International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

Comparative Analysis of Machine Learning Algorithms for Brain Stroke Risk Prediction

Ravi[a)], Dr. Vishal Verma and Mr. Sonu
MCA Student[1], Professor[2], Asst. Prof.[3]
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Brain stroke is a major cause of mortality and disability in the world. The ability to predict stroke risk at an early stage from routinely available clinical information allows timely preventive intervention to be provided. In this paper, three supervised machine learning techniques such as Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) are systematically applied and analyzed on a publicly available brain stroke dataset with 4,981patients records. A full preprocessing pipeline was applied such as label encoding, IQR based outlier capping (avg_glucose_level cap at 168.805 mg/dL), stratified train-test split (3,984/997), standard scaling, SelectKBest feature selection (all 10 features retained) and finally, SMOTE to overcome the class imbalance problem (original ratio 3,786:198, balanced ratio 3,786:3,786). Random Forest performed best with respect to test accuracy (87.06%), F1 score (89.46%) and cross validation accuracy (92.02%), beating LR performance whose accuracy is 74.82% and SVM has accuracy 78.44%. Logistic Regression had the best ROC-AUC (0.8435). The confusion matrix obtained by the Random Forest classifer gave the values TP=20, FN=30, FP=99 and TN=848. The age and the average glucose level and the BMI were the most important stroke predictors as determined by feature importance analysis. These results show the feasibility of using ensemble learning for risk screening in clinical stroke.

Keywords- Brain Stroke Prediction, Machine Learning, Random Forest, Logistic Regression, SVM, SMOTE.

I. INTRODUCTION

In the world, stroke is the second biggest cause of mortality, causing around 15 million cases a year, of which around 5 million people pass away and 5 million are permanently disabled [1]. It may develop as a result of an ischemic blockage of an artery (>80%) or haemorrhage of the vessels. Hypertension, diabetes, sedentary lifestyle and the use of tobacco are the leading causes of stroke, which is the fourth supreme root of demise in India and is becoming a growing problem in the adult population below the age of 45.

Traditionally, diagnosis of stroke is made using the CT and MRI scans which are resource-intensive and not accessible in primary care. Structured clinical information that is collected routinely from patients (age, BMI, glucose, hypertension) can be fed into trained classifiers to predict the individual patient's stroke risk prior to the event. Machine learning provides an attractive alternative. This allows for cost effective, scalable population screening.

Previous research has shown that ensemble methods and deep learning can achieve over 90% accuracy [1], [2]. However, there is a lack of standardization across different research works in terms of preprocessing, class imbalance treatment and the single-metric evaluation of accuracy which restricts the comparability and clinical usefulness of published results. This paper aims to bridge these gaps by designing and implementing a controlled and reproducible comparative experiment.

The major benefactions of this work are: (1) a well ordered comparison of LR, RF and SVM with the same preprocessing method on the same dataset of 4,981 stroke cases; (2) explicit class balancing using SMOTE method, to balance the classes with training data only; (3) reporting classification results per class to reveal minority class performance; and (4) linking predictive ranks to clinical evidence by using feature importance analysis.

II. RELATED WORK

Table I summarises six directly relevant studies published in 2024–2025, all using publicly available stroke prediction datasets.

TABLE I

Summary of Related Work on ML-Based Stroke Prediction

Ref.	Algorithms	Dataset	Best Acc.	Key Finding
[1] Noor (2025)	LR, DT, RF, NB, KNN	Kaggle 4,981	RF 95.05%	RF best; RapidMiner; class imbalance noted
[2] Kitova (2024)	SVM, RF, DSE, XGBoost	Kaggle 5,110	SVM 97.82%	SMOTE+MICE; DSE error 2.81%; no universal best
[3] Kundu (2025)	DT, LR, KNN, SVM, GB, XGB	Kaggle 5,110	SVM 95%	SVM best; web interface deployed
[4] Khan (2025)	Hybrid KNN+RF+GA	Kaggle 5,110	94.51%	GA feature selection; age top predictor
[5] Dhiman (2025)	GNB, SVM, RF, KNN, XGB, NN	Kaggle 5,110	RF 95.18%	RF best precision; NN best recall
[6] Basu (2025)	10 classifiers + ANN	43,400 records	ANN 99.37%	ANN best AUC; computationally expensive

The best results according to Noor et al. [1] were obtained using Random Forest with 95.05% accuracy on 4981 records on RapidMiner. Kitova et al. [3] contrasted three approaches to the problem of class imbalance and missing data, and found SVM with synthetic augmentation to have 97.82% accuracy and an ROC of 0.9975. Kundu et al. [4] used SVM for achieving 95% accuracy along with a web interface. Khan et al. [5] presented a hybrid GA+KNN+RF approach to achieve a 94.51% accuracy with SMOTE. Dhiman et al. [6] concluded that the RF algorithm is the best with 95.18% precision. While Basu et al.[2] showed dominance of ANN at 99.37% for a sample database size of 43,400, it required an excessive amount of computation.

One common problem in all these studies is the inconsistent treatment of class imbalance and the lack of attention for the clinically relevant metric that is the per-class recall, which is the relevant measure for stroke screening.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

III. DATASET DESCRIPTION

The Brain Stroke Prediction Dataset was taken from the public repository of Kaggle. The shape of dataset is (4981, 11): 4,981 patient records with 10 independent features and 1 binary target. All features are described in Table II.

