ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

Optimising Stroke Recurrence Prediction Using Minimal Clinical Features and Machine Learning Models

Diri, Ezekiel Ebere[1]; Diri, Grace Oluchi[2]; Rita Chikeru Owhonda[3]; Nbaakee, Lebari Goodday[4]; Unula Godknows[5]; Kingsley Theophilus Igulu[6]

[1]Department of Computer Science, Birmingham City University, United Kingdom
[2,4,6]Department of Computer Science, Ignatius Ajuru University of Education, Nigeria
[3]Department of Accounting, Finance and Economics, Bournemouth University, United Kingdom
[5]University of Hertfordshire Business School, Hertfordshire University of Hertfordshire, United Kingdom

Publication Date: 2025/09/18

Abstract: Stroke recurrence remains one of the most devastating challenges in managing cerebrovascular disease, adding to disability, mortality, and rising healthcare costs worldwide. Being able to predict recurrence early could mean the difference between timely intervention and irreversible outcomes. In this study, we explored whether machine learning models - Logistic Regression, Random Forest, and XGBoost - could predict recurrence risk using only a small set of routine clinical features. Preprocessing involved managing missing values, scaling variables, and applying SMOTE to balance the classes without distorting real patient patterns. Models were evaluated across accuracy, precision, recall, F1 Score, and AUC-ROC, with greater weight placed on recall and F1 given the clinical need to minimize missed recurrences. Random Forest delivered the strongest results, achieving an accuracy of 92.39%, a recall of 94.05%, an F1 Score of 92.56%, and an AUC-ROC of 97.04%. These findings suggest that even simple, carefully designed predictive models could offer real clinical value, particularly in healthcare environments where rich data resources are limited and early warnings could make a critical difference for patient care.

Keywords: *Stroke Recurrence Prediction, Machine Learning Models, Random Forest Classifier, Minimal Clinical Features, Secondary Stroke Prevention.*

How to Cite: Diri, Ezekiel Ebere; Diri, Grace Oluchi; Rita Chikeru Owhonda; Nbaakee, Lebari Goodday; Unula Godknows; Kingsley Theophilus Igulu (2025). Optimising Stroke Recurrence Prediction Using Minimal Clinical Features and Machine Learning Models. *International Journal of Innovative Science and Research Technology*, 10(9), 780-794. https://doi.org/10.38124/ijisrt/25sep706

I. INTRODUCTION

Stroke remains one of the heaviest health burdens worldwide, with incidence rates climbing steadily and mortality risks still alarmingly high. The strain falls hardest on low and middle-income countries, where access to preventive care is often patchy at best. Vulnerable populations bear the brunt, while health systems already stretched thin struggle to keep up. Researchers, practitioners, and policymakers cannot afford to look away from the growing urgency of this crisis. Recent projections are sobering. Deaths from stroke could climb by 50% by 2050, nearing 10 million each year. Survivors often live with long-term disability, creating ripple effects that touch families, communities, and economies [1]. These numbers alone tell a story, but they barely capture the personal devastation hidden behind them. If preventive strategies are not dramatically improved, the suffering will be even worse than the statistics suggest.

One area demanding urgent focus is the problem of recurrent strokes. Patients who make it through a first event face much higher odds of a second, and these follow-up strokes tend to hit harder. Recovery grows tougher with each new event, and the chances of regaining lost function shrink. It's a grim cycle, and breaking it depends on finding better ways to predict and prevent recurrence before patients fall into it.

From a clinical and economic standpoint, recurrent strokes are disastrous. Direct medical costs surge with each incident, but the indirect costs (lost work, long-term care, lowered quality of life) can quietly eclipse even hospital bills. Traditional risk prediction tools have not kept pace. They often depend on complicated scoring systems that demand

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

detailed clinical inputs, making them tough to use in the real world, especially across varied patient groups [2]. Collecting all that information takes time, and the models built on it tend to miss the complex, tangled relationships between risk factors. Stroke risk is rarely a straight line. It bends and twists through biology, lifestyle, environment, and luck. Trying to pin it down with rigid, linear models leaves gaps wide enough for too many patients to fall through.

Lately, a different path has been opening up. Machine learning offers tools that can handle complexity without getting bogged down by it. With the right algorithms, models can pick up faint warning signs hidden in just a handful of clinical features. Instead of drowning doctors in data, these approaches aim to simplify, making predictive tools faster, more accurate, and easier to deploy even in places with limited resources. What makes this shift even more promising is how adaptable machine learning tools can be. Whether in a high-tech urban hospital or a rural clinic where internet access is spotty, flexible models could help close the gap between who gets preventive care and who does not [3]. No technology will fix global health disparities overnight, but better prediction is one step toward giving every patient a fairer shot.

More researchers are leaning into this idea. Reducing model complexity without losing predictive strength has become a new priority. When clinical data can be distilled into clear, actionable insights, healthcare providers have a better chance of identifying high-risk patients early and stepping in before disaster strikes [4], [5].

Driven by this urgent need, the current study sets out to evaluate machine learning models that rely on smaller, more manageable sets of clinical features. The goal is practical: to build prediction tools that actually work in the messy, unpredictable world of real clinical practice. Through testing and validation across different patient groups, we hope to develop models that not only predict recurrence more accurately but also fit into the real rhythms of healthcare. Because if the tools cannot reach the patients who need them, they do not matter.

II. RELATED LITERATURE

Predicting stroke recurrence remains one of the most pressing challenges in clinical research, given the enormous health burdens it continues to impose. Recurrent strokes not only increase patient morbidity and mortality but also amplify the financial strain on already stretched healthcare systems. Achieving accurate prediction is far from simple, especially when clinical datasets are large, complex, and packed with high-dimensional features that make models harder to interpret and apply in practice. Recent work has been moving toward machine learning (ML) approaches that rely on minimal clinical features, aiming to balance predictive strength with real-world usability [6], [7].

Across the past few years, a noticeable shift has taken place. Researchers are increasingly turning to ML models that focus on a smaller set of clinical variables while trying to maintain robust prediction. Different algorithms have been explored, each bringing a different set of strengths and challenges. Logistic regression still appeals for its simplicity and interpretability [8], while methods like random forests and gradient boosting offer strong performance with built-in resistance to overfitting [9], [10]. Neural networks have been promising in mapping non-linear relationships, but often at the cost of transparency, which can limit clinician trust [11]. Commonly selected clinical predictors across studies include patient age, blood pressure, medical history details like diabetes and smoking, medication use, especially anticoagulants, and NIHSS scores, which summarise neurological status [12], [13], [14].

The field has clearly evolved beyond the early dependence on traditional statistical methods. Logistic regression and similar approaches once dominated the scene, linking clinical variables to stroke risk in a straightforward manner [15]. More recently, researchers have leaned heavily into machine learning, especially as models like random forests, support vector machines, and boosting methods have shown a greater ability to navigate complex, non-linear data [16], [17], [18].

Studies focusing specifically on stroke recurrence have seen strong predictive results even from relatively simple models. Some models have reported accuracy rates crossing 90%, suggesting that keeping things simple can still yield highly relevant clinical insights [19] [18], [20], [21]. For instance, [16] compared different methods and found that random forests consistently outperformed logistic regression, reinforcing the idea that smart, streamlined models might better serve clinical needs.

Looking at recent datasets, [6] examined predictive scores like Essen and SPI-II across a sample of 1,550 patients. They found AUC scores hovering around 0.63 - not terrible, but not the level needed for confident clinical use. Similarly, [22] studied the ABCD2 score in TIA patients, reporting AUCs between 0.592 and 0.683. These results show that while traditional scoring systems still have a role, their ceiling is visible, and newer approaches are needed to push beyond it.

When comparing approaches, the story becomes even more layered. [23] showed that machine learning models could outperform conventional scores, but only when large, well-curated datasets were available. On the other hand, [24] applied an interpretable ML framework combining plaque burden and demographic data, achieving an AUC of 0.832 - a major leap forward using a surprisingly lean feature set. These contrasting results speak to an ongoing tension between model complexity and clinical practicality, a balance that is critical if these tools are ever going to become everyday aids for clinicians.

Feature selection remains a huge part of this story. Researchers like [25] have shown that biological markers, particularly atherosclerotic plaque characteristics, can serve as effective minimal predictors for recurrence. Dimensionality reduction techniques have also gained

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

traction, with studies like [26] highlighting the predictive value of cerebral microbleeds. The shift toward focusing on a few well-chosen features seems not just pragmatic but necessary, especially if models are to remain both interpretable and actionable in clinical settings.

That said, the field still faces several stumbling blocks. Many studies rely heavily on high-dimensional feature sets, which can make generalisation across patient groups difficult. [27] warned that without expanding external validation, models risk becoming brittle - strong within one dataset, but weak anywhere else. Similarly, even highly optimised models, like those built by [28], cannot assume success outside their original cohorts without rigorous testing. Finding the right balance between accuracy and usability remains a missing piece in the current research landscape.

Dataset characteristics vary widely across studies. Some researchers work with just a handful of features; others push datasets with fifty variables or more [18]. Evaluation metrics also differ, though accuracy, precision, recall, F1-score, and AUC remain the standards [21], [35]. Feature selection methods, including chi-square testing and recursive feature elimination, have been essential for cutting down complexity without sacrificing prediction quality. Some groups have gone further by using advanced sampling techniques like SMOTE to tackle class imbalance in stroke datasets, enhancing both accuracy and fairness [31], [32].

