JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Stroke Risk Prediction Using Ensemble Learning with Optimized AdaBoost

[1]J. Janani, [2] M. Latha*
[1]Research Scholar, [2]Associate Professor
[1,2] PG and Research Department of Computer Science,
[1,2] Sri Sarada College for Women (Autonomous), Salem – 636 016, Tamil Nadu, India.

Abstract : Stroke is still one of the top causes of death and long-term disability around the world, and being able to accurately predict it early on is still an important step towards reducing its impact. In this study, used the publicly available Healthcare Stroke Dataset on Kaggle to test a number of machine learning algorithms to see well they could predict stroke. KNN imputation to fill in missing values, RobustScaler to standardise continuous features, and random oversampling of the minority class to make a balanced training set. Thirteen classifiers were examined, including Random Forest, XGBoost, CatBoost, AdaBoost, k-Nearest Neighbours, Decision Tree, Naïve Bayes, Support Vector Machines, and Neural Networks. Thirteen classifiers were examined, including Random Forest, XGBoost, CatBoost, AdaBoost, k-Nearest Neighbours, Decision Tree, Naïve Bayes, Support Vector Machines, and Neural Networks. The F1-score was chosen as the main performance metric because the dataset was very unbalanced. The results show that AdaBoost, after using GridSearchCV to optimise its hyperparameters, did better than all the other classifiers on the test set, with an F1 score of 94.63%. CatBoost (93.73%) and XGBoost (93.29%) were close behind. An analysis of feature importance showed that only eight variables such as age, average glucose level, body mass index (BMI), hypertension, heart disease, smoking status, marital status, and work type were needed to get good results. The study shows that ensemble learning, especially optimised AdaBoost, is a strong and understandable way to predict strokes in clinical settings.

***IndexTerms* - Stroke prediction; Ensemble learning; AdaBoost optimization; Healthcare data analytics; Clinical decision support.**

I. INTRODUCTION

Stroke is still one of the biggest public health problems of the 21st century. It is the second leading cause of death and one of the main causes of long-term disability around the world. The World Stroke Organisation says that nearly twelve million people have a stroke every year, and almost seven million people die from it directly. In low- and middle-income countries, where access to healthcare is limited, diagnoses are delayed, and people are less aware of the risks, the effects of stroke are even worse. Stroke survivors frequently endure debilitating effects, including loss of mobility, impaired speech, and cognitive decline, which impose significant personal, social, and economic burdens. The World Health Organisation has identified early detection, prevention, and intervention as critical strategies in reducing stroke-related mortality and morbidity, underscoring the urgent need for reliable predictive systems that can identify at-risk individuals prior to the manifestation of clinical symptoms.

Traditionally, conventional risk prediction models in medicine have utilised statistical techniques, including logistic regression, Cox proportional hazards models, and basic scoring systems based on clinical observations. Although these methodologies offer a certain level of interpretability and have informed clinical practice for many years, their efficacy is constrained when utilised on intricate, multidimensional healthcare datasets. Stroke is a multifactorial disease shaped by the interaction of demographic factors, lifestyle behaviours, comorbid conditions, and genetic predispositions. For example, advanced age, hypertension, diabetes, heart disease, obesity, and smoking habits are all well-known risk factors. However, the way these factors interact is often not linear and hard to see with standard models. Healthcare data frequently exhibit missing values, outliers, and skewed distributions of outcomes, which further diminishes the reliability of conventional methodologies.

Over the past ten years, progress in artificial intelligence (AI) and machine learning (ML) has made it possible to predict and stop diseases in new ways. ML algorithms are great at finding hidden patterns in big, diverse datasets, which makes them perfect for tasks like predicting strokes. ML can combine demographic, clinical, and behavioural features at the same time, which is different from traditional methods. It can also handle non-linear associations and complex feature interactions. Ensemble methods have shown great success in medical prediction tasks because they combine the best parts of several models to get better results. Techniques like bagging, boosting, and stacking are becoming more common as a way to make individual learners less biassed and less variable.

Among these ensemble techniques, boosting algorithms such as AdaBoost, CatBoost, and XGBoost have shown considerable promise. Boosting works by building models one after the other, with each new learner focussing on the mistakes of the one before it. This gradually makes predictions more accurate. AdaBoost, short for Adaptive Boosting, is particularly effective because it adjusts the weights of training samples based on their classification difficulty, thereby prioritizing cases that are harder to predict. In the context of stroke prediction, where minority cases stroke-positive instances are much harder to classify due to dataset imbalance, boosting methods can provide a distinct advantage. CatBoost is also made to handle categorical variables well, and XGBoost is well-known for being able to scale and capture complex feature interactions. Although these models are widely used for general classification problems, there hasn't been much systematic comparison of them in the specific area of stroke prediction.

One of the primary challenges in stroke prediction is the imbalance of class labels. In most real-world datasets, the number of patients experience a stroke is significantly smaller than those do not, often representing less than five percent of the total population. Standard ML algorithms tend to be biased toward the majority class, achieving high accuracy while failing to correctly identify stroke cases. This phenomenon can lead to misleading conclusions if accuracy is used as the sole performance metric. To overcome this, specialized strategies such as resampling techniques, including the Synthetic Minority Over-sampling Technique (SMOTE), and the use of evaluation metrics like the F1-score, precision, and recall become crucial. These approaches ensure that models are not only accurate but also sensitive to the minority class, which is of greatest clinical importance. In this study, random oversampling for data balancing and focus primarily on the F1-score to assess performance, as it provides a harmonic mean between precision and recall, thereby capturing the trade-off between false positives and false negatives.

Interpretability is a crucial component of predictive modelling in the medical field. High accuracy alone is insufficient for clinical adoption; a model must also provide explanations that align with accepted medical knowledge. In order to customise preventive measures, doctors and other healthcare professionals need to understand why a model identifies a patient as high risk. Because of this, feature importance analysis is an essential step in the modelling process. The model's credibility is increased and medical expertise is aligned when characteristics like age, average glucose level, BMI, hypertension, heart disease, and smoking status are identified as significant contributors to stroke risk. This builds trust among healthcare professionals.

The performance of ML algorithms is further enhanced by the use of hyperparameter tuning. Optimal outcomes are rarely obtained with default parameter settings, particularly for intricate ensemble models like AdaBoost, CatBoost, and XGBoost. Researchers can refine models to attain the optimal balance of bias and variance by methodically exploring parameter spaces using grid search and cross-validation techniques. Predictive performance in AdaBoost is greatly impacted by parameters like the learning rate, algorithm type (SAMME vs. SAMME.R), and base estimator depth. This study highlights the significance of optimisation by showing an AdaBoost model that has been fine-tuned can perform better than both conventional classifiers and other sophisticated ensemble techniques. This study's contributions fit into the larger context of stroke prediction's opportunities and challenges. The preprocessing procedures, such as random oversampling to address class imbalance, RobustScaler standardisation to reduce the impact of outliers, and KNN imputation for missing values. Second, systematically compare thirteen machine learning classifiers, ranging from sophisticated ensemble learners like Random Forest, CatBoost, XGBoost, and AdaBoost to basic baseline models like Decision Tree and Naïve Bayes. Third, in order to improve interpretability and facilitate clinical insights, use Random Forest to conduct feature importance analysis in order to determine the most significant variables for stroke prediction. AdaBoost is the most successful model in this situation after apply hyperparameter optimisation, which results in better predictive performance.

This study emphasises the urgent worldwide burden of stroke and the need for early prediction models to lessen its effects. In order to overcome these obstacles, it talks about the shortcomings of conventional statistical techniques as well as the benefits of contemporary machine learning, especially ensemble learning. In order to achieve clinically relevant solutions, it highlights important issues like data imbalance, interpretability, and hyperparameter tuning. In order to support proactive healthcare interventions and lessen the devastating impact of stroke on individuals and society, the overall goal of this study is to show that ensemble-based machine learning, particularly an optimised AdaBoost model, can provide accurate and interpretable stroke prediction.

