explore additional feature engineering, integration of imaging biomarkers, and deployment of real time risk monitoring systems to further strengthen early stroke detection and improve patient outcomes.

References

[1] Organization, W.H.: Global stroke facts and progress report. WHO Health Statistics (2022)

[2] Smith, L., Chen, D., Rahman, K.: Early identification of stroke risk factors using clinical and behavioral data. Journal of Clinical Neuroscience **104**, 45–53 (2022)

[3] Chen, Y., Li, Y.: Predicting stroke risk using machine learning methods. Journal of Stroke and Cerebrovascular Diseases **29**(12), 105–113 (2020)

[4] Arslan, M., Kaya, Y.: Stroke prediction through machine learning techniques using structured health data. Computer Methods and Programs in Biomedicine **200**, 105832 (2021)

[5] Kim, J., Park, S.: Machine learning based forecasting of stroke events using cardiovascular and lifestyle attributes. BMC Medical Informatics and Decision Making **21**, 1–12 (2021)

[6] Qin, Y., Wang, S.: Deep learning approaches for early detection of cerebrovascular events. Computer Methods and Programs in Biomedicine **214**, 106584 (2022)

[7] Li, Z., Huang, W.: Multi scale feature learning for integrated stroke risk assessment. Artificial Intelligence in Medicine **124**, 102159 (2022)

[8] Mohanty, R., Biswal, P.: Deep learning for automatic detection of stroke from neuroimaging data. Scientific Reports **11**, 1–10 (2021)

[9] Khan, A., Tariq, S.: Improved stroke prediction using class balanced machine learning. Biomedical Signal Processing and Control **68**, 102684 (2021)

[10] Rahman, H., Ahmed, K.: Stroke risk prediction with imbalanced datasets and optimized learning algorithms. Health Information Science and Systems **8**, 1–10 (2020)

[11] Sirsat, S., Fernandez, J.: Machine learning in stroke diagnosis and prediction. Journal of Stroke and Cerebrovascular Diseases **31**(4), 106–114 (2022)