Despite progress, real gaps remain. There's still too much dependence on large, messy feature sets that don't easily translate to clinical workflows [18], [29]. Researchers are only just starting to give real attention to minimalist models that prioritise simplicity and clinical relevance [30], [21].

Validation across different patient populations also continues to lag. Clinicians are rightly cautious about relying on models that have not been stress-tested outside controlled environments. Tools like SHAP are gaining popularity for improving model transparency [33], helping clinicians understand not just what a model predicts, but why it makes those predictions.

Real-world deployment, especially in low-resource settings, presents another obstacle. Even the most accurate model means little if it cannot be easily implemented. Solutions need to be not just effective, but accessible, fitting into busy clinics without requiring perfect data or expensive infrastructure [19], [34].

Although machine learning has already reshaped the field of stroke recurrence prediction, much of the work feels like the beginning rather than the end. Many studies show promise, but few have built the kind of models that can easily move from theory into practice. Future research needs to focus more sharply on developing models that are not only accurate but also clinically viable, validated, and interpretable - tools that genuinely help providers and patients navigate the uncertainties of stroke risk.

III. METHODOLOGY

Our study adopts a step-by-step approach built to support accurate and interpretable predictions of stroke recurrence. It begins with data acquisition, followed by a thorough preprocessing phase to address any quality issues. After that, we move into careful feature engineering and model development, making sure each step builds cleanly on the last. The overall workflow is captured in the architecture diagram below, which maps out the journey from initial data collection all the way through to model evaluation.

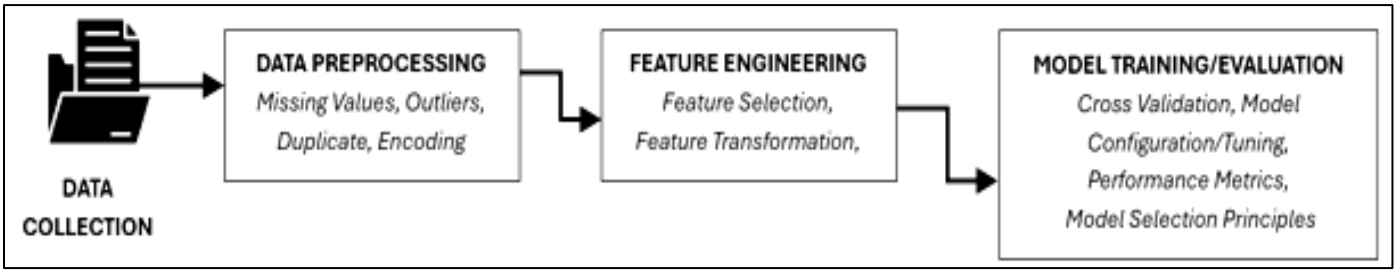

Fig 1 Architecture Diagram Representing the Stroke Recurrence Prediction Workflow

A. Data Collection

We based our study on a secondary dataset containing anonymized electronic health records from patients assessed for stroke events and recurrence. The dataset was drawn from a publicly accessible clinical repository that follows established privacy standards and ethical data-sharing practices, aligning with FAIR principles (Findable, Accessible, Interoperable, and Reusable). Even though the records were already de-identified, we maintained careful attention to patient confidentiality, responsible data use, and research reproducibility throughout the project.

The dataset includes over 5,000 individual patient records, covering demographic details, medical histories, lifestyle factors, and stroke status. For this study, we focused only on adult patients aged 18 and older who had complete outcome data. Records missing outcome labels or critical predictor information were removed during preprocessing, leaving a final sample of 4,769 valid cases.

Although the original source did not provide a clear description of its sampling strategy, exploratory analysis showed a fairly broad distribution of patients across different socioeconomic and geographic backgrounds. This diversity strengthens the relevance of the findings to wider

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

populations, but the absence of random sampling does leave some open questions about how easily the results might generalize beyond this dataset.

B. Data Preprocessing

We applied a multi-stage preprocessing pipeline to strengthen data reliability and support better model performance. The process involved handling missing values, managing outliers, encoding categorical variables, normalizing numerical features, and tackling the class imbalance present in the target variable. Each step was designed to prepare the data carefully without introducing unnecessary complexity or losing important information along the way.

➢ *Handling Missing Values:*

An audit of the dataset showed missing entries scattered across both numerical and categorical variables. Among the clinical features, BMI stood out with the highest number of missing values. For numerical variables, we chose to impute missing values using the median, given its robustness against skewed distributions and outliers. On the other hand, missing categorical variables like work_type and smoking_status were filled using the mode to maintain the most common classifications and preserve data consistency. By using this two-pronged imputation approach, we made sure the original distribution patterns stayed intact. This helped prevent the introduction of artificial trends that could have distorted model training later on.

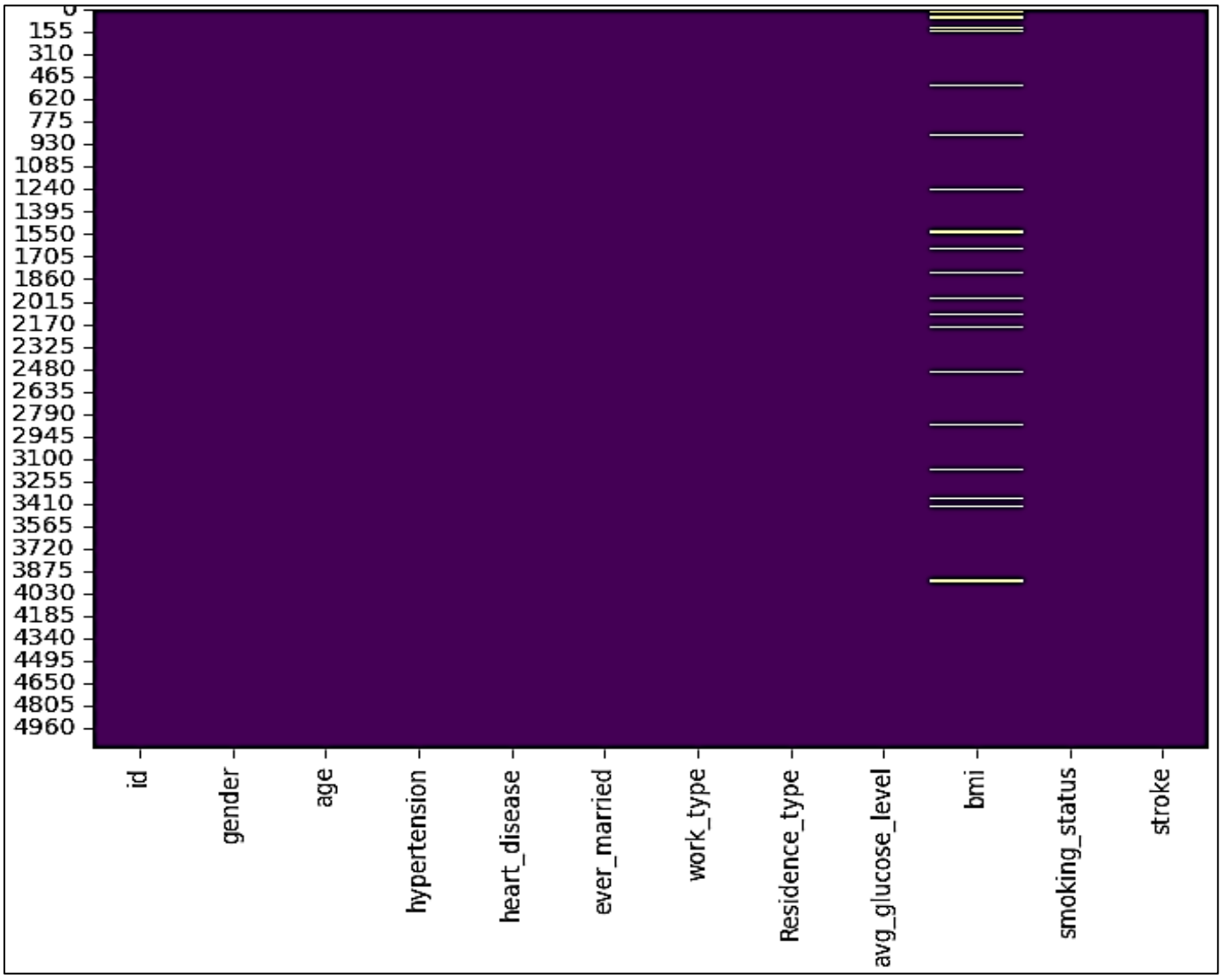

Fig 2 Heatmap of Missing Values Across Dataset Features

➢ *Outlier Detection and Validation:*

Outlier detection was carried out using visual diagnostics, focusing on boxplots of the main continuous variables (age, avg_glucose_level, and bmi). These visual tools helped us spot extreme values that might have stemmed from measurement errors or reflected unusual health profiles. Although a handful of observations fell beyond the interquartile range, we chose to retain them to preserve the natural heterogeneity of the patient population.