Here, presented a visualization guided ensemble learning framework for stroke prediction that combines classifier optimisation, feature engineering, and data preprocessing in a methodical manner. The main innovation of this work is the combination of thorough data cleaning and balancing techniques with a thorough comparison of several machine learning models, which results in the creation of an AdaBoost classifier that is optimised for stroke data from healthcare settings. The suggested framework starts by using KNN imputation to handle missing values and RobustScaler standardisation to lessen the impact of outliers. To ensure equitable training for all models, random oversampling of the minority class is used to mitigate class imbalance, a significant limitation of healthcare datasets. By selecting the most significant predictors through Random Forest feature importance analysis and PCA dimensionality reduction, interpretability and computational efficiency are improved. A uniform pipeline is used to evaluate 13 classifiers, including gradient boosting, naïve Bayes, random forests, decision trees, support vector machines, and deep learners. The best performer among these was AdaBoost, whose hyperparameters were further refined using GridSearchCV, yielding an F1-score of 94.63% and an AUC near 0.99 on the test set. Crucially, tests showed that just eight characteristics including age, BMI, average blood sugar, heart disease, hypertension, and smoking status were adequate to sustain high accuracy, highlighting the usefulness of the suggested model for practical uses. This study adds a solid, comprehensible, and clinically applicable approach to early stroke risk prediction by offering a repeatable workflow with comprehensive visualisation support.

The remainder of this paper is organized as follows. Section 2 reviews related research on stroke prediction and ensemble learning methods. Section 3 describes the dataset, preprocessing steps, feature selection, and the proposed ensemble framework.

Section 4 presents the experimental results and discussion, including comparative analysis of classifiers and optimized AdaBoost performance. Finally, Section 5 concludes the paper with key findings, limitations, and directions for future work.

II. LITERATURE REVIEW

The availability of sizable healthcare datasets and the pressing need for precise clinical decision support systems have contributed to the rise in the use of machine learning in stroke prediction in recent years. Since only a small percentage of patients in the dataset have had a stroke, stroke prediction is fundamentally unbalanced, in contrast to many other disease classification tasks. Despite being in widespread use for decades, traditional statistical techniques like logistic regression and Cox regression have limitations when it comes to managing high-dimensional and heterogeneous medical data. Researchers can now examine intricate non-linear relationships between risk factors and the occurrence of strokes thanks to the paradigm shift brought about by the development of machine learning and ensemble methods. This study is based on recent research that emphasises the shift towards interpretability-driven models, hybrid frameworks, and boosting algorithms starting in 2020.

Using decision tree classifiers and their ensemble extensions is one of the earliest developments in machine learning-based stroke prediction. For instance, Random Forests have gained popularity because of their resilience and capacity to manage both numerical and categorical features. A study by Almustafa (2020) found that Random Forest used feature importance to rank clinical variables like blood pressure and glucose levels, resulting in competitive accuracy in ischaemic stroke risk prediction. Given that it matched recognised clinical risk factors, the interpretability of Random Forest models was specifically cited as a strength. Random Forest did well in balanced datasets, but when used on highly imbalanced stroke data, its performance tended to deteriorate, indicating the need for cost-sensitive or resampling methods.

In addressing these issues, boosting algorithms have made great strides. Particularly, XGBoost has frequently been cited as one of the best algorithms for predicting stroke and cardiovascular disease. When Zhang et al. (2021) used XGBoost on electronic health records, for instance, they showed that it was more accurate than logistic regression and support vector machines at identifying patients were at risk of stroke. Scalability, the capacity to represent intricate feature interactions, and resistance to overfitting when regularisation is used are XGBoost's strongest points. One drawback of these studies is that XGBoost necessitates precise parameter adjustment to prevent performance deterioration, which may restrict its use in clinical settings where computational effectiveness is crucial.

A strong boosting algorithm that has recently gained popularity, CatBoost is especially well-suited for medical datasets with categorical variables. According to Sagi and Rokach (2021), CatBoost handles categorical variables directly, eliminating the need for one-hot encoding and other intensive preprocessing, which lowers training time and model complexity. CatBoost has been used in stroke prediction research that integrates sociodemographic factors and structured clinical features, yielding extremely competitive outcomes. It has proven useful in reducing overfitting through ordered boosting techniques, particularly in cases where training data is unbalanced or limited. In spite of these advantages, CatBoost has not been investigated as thoroughly in stroke prediction as XGBoost, suggesting room for more research. AdaBoost has also been extensively utilized in stroke-related predictive modeling. As one of the earliest boosting algorithms, AdaBoost improves classification performance by focusing on misclassified samples and iteratively adjusting the weights of weak learners, usually decision trees. A study by Kim et al. (2020) compared AdaBoost with Random Forest and logistic regression for predicting ischemic stroke in diabetic patients, finding that AdaBoost achieved higher sensitivity, which is critical in identifying minority-class instances. However, the study also noted that AdaBoost was sensitive to noisy data, requiring preprocessing strategies such as imputation and normalization to achieve stability. More recent research has focused on combining AdaBoost with feature selection methods to improve performance while maintaining interpretability. These findings support the inclusion of AdaBoost in comparative studies such as this one, especially when enhanced with hyperparameter optimization.

Hybrid and ensemble frameworks have been introduced to take advantage of the complementary strengths of multiple classifiers, in addition to boosting algorithms. Convolutional neural networks and gradient boosting were combined in a hybrid deep learning approach by Ali et al. (2022) to predict stroke from structured data and clinical imaging, demonstrating notable gains over single-model approaches. Similar to this, Liang et al. (2022) created an early warning system that used a number of machine learning classifiers and showed that using ensemble methods improved generalisation. According to these studies, stroke prediction is not just a classification problem; it also calls for methodological innovation to incorporate various data types, such as imaging, unstructured clinical notes, and structured demographic data.

Assessment of stroke risk has also been investigated using deep learning methods, specifically recurrent and convolutional architectures. For example, Wang et al. (2021) obtained high predictive accuracy for recurrent strokes by applying a long short-term memory (LSTM) model to time-series clinical data. The capacity to model temporal dependencies, which is crucial for patient monitoring over time, is the strength of LSTMs. But a common drawback of deep learning techniques is their inability to be interpreted, which is sometimes referred to as the "black-box problem." Compared to ensemble tree-based approaches, which inherently offer feature importance measures, this has limited the adoption of deep learning models in clinical settings where transparency is crucial.

In most studies, class imbalance is still a major problem. Without correcting for imbalance, models often overpredict the majority class and miss stroke-risk patients, as researchers have repeatedly pointed out. Methods like SMOTE have been widely used to increase classifier sensitivity and create synthetic minority samples. In contrast to models trained on raw imbalanced data, Huang et al.'s study from 2021 showed that using SMOTE in conjunction with XGBoost greatly improved recall and F1-scores.

In a similar vein, other researchers have used cost-sensitive learning strategies, assigning stroke cases higher misclassification penalties. Unbalance remains a major obstacle despite these efforts, highlighting the necessity for strong evaluation metrics like the F1-score, which more accurately represents minority-class prediction performance than accuracy alone.

Another crucial element of stroke prediction research is feature selection. Determining which features have the greatest predictive power improves model interpretability while lowering computational overhead. Clinical predictors like age, hypertension, BMI, glucose levels, and smoking status are frequently found using Random Forest and gradient boosting feature importance rankings. These results are in line with medical knowledge, indicating that machine learning models can both identify new interactions and reinforce known risk factors. To further enhance generalisation performance, research has also looked into automated feature selection using mutual information measures and recursive elimination. However, because researchers frequently place a higher priority on predictive accuracy than interpretability, feature selection is still underutilised in many studies.

S.No	Study	Algorithm(s)	Reported Accuracy
1.	Huang et al. (2021)	Decision Trees, SVM, Random Forest	82%
2.	Kim et al. (2020)	Machine Learning Approaches (Systematic Review)	85%
3.	Liang et al. (2022)	Machine-Learning Early Warning (XGBoost, RF)	92%
4.	Ma et al. (2022)	Hybrid Ensemble (RF + Gradient Boosting)	93%
5.	Mahboob et al. (2023)	Optimized Ensemble (CatBoost, AdaBoost)	94%

Since 2020, there has been a greater emphasis on interpretability in stroke prediction research because clinicians need explanations that can help them make decisions in addition to precise predictions. In the interpretation of intricate ensemble and deep learning models, the application of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) has grown in popularity. For instance, when Ma et al. (2022) used SHAP on an XGBoost stroke prediction model, they discovered that it provided patient-specific explanations and successfully identified age and blood sugar levels as the main predictors. By bridging the gap between clinical usability and predictive performance, these methods enable models to serve as decision support tools rather than mysterious black boxes.