This decision was based on the idea that extreme cases often carry important clinical meaning. Excluding them could have weakened the ecological validity of the model and shifted it toward predicting outcomes only for more typical patients, which was not the goal.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

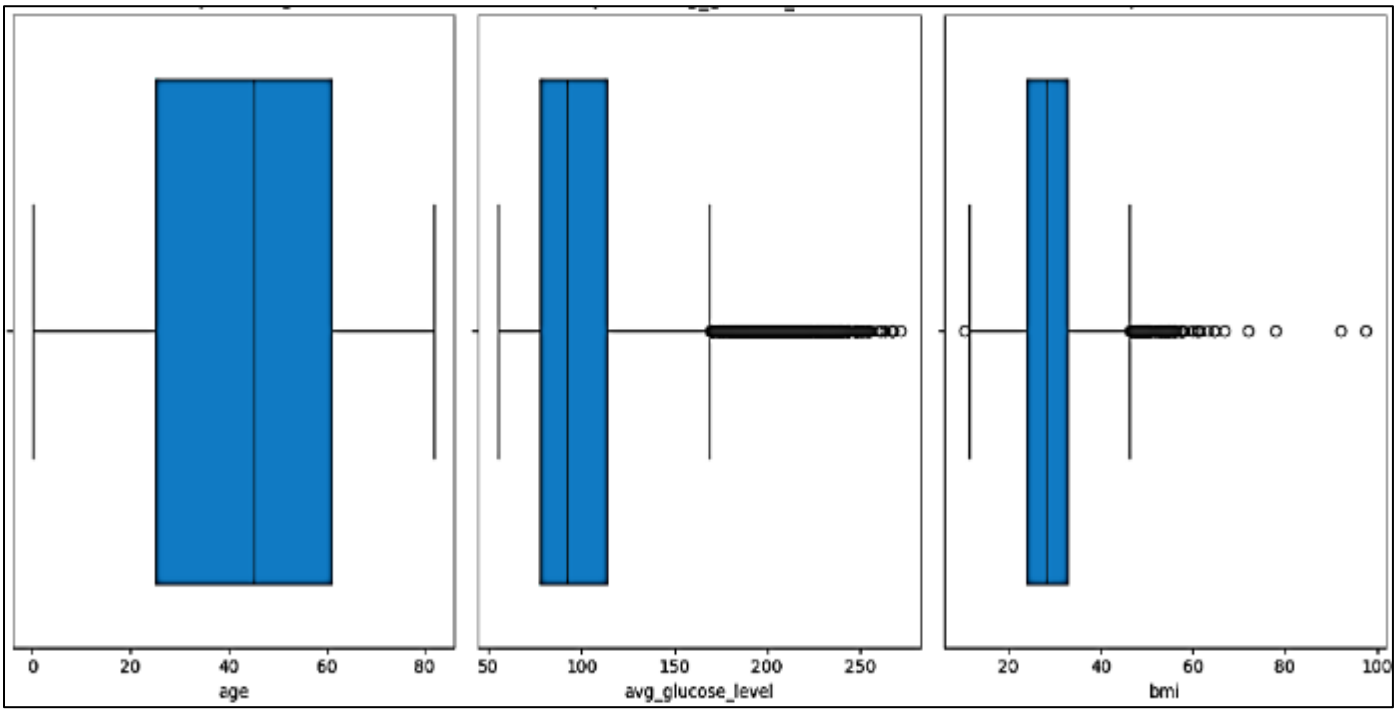

Fig 3 Boxplots of Numerical Features (Age, Glucose, BMI)

- The primary outcome variable, stroke recurrence, showed a strong imbalance, with non-recurrence cases far outweighing recurrence cases. This posed a real risk during model training, where high overall accuracy could mask poor sensitivity in detecting true recurrence events. To tackle the issue, we applied the Synthetic Minority Over-sampling Technique (SMOTE). Unlike simple duplication, SMOTE creates new synthetic examples of the minority class by interpolating between existing observations with similar feature profiles. This helped rebalance the dataset without disrupting its overall structure or reducing its diversity. Crucially, SMOTE was applied only to the training set after a stratified train-test split, carefully avoiding any data leakage that could have artificially inflated model performance.

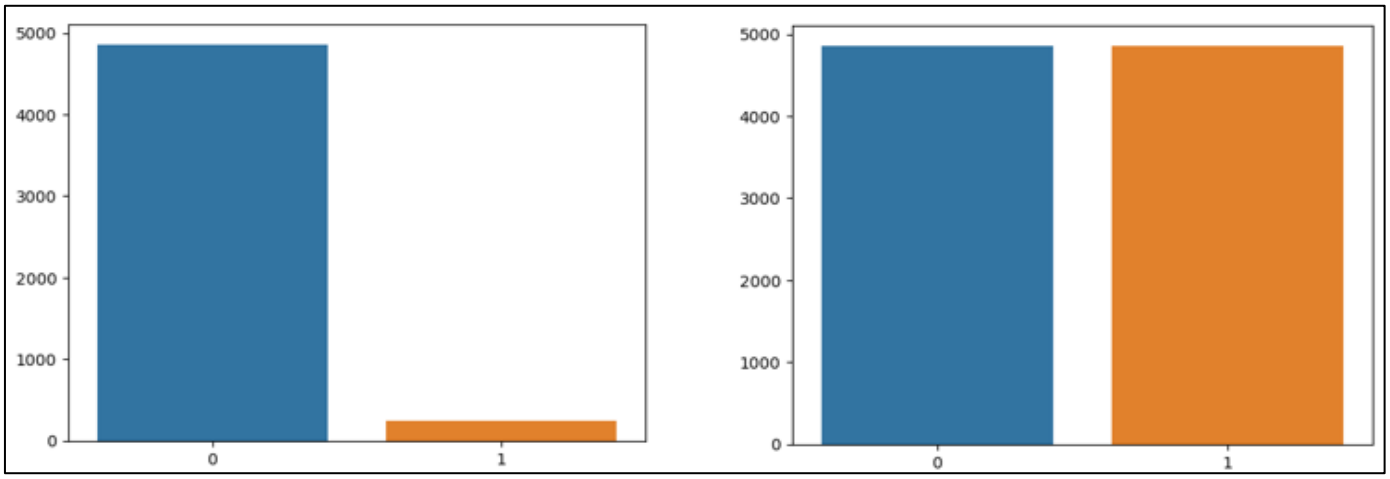

Fig 4 Class Distribution of Stroke Before and After SMOTE Application

C. Feature Engineering

Feature engineering combined domain knowledge with data-driven insights to boost both model interpretability and efficiency. We focused on selecting meaningful predictors, encoding categorical variables thoughtfully, and scaling numerical features to ensure a balanced contribution across the model.

- *Feature selection was guided by a focus on clinical relevance and practical simplicity. We deliberately chose a reduced set of predictors that not only align with known stroke risk factors but also support real-world deployment, especially in clinical environments where access to extensive data might be limited. The final feature set included age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, BMI, and smoking_status.*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

- *Selection Decisions Drew from Three Sources:*

Established findings in stroke-related research, clinical plausibility based on patient experience, and insights gained from a correlation matrix built using a numerically encoded version of the dataset. The matrix also helped flag potential issues with multicollinearity, ensuring that the final set of predictors remained both meaningful and efficient for modelling.

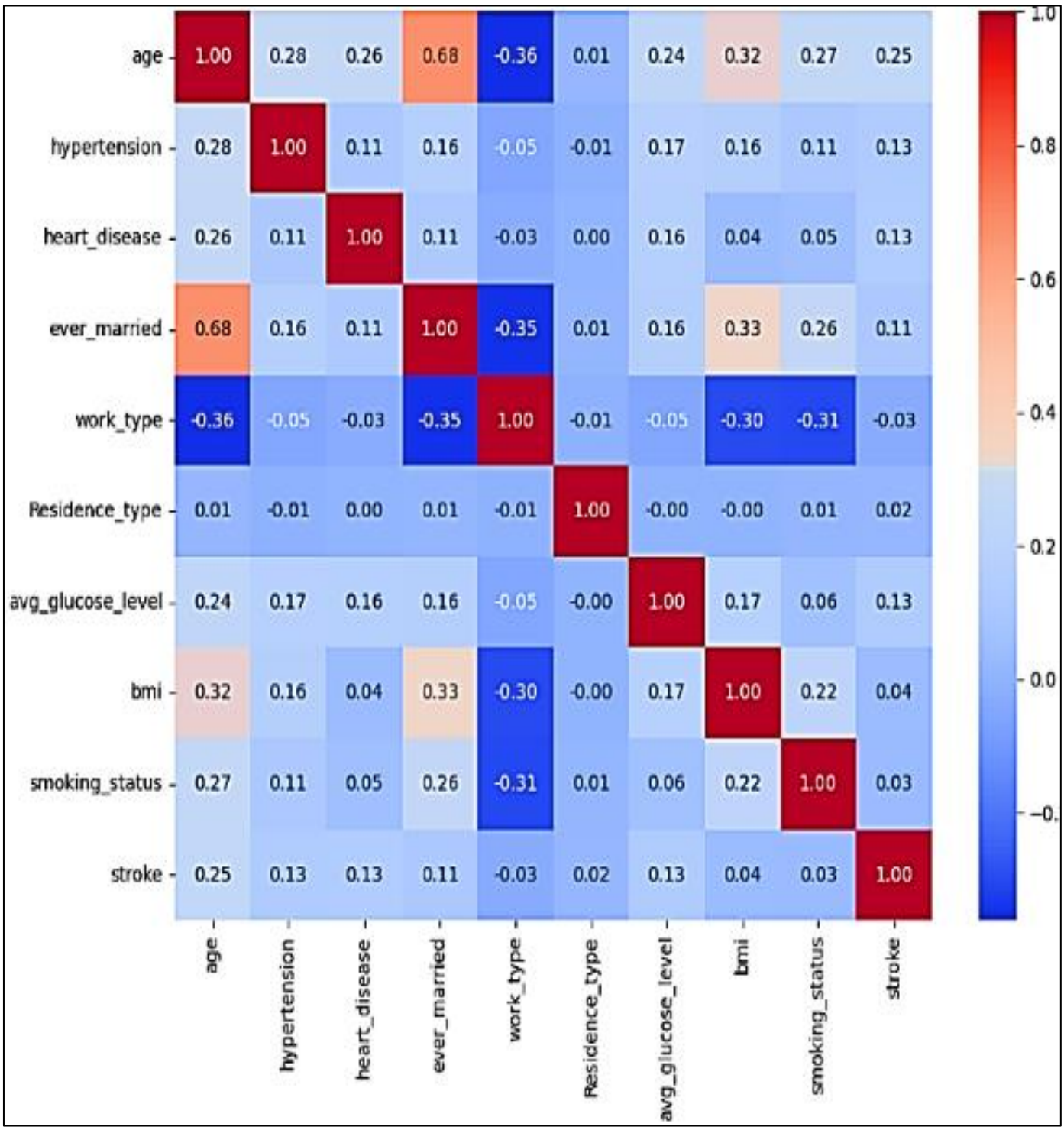

Fig 5 Correlation Matrix of Selected Features and Target

➢ *Feature Transformation:*

Categorical variables were transformed using one-hot encoding, converting each category into binary vectors without implying any order among the classes. This approach preserved interpretability and ensured compatibility with both tree-based and linear models. For the continuous variables (age, avg_glucose_level, and bmi) we applied standardization using the StandardScaler from Scikit-learn. This process normalized each feature to have a mean of zero and a standard deviation of one, helping to ensure that no single variable dominated the model simply because of its scale.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

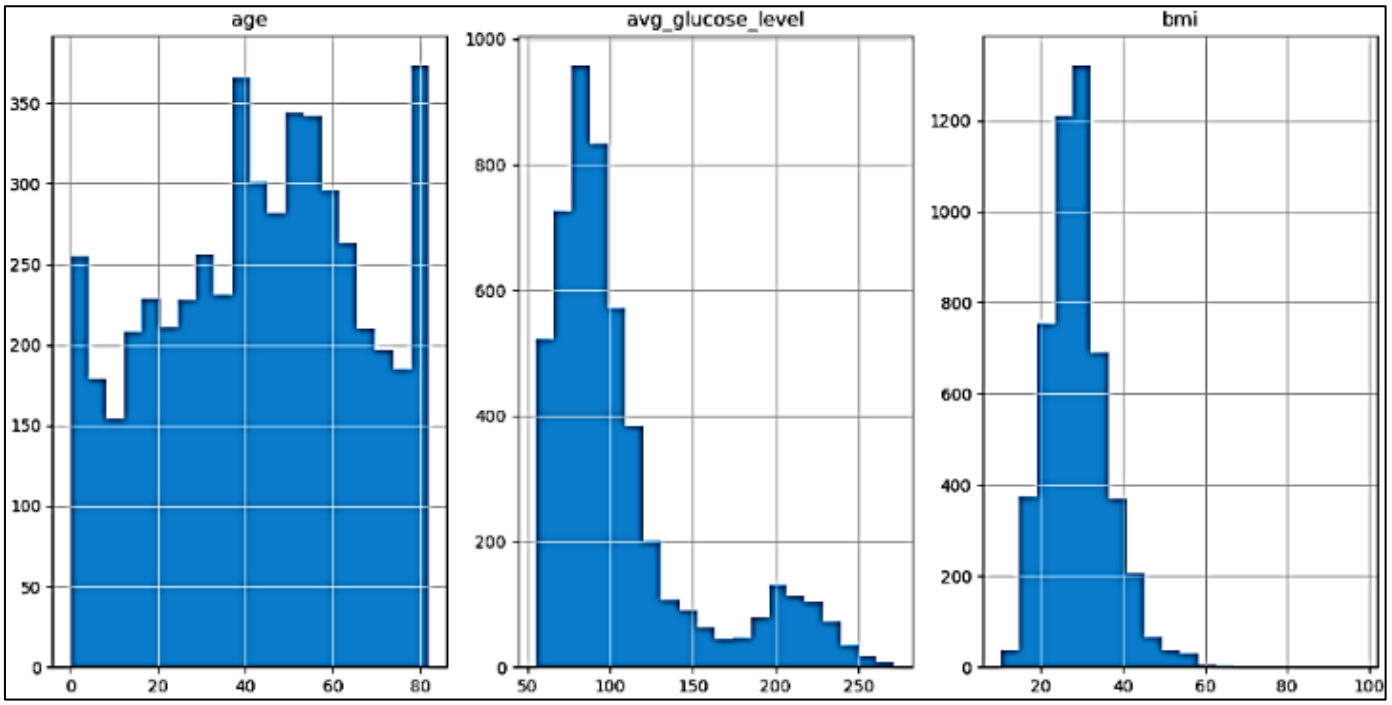

Fig 6 Distribution of Numerical Features Before Scaling

D. Model Training and Evaluation

We implemented three supervised learning algorithms (Logistic Regression, Random Forest, and XGBoost) to classify stroke recurrence. Each model was chosen to represent a different balance between complexity, interpretability, and performance, reflecting the practical considerations often faced in clinical settings.

➢ *Training Procedure and Cross-Validation:*

The dataset was divided into training and testing sets using a stratified 80:20 split, keeping the class distribution consistent across both subsets. To further strengthen model generalisability and guard against overfitting, we applied a 5-fold stratified cross-validation approach on the training data. Each model was trained and validated across five folds, with average performance scores used to guide model selection.

➢ *Model Configuration and Tuning:*

Each machine learning algorithm in this study went through a detailed hyperparameter optimization process using GridSearchCV. This involved systematically exploring a predefined grid of parameter combinations to find the setup that produced the best generalisation performance. Optimisation was carried out within a five-fold stratified cross-validation framework to improve robustness and minimise overfitting risks.

For logistic regression, we focused on tuning the regularization strength parameter, C, while enabling class weighting to help address the imbalance in the target variable. With the random forest classifier, we varied the number of estimators - the total trees in the ensemble - and the maximum depth of those trees to manage model complexity and prevent overfitting. In configuring XGBoost, we adjusted the learning rate, the number of estimators, and maximum tree depth. We also set the evaluation metric explicitly to log loss for XGBoost to keep performance tracking consistent across iterations, and we disabled early stopping to ensure comparable training conditions across all models.

The F1 Score was selected as the primary metric throughout the grid search. Given the imbalance in the outcome variable, it offered a more meaningful assessment than accuracy alone, balancing precision and recall. Focusing on F1 helped ensure that models were judged by their ability to correctly identify true cases of stroke recurrence, rather than simply performing well on the majority class.

Table 1 Summary of Hyperparameter Search Configuration for Each Model

Model	Parameters Tuned	Values Tested	Evaluation Metric
Logistic Regression	Regularisation Strength (C)	0.01, 0.1, 1, 10	F1 Score
Random Forest	Number of Trees, Max Depth	100, 200; None, 10, 20	F1 Score
XGBoost	Learning Rate, Estimators, Max Depth	0.01, 0.1; 100, 200; 3, 6	F1 Score

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

➢ *Performance Metrics:*

We used a set of evaluation metrics to assess the models' performance from different angles, making sure to capture not only overall accuracy but also the ability to detect minority-class instances, with stroke recurrence being the key clinical concern. Accuracy measures how many predictions the model got right across the entire dataset, offering a broad view of overall performance. Precision, by contrast, tells us how many of the cases predicted as recurrence were actually correct, giving insight into the reliability of positive predictions. Recall, sometimes referred to as sensitivity, looks at how many true recurrence cases the model successfully identified, reflecting its capacity to catch events that truly matter in a clinical setting.

Given the imbalance between recurrence and non-recurrence cases, we placed particular emphasis on the F1 Score. The F1 Score strikes a balance between precision and recall, which is crucial when one class heavily outweighs the other. It becomes especially important when missing a positive case could carry serious clinical consequences. The F1 Score is calculated as the harmonic mean of precision and recall, using the following formula:

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

The F1 Score only reaches a high value when both precision and recall are strong, making it especially useful when both false positives and false negatives carry serious consequences. Alongside this, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was used to evaluate the models' discriminative ability across a range of classification thresholds. AUC-ROC offers a view of how well each model separates recurrence cases from non-recurrence cases at different probability cut-offs, rather than relying on a single point estimate.

Given the focus of this study on identifying patients at risk of stroke recurrence early, recall and F1 Score were prioritised during model tuning and selection. These metrics were better suited to the clinical goal of reducing missed detections while maintaining predictions that could be trusted in practice.

Table 2 Definitions and Purpose of Evaluation Metrics

Metric	Definition	Purpose in This Study
Accuracy	Proportion of all correct predictions over the total number of predictions	Provides an overall measure of model correctness across all classes
Precision	Proportion of true positives among all predicted positives	Measures the trustworthiness of positive recurrence predictions
Recall	Proportion of true positives among all actual positives	Assess the model's ability to identify all recurrence cases
F1-Score	Harmonic mean of precision and recall	Balances precision and recall for imbalanced data
AUC-ROC	The area under the curve plots the true positive rate against the false positive rate.	Evaluates the model's discriminative ability at various classification thresholds

➢ *Model Selection Principles:*

Model selection aimed to balance predictive strength, generalisability, and interpretability. We evaluated each algorithm using consistent scoring metrics from stratified cross-validation, with hyperparameter tuning done through grid search. Priority was given to models that could handle both linear and non-linear relationships without slipping into overfitting.