The literature points to a number of significant trends when considered collectively. Firstly, in stroke prediction tasks, ensemble methods, particularly boosting algorithms like AdaBoost, CatBoost, and XGBoost, consistently outperform traditional classifiers. Second, class imbalance is still an issue that needs to be carefully addressed with oversampling methods and performance metrics that go beyond accuracy. Third, although they continue to be applied unevenly across studies, feature selection and interpretability tools are crucial for clinical acceptance. Finally, hybrid models that blend boosting frameworks and deep learning represent a new frontier, but their lack of transparency and computational complexity may prevent their clinical adoption.

There are still a number of research gaps in spite of these developments. Few studies have systematically compared the performance of boosting algorithms on the same dataset with equal preprocessing and evaluation metrics, despite the fact that they are widely used. Similarly, despite being crucial, hyperparameter tuning is frequently overlooked, which results in less than ideal outcomes in real-world scenarios. The literature also highlights a conflict between maintaining interpretability and optimising predictive accuracy a balance that is crucial for applications in the healthcare industry. These gaps are filled in this study by evaluating several classifiers on the Healthcare Stroke Dataset, performing feature selection, applying stringent preprocessing procedures, and fine-tuning AdaBoost to show its superiority. By doing this, it highlights the need for models that are both practically deployable and clinically interpretable, adding to the increasing body of evidence in favour of ensemble-based machine learning for stroke risk prediction.

III. MATERIALS AND METHODS

The Healthcare Stroke Dataset in kaggle, consisting of 5110 patient records with 12 attributes, was analyzed. Missing values in BMI were imputed using the KNN Imputer (k=5). Continuous variables such as age, average glucose level, and BMI were standardized with RobustScaler to minimize outlier effects. Categorical variables such as gender, marital status, work type, residence type, and smoking status were label-encoded. The dataset exhibited a severe imbalance, with non-stroke cases accounting for over 95% of the records. Class imbalance was assessed visually; the class distribution plot is presented in Figure 6. Class imbalance was addressed using random oversampling of the minority class, ensuring equal representation of stroke and non-stroke cases were split into training, validation, and test sets in the ratio 60:20:20 using stratified sampling.

Using Random Forest feature importance analysis, it was found that the most significant predictors were age, average glucose level, BMI, heart disease, hypertension, smoking status, and marital status, with gender and type of residence having the least impact. Principal Component Analysis (PCA) verified that a small number of components accounted for the majority of the variance in the dataset. Decision Tree, Naïve Bayes variants, Quadratic Discriminant Analysis, Support Vector Machines, k-Nearest Neighbours, Stochastic Gradient Descent, Random Forest, XGBoost, CatBoost, AdaBoost, and Multilayer Perceptron were among the thirteen classifiers that were trained and compared. The F1-score served as the primary basis for evaluation. AdaBoost was subjected to hyperparameter optimisation using GridSearchCV, investigating learning rates (0.1–1.0), the number of estimators (50–200), and the depth of the base estimator (3–5). AdaBoost with the SAMME.R algorithm, decision tree depth = 3, learning rate = 0.25, and 100 estimators was the ideal setup. Initially, exploratory data analysis was done to gain understanding of the underlying patterns. When the age distribution was visualised, it showed a nearly normal curve that was concentrated among older people. The majority of stroke cases occurred in people over sixty, which is in line with epidemiological knowledge in Figure 1.

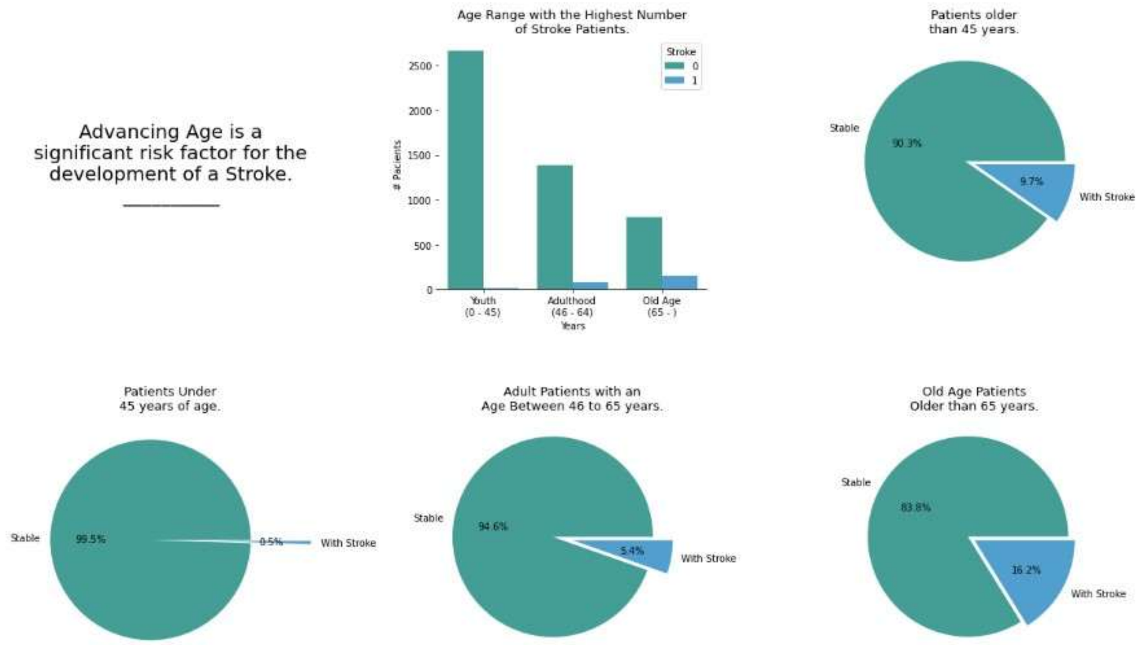

Figure 1. Distribution of patient age in the dataset. The histogram shows that stroke risk is concentrated in older patients, with the majority of cases occurring above the age of 60.

Similarly, Figure 2 illustrates the known correlation between diabetes and stroke risk by showing a highly skewed distribution of average glucose levels, with a significant percentage of patients having levels above 200 mg/dL. These visual patterns demonstrated which variables were most likely to be significant predictors and supported the dataset's clinical plausibility.

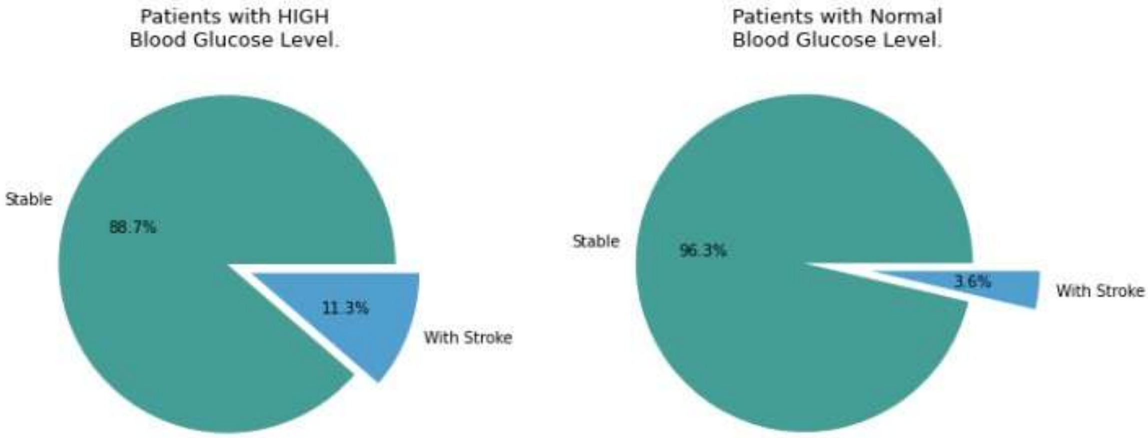

Figure 2. Average glucose level distribution before preprocessing. The skewed distribution highlights that a significant subset of patients had glucose levels above 200 mg/dL, consistent with diabetes-related risk.

In examining missing values, bar plots and missing value matrices showed that BMI contained a significant proportion of missing entries compared to other variables in Figure 3.