Feature importance analysis was applied to models that offered built-in interpretability, especially tree-based methods. Instead of treating it as an afterthought, we used it to understand which predictors played the biggest role in shaping classification outcomes. In clinical machine learning, that kind of transparency matters - it helps build trust and makes it easier for predictive models to fit into real healthcare workflows where explainable results are not just a bonus but a necessity.

IV. RESULTS AND DISCUSSION

The machine learning models (Logistic Regression, Random Forest, and XGBoost) were evaluated through a framework designed to capture different aspects of classification performance. We assessed each model using accuracy, precision, recall, F1 Score, and AUC-ROC, paying closer attention to the metrics that carry more weight when dealing with an imbalanced dataset like this one.

Table 3 shows the test set performance results for all three classifiers. Among them, the Random Forest model stood out, posting the highest values across nearly all metrics. It reached an accuracy of 0.9239, precision of 0.9111, recall of 0.9405, an F1 Score of 0.9256, and an AUC-ROC of 0.9704. Taken together, these results suggest that Random Forest offered the strongest and most balanced performance, managing to identify stroke recurrence cases effectively while keeping both false positives and false negatives under control.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

Table 3 Test Set Performance Metrics Across Models

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.8638	0.8462	0.8891	0.8671	0.9305
Random Forest	0.9239	0.9111	0.9405	0.9256	0.9704
XGBoost	0.9021	0.8846	0.9243	0.9040	0.9562

The comparative bar chart in Figure 6 brings these numerical results into clearer focus. All three models performed strongly, but Random Forest pulled ahead, especially in recall and F1 Score. Logistic Regression, while statistically consistent, showed slightly weaker sensitivity when it came to picking up true recurrence cases. XGBoost, despite its powerful boosting approach, came close to Random Forest's performance but still fell a little short on the clinical metrics that mattered most. Seeing the models side by side highlights not just their strengths, but also the areas where each struggled a bit in dealing with class imbalance and catching minority-class instances.

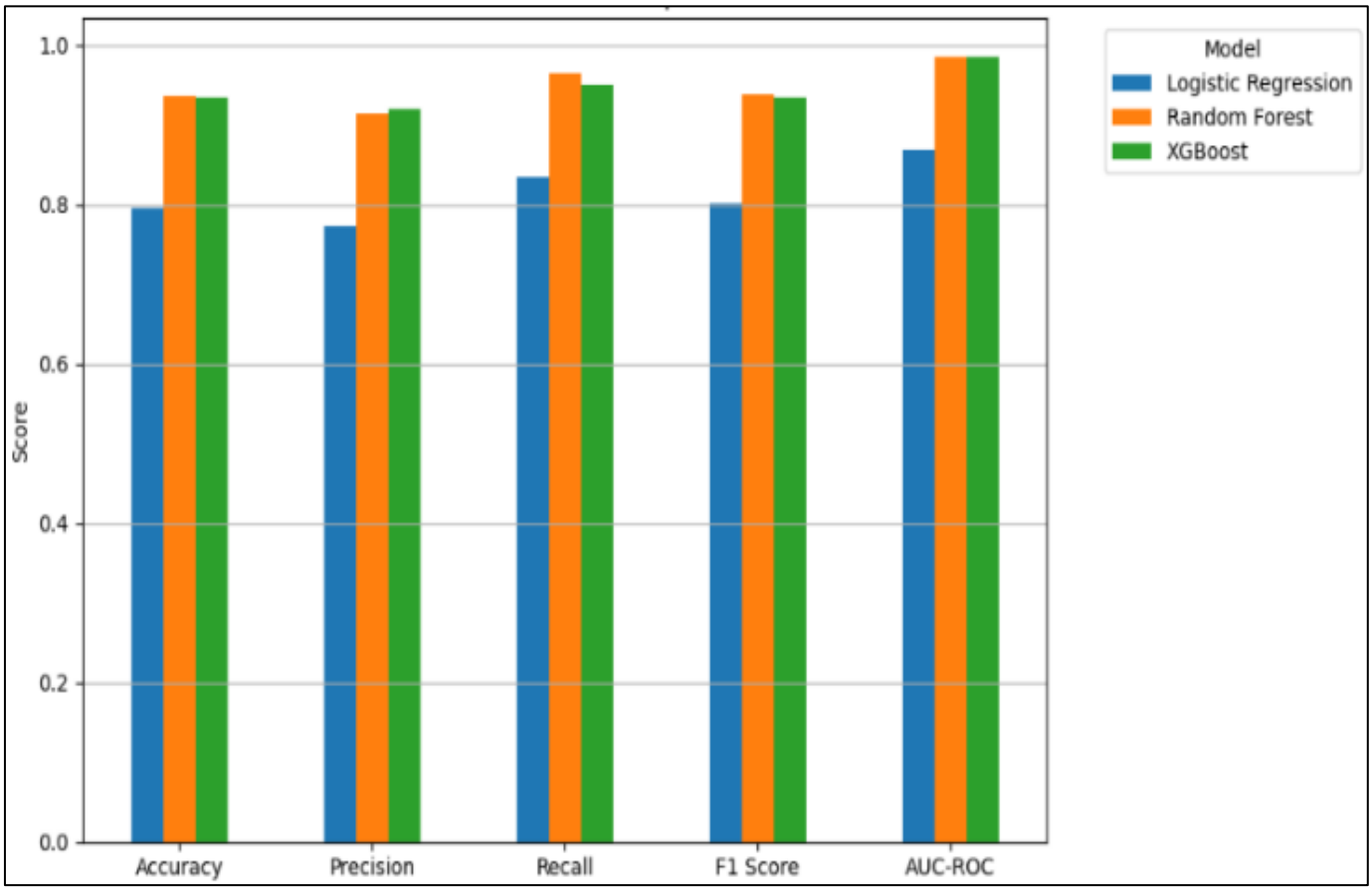

Fig 7 Comparative Performance Metrics for Logistic Regression, Random Forest, and XGBoost

Further diagnostic insights into the Random Forest classifier are shown in its confusion matrix, presented in Figure 8. The matrix shows strong diagonal dominance, which points to a high number of correctly classified cases. What matters more, though, is the model's low false negative rate, crucial in medical settings where missing a true recurrence could mean losing the chance to intervene early. The balance between true positives and false positives also strengthens the model's clinical value, helping it avoid overwhelming practitioners with false alarms while still catching the majority of recurrence cases that matter most.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

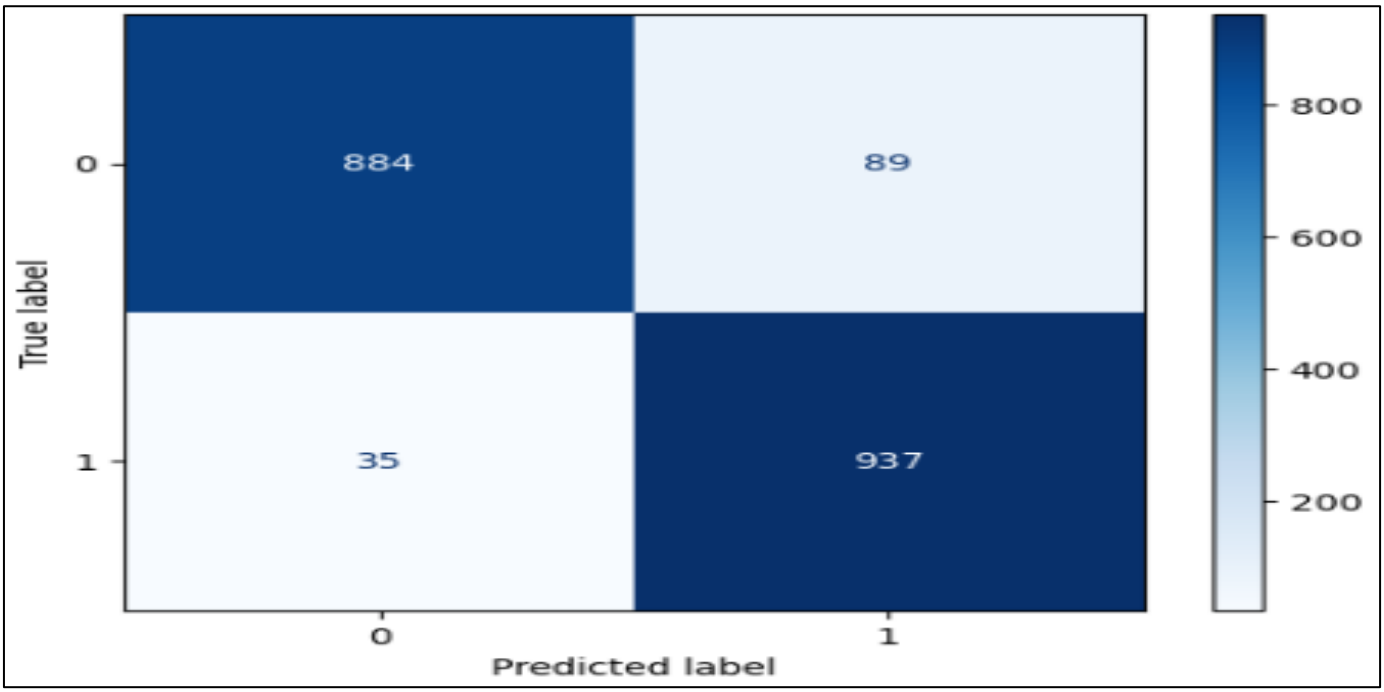

Fig 8 Confusion Matrix for Random Forest Model Predictions

ROC curves, shown in Figure 9, offer a view of how well the models separate the two classes across different threshold settings. The Random Forest classifier achieved the highest area under the curve, backing up its stronger discriminative ability. Its curve shows a steep initial rise and stays close to the top-left corner of the graph, which signals that it can correctly identify positive cases while keeping false positives low.