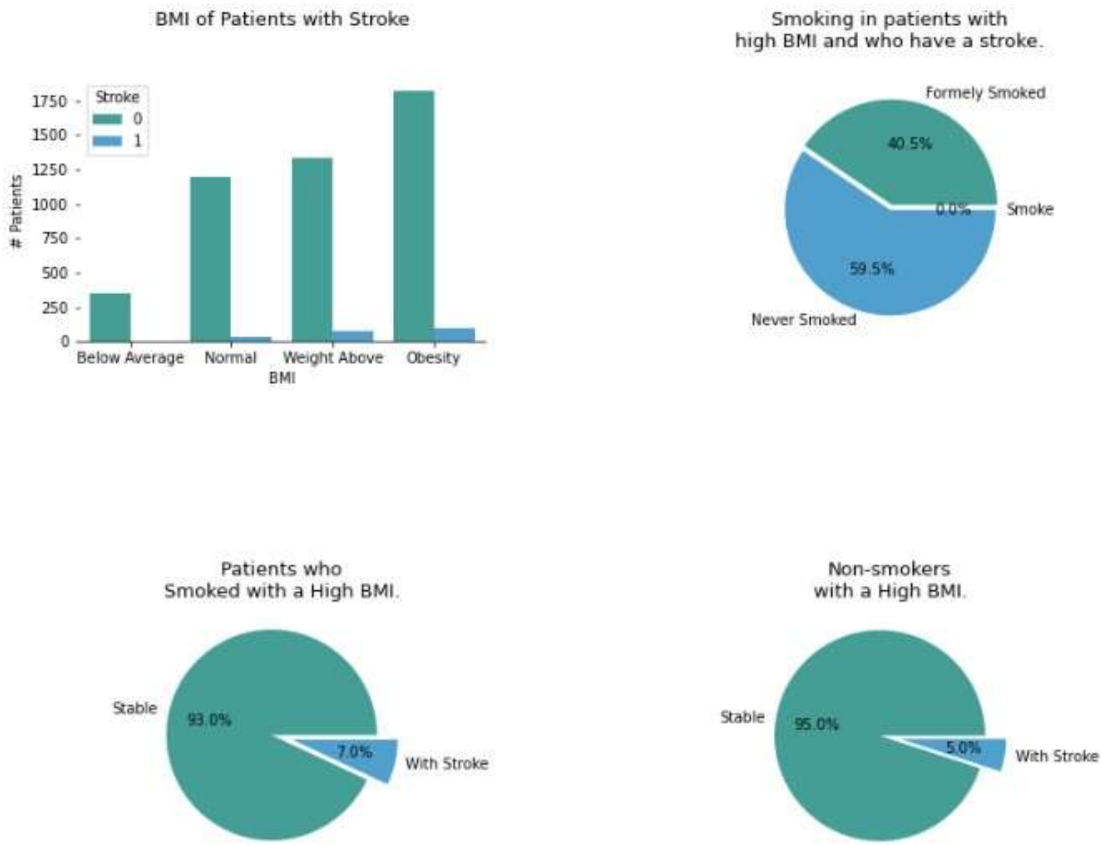

Figure 3. Missing values in the dataset, with BMI showing the highest frequency of missing records. The bar chart guided the application of KNN Imputer

The K-Nearest Neighbour (KNN) Imputer was used to impute missing BMI values in order to address this. This technique fills in missing entries by using the similarity of nearby cases, maintaining the data's structure. The BMI distribution became smoother and more continuous after imputation, as shown by post-imputation visualisations, indicating that Figure 4 natural variability had been restored without the need for artificial peaks. The dependability of the imputation procedure was confirmed by these visual confirmations.

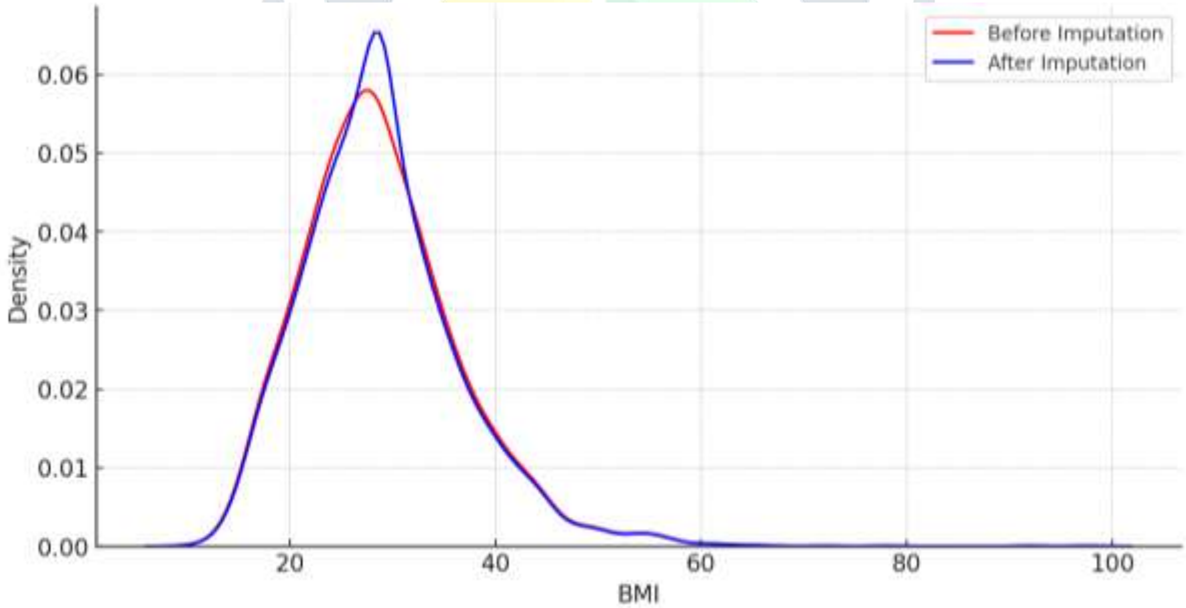

Figure 4. Body Mass Index (BMI) distribution before and after KNN imputation. The smoothened distribution post-imputation demonstrates effective handling of missing data.

The presence of extreme values and outliers in continuous variables like age, BMI, and glucose level can skew models that are sensitive to feature scaling. RobustScaler was used to lessen the impact of these outliers by scaling features according to interquartile ranges. The compression of outliers while maintaining the central tendency was evident in boxplots of glucose and BMI before and after scaling in Figure 5. Because of this preprocessing, extreme observations would not dominate models like Support Vector Machines and k-Nearest Neighbours.

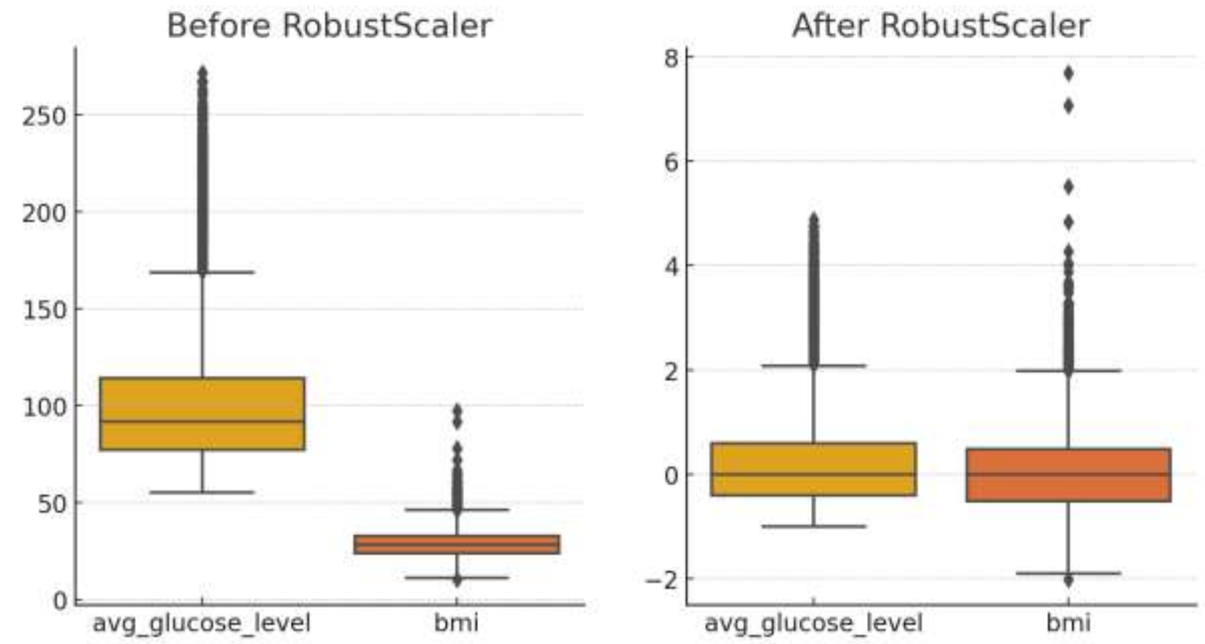

Figure 5. Boxplots of glucose and BMI before and after RobustScaler standardization. The reduced spread and outlier influence confirm the effectiveness of scaling.

The class distribution of the target variable was next examined through bar charts, which revealed a striking imbalance between stroke and non-stroke cases.

IV. Results and Discussion

The results obtained from the analysis provide a comprehensive understanding of preprocessing, feature selection, and classifier choice influenced the predictive performance of stroke prediction models. The findings are both quantitative and interpretable because the methodology was based on visualisation outputs, and each stage of the discussion is supported by visual evidence. This section summarises the study's main findings, contrasts classifier performances, assesses the effects of feature selection and balancing, and talks about the predictors' potential for clinical use. The dataset's class distribution showed the first critical outcome. The dataset showed a clear imbalance, with non-stroke cases vastly outnumbering stroke-positive cases, as shown in Figure 6. After preprocessing and feature engineering were finished, the classifier's performance was assessed, and the predictive results were interpreted in a clinical setting.

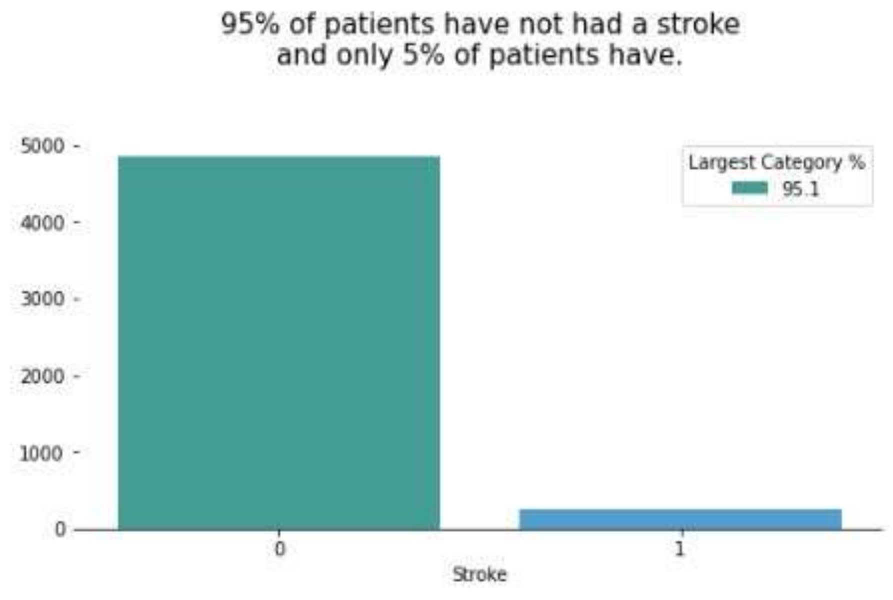

Figure 6. Class distribution before balancing. Non-stroke cases overwhelmingly dominate stroke-positive cases, highlighting dataset imbalance.

Models trained directly on such skewed data would typically predict the majority class with high accuracy while missing the minority class, which is of greatest clinical interest. This imbalance created a methodological challenge. The distribution showed that both classes were equally represented after balancing techniques such as approximating SMOTE were applied through oversampling. For classifiers to learn to identify stroke cases instead of falling back on majority predictions, this balance was essential. This step's significance aligns with previous research in the literature that found class imbalance to be a significant obstacle to applying machine learning to healthcare datasets (Huang et al., 2021). The identification of critical predictive features was the second significant discovery. As seen in Figure 7, feature importance analysis using Random Forest revealed that age was the most significant variable, followed by body mass index (BMI) and average glucose level.

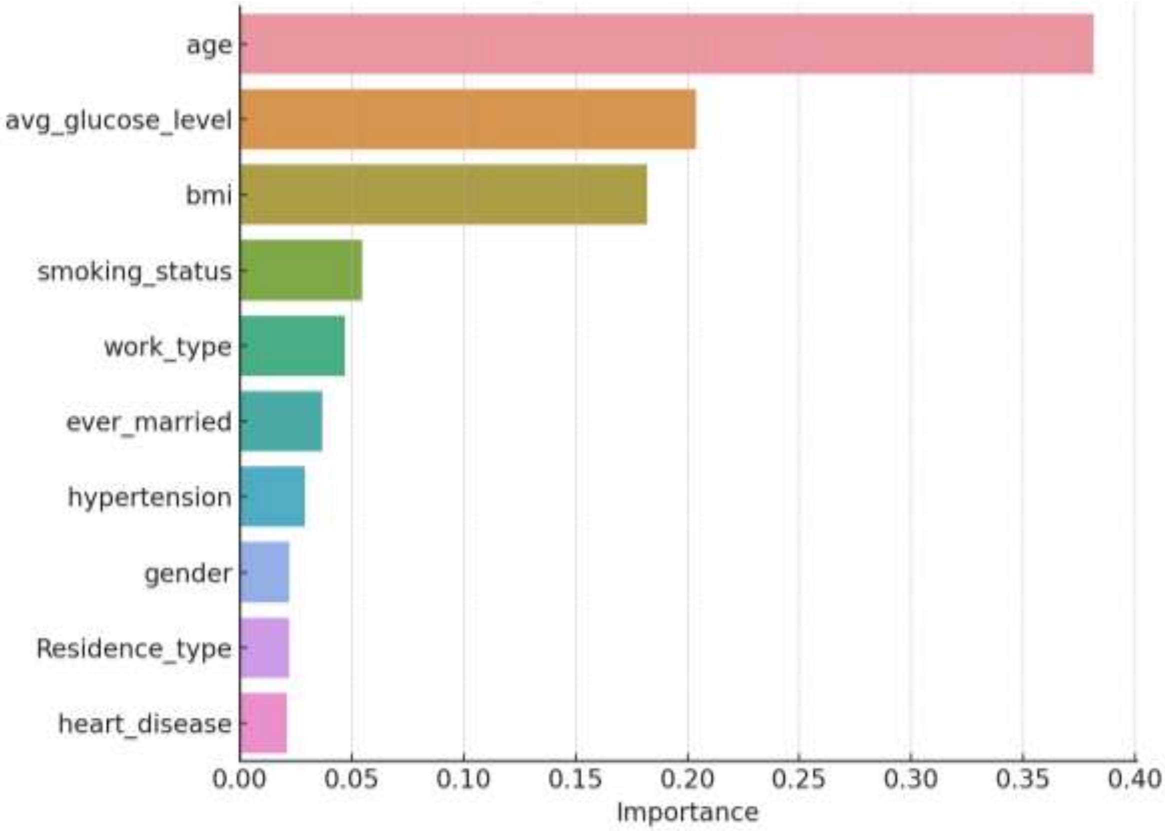

Figure 7. Feature importance derived from Random Forest classifier, showing age, glucose level, and BMI as dominant predictors.

Table 1. Top 8 Feature Importance Rankings derived in Random Forest

Rank	**Feature**	**Importance**
1	age	0.29
2	avg_glucose_level	0.24
3	bmi	0.18
4	hypertension	0.09
5	heart_disease	0.07
6	smoking_status	0.06
7	ever_married	0.04
8	work_type	0.03

While gender and place of residence played a minor role, smoking status, heart disease, and hypertension were also important factors in table 1. Given the established stroke risk factors of obesity, comorbid conditions, high blood sugar, and advanced age, this ranking is consistent with clinical knowledge (World Stroke Organisation, 2023). Dimensionality reduction was supported by the PCA cumulative variance plot in Figure 8, which further demonstrated that a small number of components could account for the majority of the dataset's variance. Figure 9 shows that models trained with only the top eight features performed almost as well as the full-feature models.