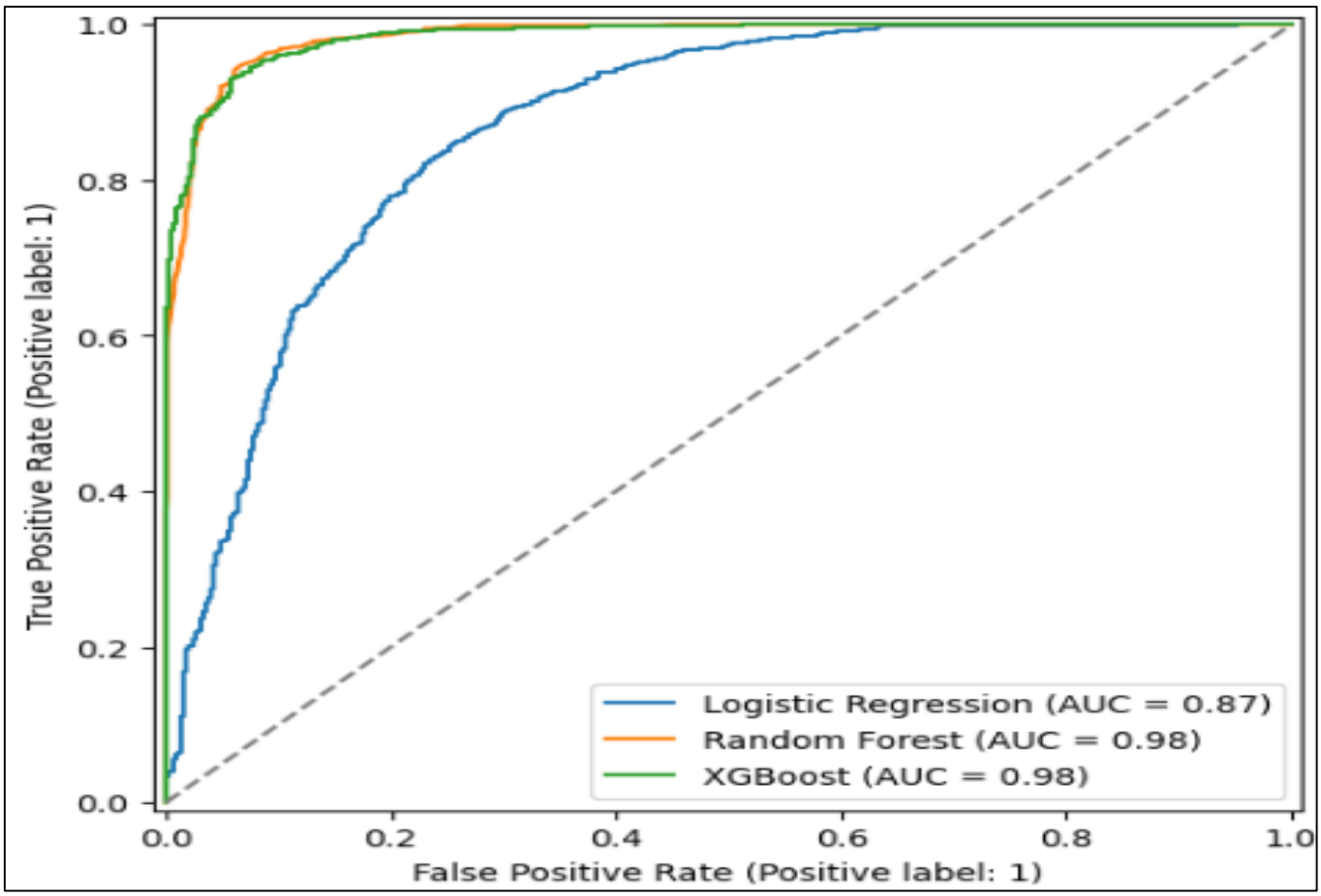

Fig 9 ROC Curves for All Models

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

Alongside the ROC analysis, precision-recall curves were plotted in Figure 10 to give a closer look at how each model handled the minority class. The Random Forest model showed the largest area under the precision-recall curve, reinforcing its strength in detecting stroke recurrence while still keeping precision high. That balance matters, especially in a clinical setting where the cost of missing a true recurrence is much higher than the risk of raising an extra precautionary alert.

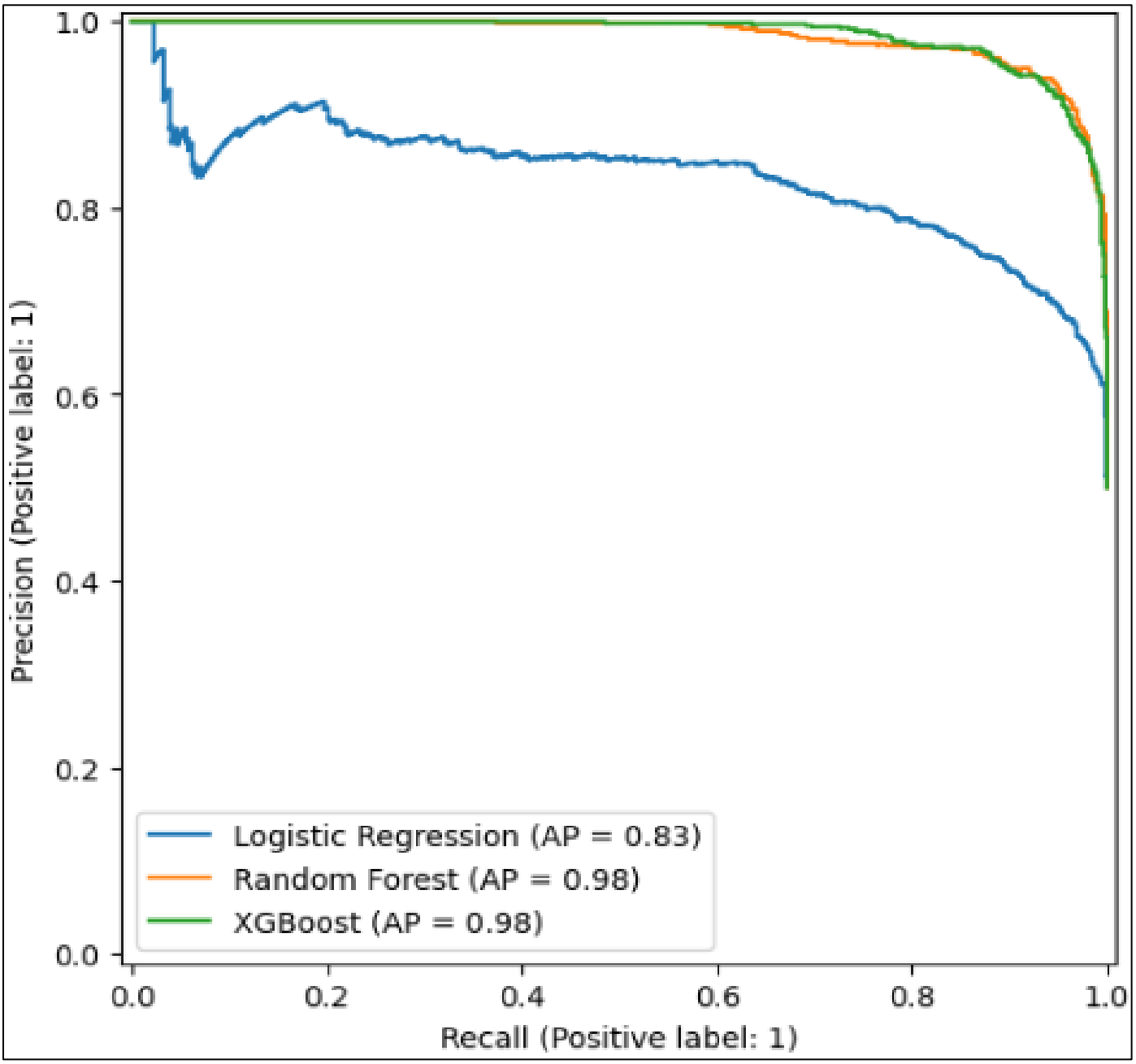

Fig 10 Precision-Recall Curves for All Models

The internal structure of the Random Forest model was also examined to assess feature importance. As shown in Figure 10, the top predictors were age, average glucose level, BMI, and history of hypertension. These findings are consistent with established clinical research, suggesting that the model identified patterns grounded in clinical relevance. The ability to see which features drive predictions improves the model's transparency and offers clinicians clearer insight into the factors most associated with recurrence risk.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