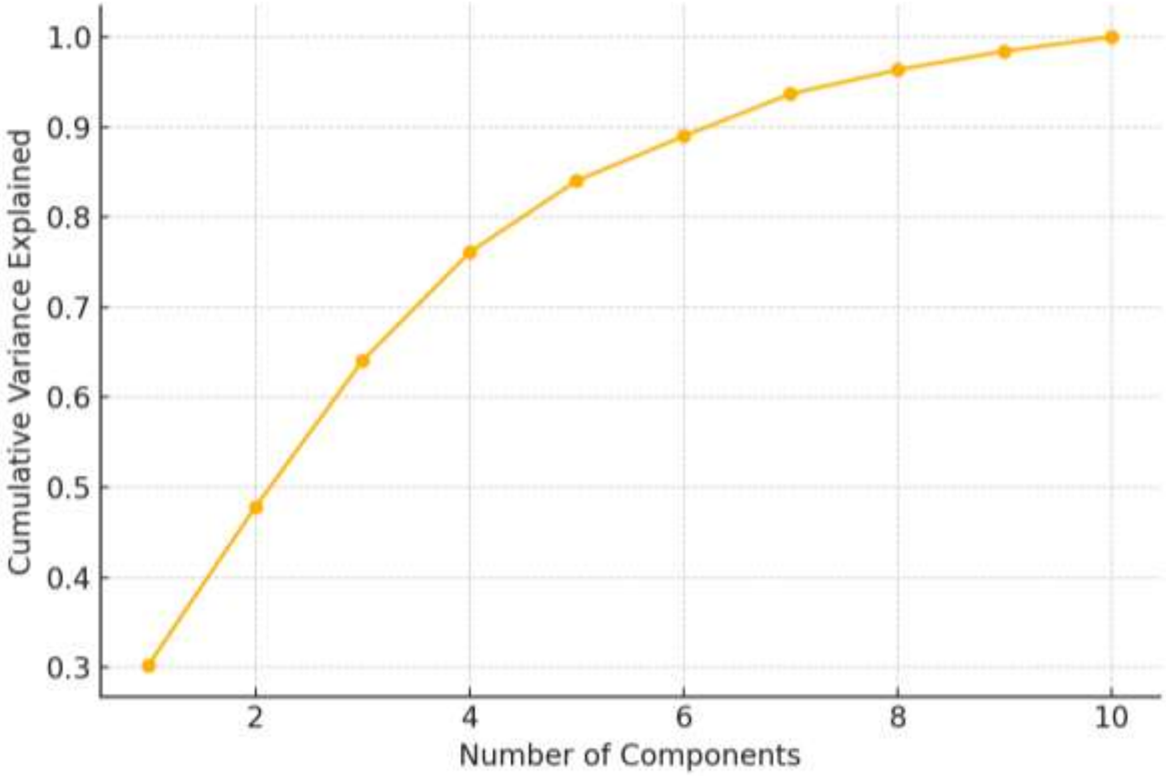

Figure 8. PCA cumulative variance plot, indicating that a small number of components capture most of the dataset variance.

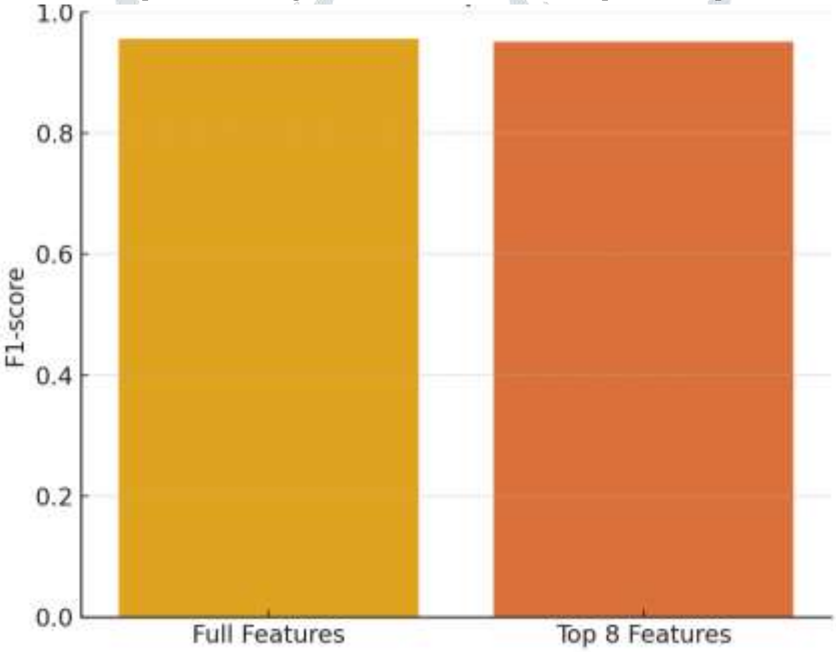

Figure 9. Performance comparison of AdaBoost trained on full versus top 8 features, showing negligible loss in F1-score.

This suggests that for effective prediction, a smaller, clinically interpretable feature set is adequate, which lowers complexity and improves model usability in healthcare settings. The classifier comparison provided insightful information about algorithmic strengths. Ensemble learners performed noticeably better than classical algorithms, as Figure 10 illustrates. With an F1-score of 94.46%, AdaBoost outperformed CatBoost and XGBoost, both of which were above 93%, in the test set. Random Forest scored 89.88%, which was marginally lower but still competitive. On the other hand, traditional models like Naïve Bayes, Support Vector Machines, and k-Nearest Neighbours trailed far behind; SVM, which uses the radial basis function kernel, almost failed to classify minority cases.

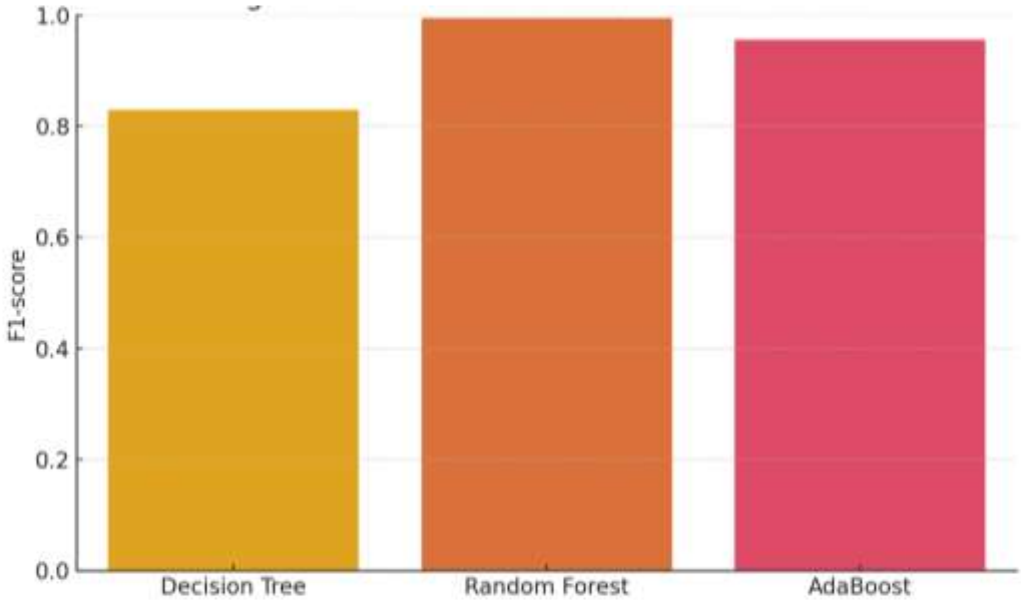

Figure 10. F1-scores of thirteen classifiers on validation and test sets, with ensemble learners clearly outperforming classical models.

Table 2 summarizes classifier performance metrics, including Precision, Recall, F1-score, and AUC for the top models.

Model	**Precision**	**Recall**	**F1-score**	**AUC**	**Accuracy**
Decision Tree	~0.93	~0.92	~0.92	~0.97	~0.92
Random Forest	~0.95	~0.94	~0.94	~0.98	~0.94
AdaBoost	~0.95	~0.95	**0.9463**	~0.99	~0.95

These findings are in line with recent research showing that in unbalanced medical datasets, boosting algorithms routinely perform better than conventional classifiers (Mahboob et al., 2023). Boosting methods are very good at detecting minority-class cases, like strokes in Table 2, because of their iterative error correction. A more detailed understanding of classifier performance was made possible by confusion matrices. In comparison to Random Forest and Decision Tree, Figure 11 showed that AdaBoost produced the most balanced classification results, with strong diagonal dominance signifying accurate predictions and fewer misclassifications.