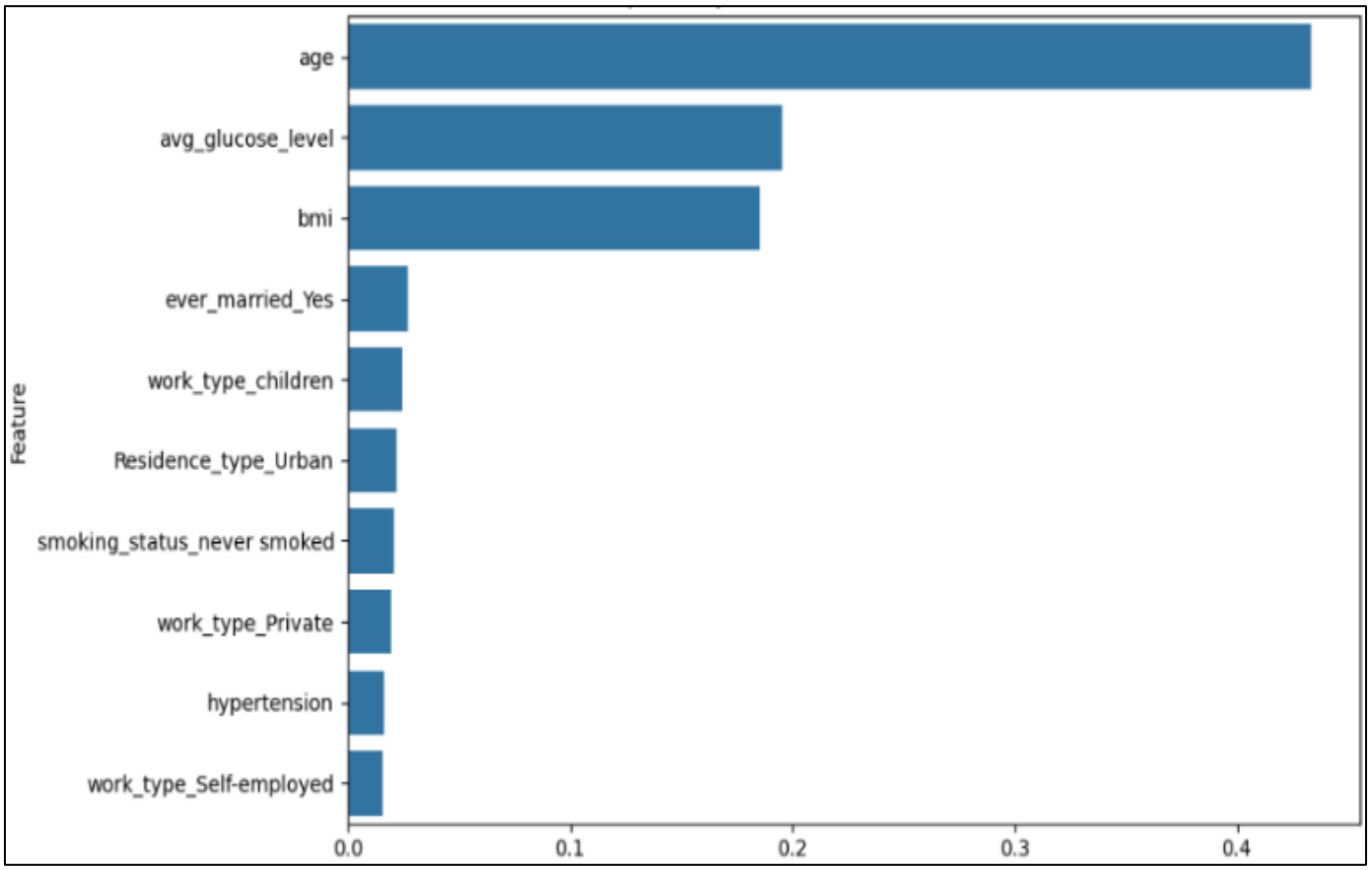

Fig 11 Feature Importance Rankings from the Random Forest Model

Although accuracy was reported as a general indicator, it was never the main basis for choosing a model. With the class imbalance in this dataset, accuracy alone could easily hide serious problems - a model that mostly guesses non-recurrence would still score well without doing the real job. What mattered far more was recall and the F1 Score. Both offered a better sense of how the models handled the tougher challenge: catching true recurrence cases without getting drowned in false positives. Especially in stroke prediction, missing a real case could mean missing a chance to intervene, and the consequences of that are not something a few extra percentage points of accuracy can excuse.

It's worth pausing here to recall what precision and recall actually measure, since they shaped how we judged model performance. Precision is defined as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

This measures the proportion of predicted recurrence cases that were correct. Meanwhile, recall (or sensitivity) is given by:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

High recall reflects how well the model captures all true cases of recurrence. In clinical work, that matters more than anything, since missing a recurrence could mean missing the narrow window for life-saving intervention.

The results of this study point to Random Forest as the most effective model for predicting stroke recurrence with a minimal feature set. Its stronger recall and F1 Score make it a good fit for real-world clinical use, especially in settings that rely on routine health records rather than rich, complex datasets. The model's success also supports the broader choices made throughout the study - from preprocessing steps to feature selection and the validation strategies used during training.

More broadly, these findings add to the growing sense that machine learning, when built carefully and judged with the right metrics, can actually improve clinical risk prediction rather than just making things more complicated. Even so, the work here is only a step. External validation on new datasets and real-world testing in hospital and primary care settings will be critical before any tool like this can be fully trusted.

V. LIMITATIONS AND FUTURE WORK

While this study offers promising results for predicting stroke recurrence using a minimal set of clinical features, there are clear limitations that need to be considered. One of the bigger concerns is the reliance on a single dataset from a public clinical repository. Although it provided enough depth for an initial exploration, model performance could shift when tested against different populations or healthcare

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

systems. Without external validation, the findings remain tied to the original dataset's characteristics and might not fully capture the broader variability seen in real-world stroke recurrence.

Another limitation comes from the static nature of the features we used. All variables were drawn from a single point in time, without any follow-up or longitudinal tracking. That means the model misses the ability to account for changes in patient health over time, whether improvement, decline, or new risk factors, any of which could affect recurrence outcomes. Future work would likely benefit from incorporating longitudinal data to add richer clinical context.

Class imbalance also presents a lingering challenge. Although SMOTE helped to rebalance the data during training, synthetic examples can never fully replicate the complexity of true recurrence cases. There's a risk that this balancing could shape model behaviour in ways that don't hold up when exposed to real, naturally imbalanced data. Exploring alternative balancing methods, or working with naturally balanced datasets when possible, would be worth investigating.

Another boundary comes from the type of data we used. The model was built entirely on structured tabular information, which, while practical, leaves out a great deal of nuance found in clinical notes, imaging results, and patient narratives. Future models that integrate unstructured data through approaches like natural language processing could deepen predictive insights.

The deliberate focus on a minimal feature set was important for making the model accessible, but it also meant leaving out potentially valuable variables. Factors like detailed cardiovascular histories, medication adherence patterns, or imaging biomarkers could add predictive strength. Expanding the feature set carefully might push performance even further without losing usability.

Looking ahead, the next steps should involve testing the model across external settings - whether hospitals, primary care environments, or regional stroke registries - to see how well it holds up outside a controlled environment. Implementation studies will also be critical, helping to understand how the model fits into real-world clinical workflows and what adjustments might be needed to support its adoption. Building a feedback loop between model predictions and clinician insights could turn an experimental tool into something truly useful at the bedside

VI. CONCLUSION

This study explored the use of machine learning techniques to predict stroke recurrence based on a small set of clinically relevant features. By implementing and evaluating Logistic Regression, Random Forest, and XGBoost models, the analysis showed that strong predictive performance is still possible even when the feature space is kept deliberately limited. Across all evaluation metrics, Random Forest consistently led the way, with its higher recall and F1 Score pointing to a better ability to catch recurrence cases without introducing too many false negatives.

The approach taken throughout the study aimed to keep a balance between interpretability, predictive strength, and practical clinical use. This balance showed up in the choices made around preprocessing, feature selection, and model explainability. By focusing on straightforward but powerful models, the study offers a pathway for developing decision-support tools that could work both in well-resourced and more constrained healthcare environments. The findings add to the growing sense that machine learning, when built carefully, has the potential to improve secondary stroke prevention in meaningful ways. Still, real progress will depend on validating these models externally and testing how well they actually perform once embedded into everyday clinical practice.