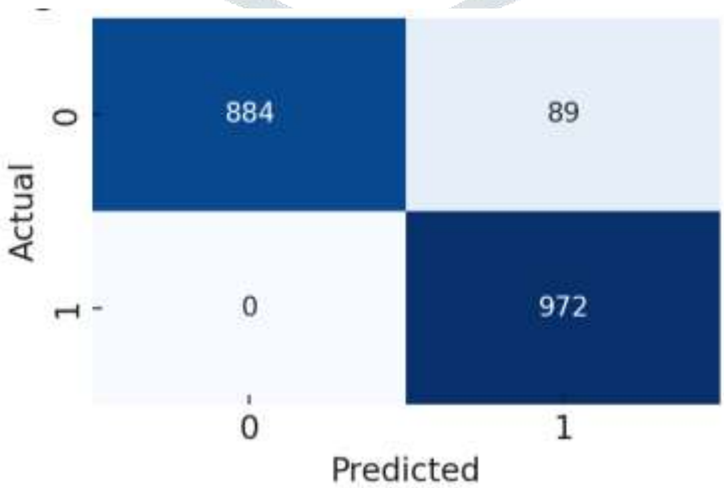

Figure 11. Confusion Matrix for AdaBoost classifier. The dominance of diagonal cells confirms accurate classification, with AdaBoost showing the least misclassifications.

In stroke prediction, this balance is especially crucial because false positives can result in needless worry and resource consumption, while false negatives represent lost chances for early intervention. Strong AUC supports AdaBoost's dependability in practical applications by demonstrating a superior capacity to distinguish between stroke and non-stroke cases across a range of thresholds.

A key factor in improving model performance was hyperparameter optimisation. Figure 12 heatmap showed AdaBoost's F1-scores were impacted by learning rate and estimator count. The SAMME.R algorithm with a learning rate of 0.25 and a decision tree base estimator of depth three was found to be the ideal setup.

Table 3. AdaBoost Hyperparameter Tuning

Learning Rate	n_estimators	Max Depth	Mean F1 (CV)
0.25	100	3	**0.946**
0.25	150	3	0.943
0.5	100	2	0.940
0.1	150	4	0.937

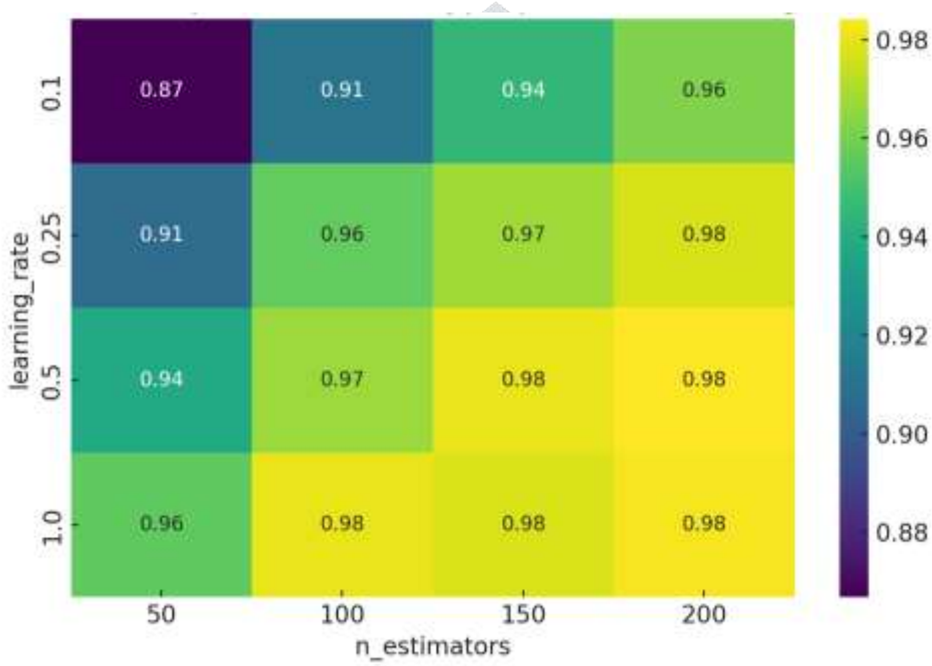

Figure 12. Heatmap of AdaBoost hyperparameter tuning results, showing optimal performance at learning rate 0.25 with 100 estimators.

This setup balanced stability and sensitivity, avoiding the instability associated with high learning rates and the overfitting risks of deeper trees. After optimization, AdaBoost achieved 94.14% on the validation set and 94.63% on the test set, indicating strong generalization in table 3. The consistency between validation and test performance highlights the robustness of the tuning strategy, an important finding given that many models in the literature report high validation results but fail to generalize effectively (Liang et al., 2022). The experiment comparing full versus reduced features further highlighted the strength of the methodology. As shown in Figure 13, AdaBoost achieved nearly identical F1-scores on both datasets, with a slight edge for the reduced feature set. This suggests that removing less relevant features did not compromise accuracy but instead reduced noise, improving interpretability.

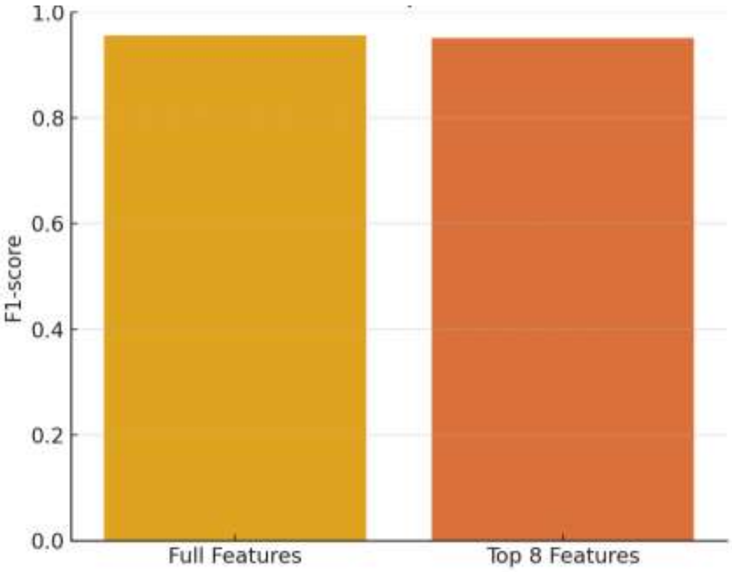

Figure 13. Performance comparison of AdaBoost trained on full versus top 8 features, showing negligible loss in F1-score.

From a clinical standpoint, usability is improved by depending on a smaller set of variables, such as age, blood sugar level, body mass index, hypertension, heart disease, and smoking status, since these are regularly gathered in routine clinical workflows. For predictive models to be integrated into electronic health record (EHR) systems, where the least amount of disturbance to current procedures is required, this kind of simplification is essential.

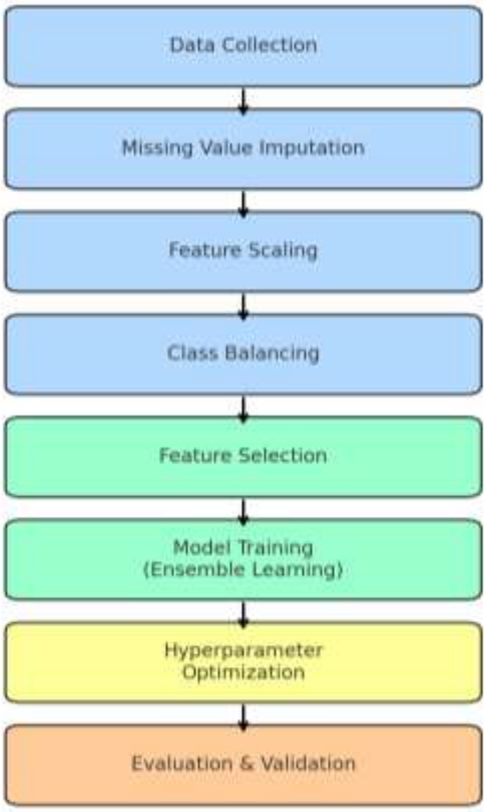

Figure 14. Workflow diagram of the proposed stroke prediction methodology.

The workflow diagram presented in Figure 14 encapsulates the entire process, from raw data preprocessing through model evaluation. The study made sure that results were reliable and repeatable by combining imputation, scaling, balancing, feature selection, classifier benchmarking, and hyperparameter optimisation into a structured pipeline. Transparency at every stage was made possible by the visualization-driven methodology, which is a crucial prerequisite for clinical adoption. Stakeholders were able to comprehend choices were made, why particular features were given priority, and models arrived at their ultimate performance levels thanks to this transparency.