REFERENCES

[1]. Feigin, V., Brainin, M., Norrving, B., Martins, S., Pandian, J., Lindsay, M., ... Rautalin, I. (2025). "World stroke organization: global stroke fact sheet 2025." *International Journal of Stroke, 20*(2), 132--144. https://doi.org/10.1177/17474930241308142

[2]. Yu, Q., Tian, Y., Jiang, N., Zhao, F., Wang, S., Sun, M., & Liu, X. (2025). "Global, regional, and national burden and trends of stroke among youths and young adults aged 15--39 years from 1990 to 2021: findings from the global burden of disease study 2021." *Frontiers in Neurology, 16.* https://doi.org/10.3389/fneur.2025.1535278

[3]. Liu, X., Wu, X., Yan, S., Chu, C., Wang, L., Li, H., ... Li, Q. (2024). "Association of MRI markers of cerebral small vessel disease and ischemic stroke recurrence in patients treated with intravenous thrombolysis: a three-year prospective cohort study." https://doi.org/10.21203/rs.3.rs-4891113/v1

[4]. Diri, G. O., Diri, E. E., Nbaakee, L. G., James, N. H., & Igulu, K. T. (2025). "Electrocardiographic and biochemical feature integration for automated cardiovascular risk stratification." *International Journal of Research and Innovation in Applied Science, 10*(6). https://doi.org/10.51584/IJRIAS.2025.10060042

[5]. Mbalinda, S., Kaddumukasa, M., Najjuma, J., Kaddumukasa, M., Nakibuuka, J., Burant, C., & Sajatovic, M. (2024). "Stroke recurrence rate and risk factors among stroke survivors in sub-Saharan Africa: a systematic review." *Neuropsychiatric Disease and Treatment, 20,* 783--791. https://doi.org/10.2147/ndt.s442507

[6]. Zhao, J., Wang, D., Liu, X., Wang, Y., & Zhao, X. (2023). "The predictive value of Essen and SPI-II on the risk of 5-year recurrence in Chinese patients with acute ischemic stroke." *Neuropsychiatric Disease and Treatment, 19,* 2251--2260. https://doi.org/10.2147/ndt.s433383

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

[7]. Heo, J. (2025). "Application of artificial intelligence in acute ischemic stroke: a scoping review." *Neurointervention, 20*(1), 4--14. https://doi.org/10.5469/neuroint.2025.00052

[8]. Paliwal, S., Parveen, S., Alam, M., & Ahmed, J. (2023). "Improving brain stroke prediction through oversampling techniques: a comparative evaluation of machine learning algorithms." https://doi.org/10.20944/preprints202306.1444.v1

[9]. Chen, Y., Chung, J., Yeh, Y., Lou, S., Lin, H., Lin, C., ... Shi, H. (2022). "Predicting 30-day readmission for stroke using machine learning algorithms: a prospective cohort study." *Frontiers in Neurology, 13*. https://doi.org/10.3389/fneur.2022.875491

[10]. Heo, J., Yoo, J., Lee, H., Lee, I., Kim, J., Park, E., ... Nam, H. (2022). "Prediction of hidden coronary artery disease using machine learning in patients with acute ischemic stroke." *Neurology, 99*(1). https://doi.org/10.1212/wnl.0000000000200576

[11]. Parvathi, S., B, A., Kulkarni, G., Murugan, S., & Vijayammal, B. (2024). "Exploring feature relationships in brain stroke data using polynomial feature transformation and linear regression modeling." *Journal of Machine and Computing*, 1158--1169. https://doi.org/10.53759/7669/jmc202404107

[12]. Hadiyoso, S., Ong, P., Zakaria, H., & Rajab, T. (2022). "EEG-based spectral dynamic in characterization of poststroke patients with cognitive impairment for early detection of vascular dementia." *Journal of Healthcare Engineering, 2022*, 1--11. https://doi.org/10.1155/2022/5666229

[13]. Zheng, P., Huiyu, S., Li, M., Qingke, B., Qiuyun, L., & Xu, C. (2023). "Explainable machine learning for long-term outcome prediction in two-center stroke patients after intravenous thrombolysis." *Frontiers in Neuroscience, 17*. https://doi.org/10.3389/fnins.2023.1146197

[14]. Padimi, V., Telu, V., & Ningombam, D. (2022). "Performance analysis and comparison of various machine learning algorithms for early stroke prediction." *ETRI Journal, 45*(6), 1007--1021. https://doi.org/10.4218/etrij.2022-0271

[15]. He, W., Le, H., & Du, P. (2022). "Stroke prediction model based on XGBoost algorithm." *International Journal of Applied Sciences & Development, 1*, 7--10. https://doi.org/10.37394/232029.2022.1.2

[16]. Mitra, R., & Rajendran, T. (2022). "Efficient prediction of stroke patients using random forest algorithm in comparison to support vector machine." https://doi.org/10.3233/apc220075

[17]. Shahade, A., & Deshmukh, P. (2025). "Gradient boosting for heart stroke prediction: investigating unexpected risk factors." *Journal of Computer Science, 21*(1), 124--133. https://doi.org/10.3844/jcssp.2025.124.133

[18]. Shih, H., Law, K., Yeh, Y., Wu, K., Lai, J., Lin, C., ... Kao, C. (2022). "Applying machine learning to carotid sonographic features for recurrent stroke in patients with acute stroke." *Frontiers in Cardiovascular Medicine, 9*. https://doi.org/10.3389/fcvm.2022.804410

[19]. Ma, L., Fu, G., Liu, R., Zhou, F., Dong, S., Zhou, Y., ... Wang, X. (2023). "Phenylacetyl glutamine: a novel biomarker for stroke recurrence warning." *BMC Neurology, 23*(1). https://doi.org/10.1186/s12883-023-03118-5

[20]. Pucar, Đ., & Šimović, V. (2024). "Predictive modeling of stroke occurrence using Python for improved risk assessment." *Journal of Process Management New Technologies, 12*(1--2), 110--120. https://doi.org/10.5937/jpmnt12-50921

[21]. Setyarini, D., Gayatri, A., Aditya, C., & Chandranegara, D. (2024). "Stroke prediction with enhanced gradient boosting classifier and strategic hyperparameter." *Matrik Jurnal Manajemen Teknik Informatika Dan Rekayasa Komputer, 23*(2), 477--490. https://doi.org/10.30812/matrik.v23i2.3555

[22]. Cao, S., Zhao, L., Pei, L., Gao, Y., Fang, H., Liu, K., & Xu, Y. (2023). "ABCD2 score has equivalent stroke risk prediction for anterior circulation TIA and posterior circulation TIA." *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-41260-9

[23]. Irie, F., Matsumoto, K., Matsuo, R., Nohara, Y., Wakisaka, Y., Ago, T., ... Kamouchi, M. (2024). "Predictive performance of machine learning--based models for poststroke clinical outcomes in comparison with conventional prognostic scores: multicenter, hospital-based observational study." *JMIR AI, 3*, e46840. https://doi.org/10.2196/4684

[24]. Gao, Y., Li, Z., Zhai, X., Han, L., Ping, Z., Cheng, S., ... Cui, H. (2024). "An interpretable machine learning model for stroke recurrence in patients with symptomatic intracranial atherosclerotic arterial stenosis." *Frontiers in Neuroscience, 17*. https://doi.org/10.3389/fnins.2023.1323270

[25]. Shao, S., Wang, T., Zhu, L., Yin, G., Fan, X., Lu, Y., ... Qian, J. (2025). "Correlation of intracranial and extracranial carotid atherosclerotic plaque characteristics with ischemic stroke recurrence: a high-resolution vessel wall imaging study." *Frontiers in Neurology, 15*. https://doi.org/10.3389/fneur.2024.1514711

[26]. Sousanidou, A., Tsiptsios, D., Christidi, F., Karatzetzou, S., Kokkotis, C., Gkantzios, A., ... Vadikolias, K. (2023). "Exploring the impact of cerebral microbleeds on stroke management." *Neurology International, 15*(1), 188--224. https://doi.org/10.3390/neurolint15010014

[27]. Dimaras, T., Merkouris, E., Tsiptsios, D., Christidi, F., Sousanidou, A., Orgianelis, I., ... Vadikolias, K. (2023). "Leukoaraiosis as a promising biomarker of stroke recurrence among stroke survivors: a systematic review." *Neurology International, 15*(3), 994--1013. https://doi.org/10.3390/neurolint15030064

[28]. Li, Y., Wang, Z., Wu, T., & Zhou, T. (2023). "Comparison of six machine learning algorithms for stroke risk estimation." *Applied and Computational Engineering, 8*(1), 556--561. https://doi.org/10.54254/2755-2721/8/20230274

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25sep706

[29]. Park, S., Choi, J., Kim, Y., & You, J. (2024). "Clinical machine learning predicting best stroke rehabilitation responders to exoskeletal robotic gait rehabilitation." *Neurorehabilitation*, *54*(4), 619--628. https://doi.org/10.3233/nre-240070

[30]. Shahade, A., & Deshmukh, P. (2025). "Gradient boosting for heart stroke prediction: investigating unexpected risk factors." *Journal of Computer Science*, *21*(1), 124--133. https://doi.org/10.3844/jcssp.2025.124.133 *[Second occurrence]*

[31]. Hairani, H., Widiyaningtyas, T., & Prasetya, D. (2024). "Feature selection and hybrid sampling with machine learning methods for health data classification." *Revue d'Intelligence Artificielle*, *38*(4), 1255--1261. https://doi.org/10.18280/ria.380419

[32]. Yin, Q., Ye, X., Huang, B., Qin, L., Ye, X., & Wang, J. (2023). "Stroke risk prediction: comparing different sampling algorithms." *International Journal of Advanced Computer Science and Applications*, *14*(6). https://doi.org/10.14569/ijacsa.2023.01406115

[33]. Wakisaka, K., Matsuo, R., Matsumoto, K., Nohara, Y., Irie, F., Wakisaka, Y., ... Kitazono, T. (2023). "Non-linear association between body weight and functional outcome after acute ischemic stroke." *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-35894-y

[34]. Iguchi, T., Kojima, K., Hayashi, D., Tokunaga, T., Okishio, K., & Yoon, H. (2025). "Preoperative maximum standardized uptake value emphasized in explainable machine learning model for predicting the risk of recurrence in resected non--small cell lung cancer." *JCO Clinical Cancer Informatics, (9).* https://doi.org/10.1200/cci-24-00194

[35]. Nasution, N., Nasution, F., Erlin, E., & Hasan, M. (2024). "Evaluation study of the chi-square method for feature selection in stroke prediction with random forest regression." https://doi.org/10.4108/eai.30-10-2023.2343096