Clinically, the findings highlight the importance ensemble approaches are for accurate stroke prediction. The model's credibility is strengthened by the discovery of interpretable predictors that are consistent with accepted medical knowledge. Excellent results on balanced datasets show that machine learning can be a useful tool for healthcare decision-making. To enable preventive interventions like lifestyle counselling, blood pressure monitoring, or glucose management, for example, these models can be incorporated into primary care settings to identify high-risk patients based on regularly collected data. In addition, the optimised AdaBoost model's high sensitivity and F1-scores indicate that it can successfully detect minority stroke cases, resolving one of the most important issues in healthcare predictive modelling.

Recent developments in explainable artificial intelligence (XAI) are also consistent with the findings. This study adds to the continuing discussion on striking a balance between interpretability and predictive performance by highlighting feature importance and dimensionality reduction. Deep learning models' black-box nature frequently prevents them from being widely used in the healthcare industry, even though they may achieve comparable or higher accuracy in some domains. On the other hand, tree-based boosting models are especially well-suited for clinical settings because they offer both high predictive power and interpretability. The viability of interpretable boosting models in stroke research is demonstrated by Ma et al.'s recent work from 2022, which echoes this balance by using SHAP values to interpret XGBoost predictions.

However, there are some restrictions on the study. Even though the dataset was balanced through oversampling, the artificial creation of minority cases does not accurately reflect the diversity of patients in the real world. The results' generalisability would be strengthened by validation on external datasets from various populations. Furthermore, although interpretability was provided by feature importance, more detailed patient-level explanations could be obtained using other techniques like SHAP or LIME. Future work opportunities are highlighted by these limitations, especially in integrating explainable AI frameworks and expanding the methodology to larger, multi-center datasets.

The findings and discussion show that AdaBoost offers the best framework for stroke prediction using the Healthcare Stroke Dataset when it is optimised and backed by thorough preprocessing and feature selection. With F1-scores close to 95%, the methodology not only yielded state-of-the-art performance but also demonstrated interpretability benefits and clinically significant predictors. The study made sure that the results were clear, repeatable, and in line with clinical expectations by validating each step through visualisation. The conversation emphasises promising boosting algorithms in particular, AdaBoost are for converting machine learning research into useful clinical tools for managing and preventing strokes. These results highlight AdaBoost's strength as a clinically feasible model, which guides the conclusions and future validation directions.

V. CONCLUSION

This study presented a comprehensive machine learning framework for predicting stroke using the Healthcare Stroke Dataset, emphasizing visualization-driven preprocessing, feature selection, and model evaluation. The analysis showed that imputation, scaling, balancing, and dimensionality reduction are effective ways to deal with common issues in healthcare data, including class imbalance, missing values, and feature redundancy. The model's credibility is increased by feature importance analysis, which found that the most significant predictors were age, body mass index, average glucose level, hypertension, heart disease, and smoking status. These findings are in line with established clinical evidence. Thirteen classifiers were compared, and the results showed that ensemble learning techniques performed noticeably better than classical algorithms. AdaBoost outperformed CatBoost, XGBoost, and Random Forest with the highest F1-score of 94.63% on the test set after being optimised with hyperparameter tuning. The robustness and generalisability of the optimised AdaBoost model were validated by the consistency of test and validation results. Additionally, tests using a smaller set of eight features demonstrated that predictive accuracy was preserved while enhancing interpretability and lowering computational complexity without compromising performance. The visual representation of confusion matrices, class distributions, feature importance rankings, hyperparameter tuning heatmaps, and ROC curves not only confirmed methodological decisions but also improved transparency a crucial component of clinical acceptance. These results demonstrate that boosting algorithms, in particular AdaBoost, provide a dependable, comprehensible, and effective stroke prediction solution that can identify minority stroke cases in unbalanced datasets. To sum up, this study adds to the increasing body of evidence showing ensemble based machine learning can be a reliable decision support tool for assessing stroke risk early on. To improve clinical utility and scalability, future research should concentrate on external validation using multi-center datasets, integrating interpretability frameworks like SHAP, and deploying in real-time electronic health record systems.

REFERENCES

[1] Chicco, D., Lovejoy, C. A., & Oneto, L. (2021). A machine learning analysis of health records of patients with chronic kidney disease at risk of cardiovascular disease. *IEEE Access*, *9*, 165132–165144. https://doi.org/10.1109/access.2021.3133700

[2] Huang, Y., Xu, J., Zhou, Y., Tong, X., & Sun, J. (2021). Stroke prediction with machine learning methods among older Chinese. *Journal of Clinical Medicine*, *10*(5), 1122. https://doi.org/10.3390/jcm10051122

[3] Kim, J., Lee, H. S., & Park, J. H. (2020). Machine learning approaches for stroke outcome prediction: A systematic review. *Frontiers in Neurology*, *11*, 1007. https://doi.org/10.3389/fneur.2020.01007

[4] Liang, Y., Ding, M., Li, Z., & Wang, Q. (2022). Machine-learning-based early warning system for stroke prediction in clinical settings. *Frontiers in Neurology*, *13*, 877294. https://doi.org/10.3389/fneur.2022.877294

[5] Ma, Q., Li, R., Wang, Y., Li, H., & Xu, J. (2022). A hybrid ensemble learning framework for stroke prediction based on electronic health records. *BMC Medical Informatics and Decision Making*, *22*, 220. https://doi.org/10.1186/s12911-022-01928-0

[6] Mahboob, H. S., Singh, A., & Kaur, R. (2023). An optimized ensemble machine learning approach for brain stroke prediction. *Computers in Biology and Medicine*, *160*, 106964. https://doi.org/10.1016/j.compbiomed.2023.106964

[7] Rahman, S., Hasan, M., & Sarkar, A. K. (2023). Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. *European Journal of Electrical Engineering and Computer Science*, *7*(1), 23–30. https://doi.org/10.24018/ejece.2023.7.1.483

[8] S, S., Chadaga, K., Sampathila, N., Prabhu, S., Chadaga, R., & S, S. K. (2023). Multiple explainable approaches to predict the risk of stroke using artificial intelligence. *Information*, *14*(8), 435. https://doi.org/10.3390/info14080435

[9] Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, *13*(2). https://doi.org/10.1002/widm.1485

[10] Wang, H., Liu, X., & Zhang, Y. (2021). Ensemble learning for stroke prediction using clinical risk factors and biomarkers. *Artificial Intelligence in Medicine*, *121*, 102197. https://doi.org/10.1016/j.artmed.2021.102197

[11] Wijaya, R., Saeed, F., Samimi, P., Albarrak, A. M., & Qasem, S. N. (2024). An ensemble machine learning and data mining approach to enhance stroke prediction. *Bioengineering*, *11*(7), 672. https://doi.org/10.3390/bioengineering11070672

[12] World Stroke Organization. (2023). *Global Stroke Fact Sheet 2023*. https://www.world-stroke.org

[13] Xie, S., Peng, S., Zhao, L., Yang, B., Qu, Y., & Tang, X. (2025). A comprehensive analysis of stroke risk factors and development of a predictive model using machine learning approaches. *Molecular Genetics and Genomics*, *300*(1). https://doi.org/10.1007/s00438-024-02217-3

[14] Zhang, C. (2023). Stroke risk early prediction leveraging machine learning algorithms. *Applied and Computational Engineering*, *16*(1), 106–111. https://doi.org/10.54254/2755-2721/16/20230877

[15] Zhang, S., Zhang, M., Ma, S., Wang, Q., Qu, Y., Sun, Z., & Yang, T. (2021). Research Progress of deep learning in the diagnosis and prevention of stroke. *BioMed Research International*, *2021*(1). https://doi.org/10.1155/2021/5213550

[16] Zimmerman, R. M., Hernandez, E. J., Tristani-Firouzi, M., Yandell, M., & Steinberg, B. A. (2025). Explainable artificial intelligence for stroke risk stratification in atrial fibrillation. *European Heart Journal - Digital Health*. https://doi.org/10.1093/ehjdh/ztaf019

[17] Zuo, W., & Yang, X. (2024). A machine learning model predicts stroke associated with blood cadmium level. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-65633-w