TABLE II

Dataset Feature Description

Feature	Description	Type / Range
age	Patient age	Numeric: 0.08–82.0 (μ=43.42)
gender	Biological sex	Categorical: Male, Female
hypertension	High blood pressure	Binary: 0/1
heart_disease	Cardiac history	Binary: 0/1
ever_married	Marital status	Categorical: Yes/No
work_type	Employment category	5 categories
Residence_type	Urban/rural	Categorical
avg_glucose_level	Avg. blood glucose mg/dL	Numeric: 55.12–168.81 (capped)
bmi	Body Mass Index	Numeric: 14.0–48.9 (μ=28.50)
smoking_status	Smoking habit	4 categories
stroke	Target variable	Binary: 0=No (95%), 1=Yes (5%)

Statistical summary of the dataset : age mean=43.42, std=22.67; avg_glucose_level mean=105.94, std=45.08; BMI mean=28.50, std=6.79. It was verified that there were zero null values and duplicate rows. The data is highly imbalanced with a target variable having 4,733 non-stroke and 248 stroke records (95.03% - 4.97%). [7]

IV. PROPOSED METHODOLOGY

Fig. (1) presents the complete processing pipeline.

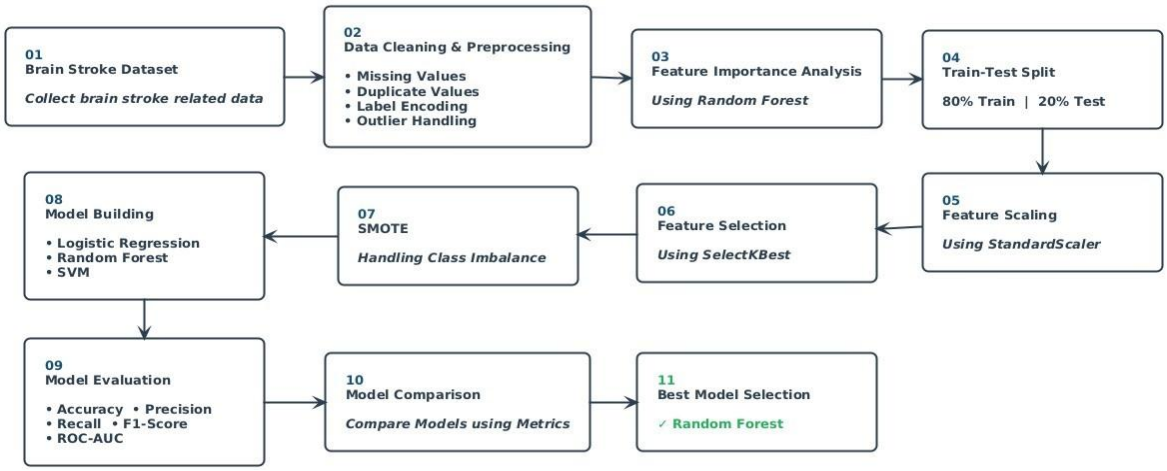


Fig. 1. Proposed Pipeline

A) Label Encoding

The categorical features are: ever_married, gender, work_type, smoking_status, Residence_type, which were encoded using LabelEncoder from scikit-learn. A successful post-encoding df.info() reveals all 11 of these columns are numeric (int64/float64), with 4,981 non-null values found in each.

B) Outlier Detection and Handling

Using IQR-based outlier detection, 602 outliers were found in output for avg_glucose_level (capped at 168.805 mg/dL) and 43 in BMI (capped at 95th percentile). The binary features (hypertension: 479, heart_disease: 275, stroke: 248) were preserved without any transformation. Following handling, there are no outliers for both continuous features.

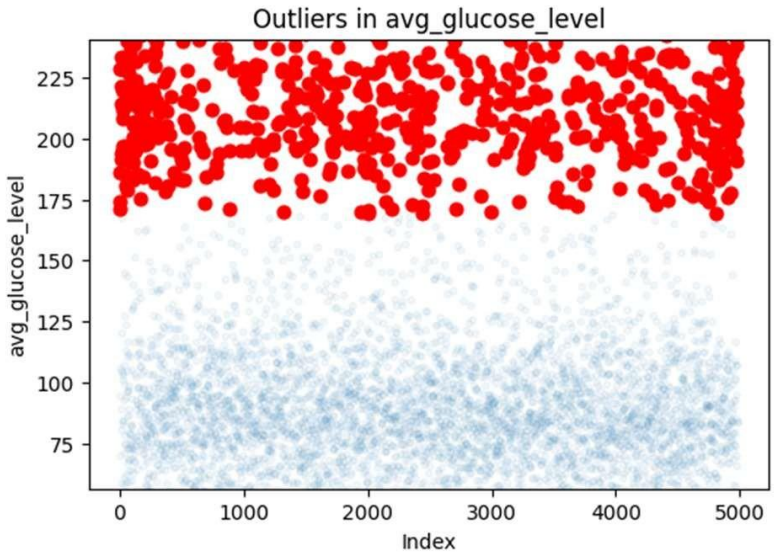
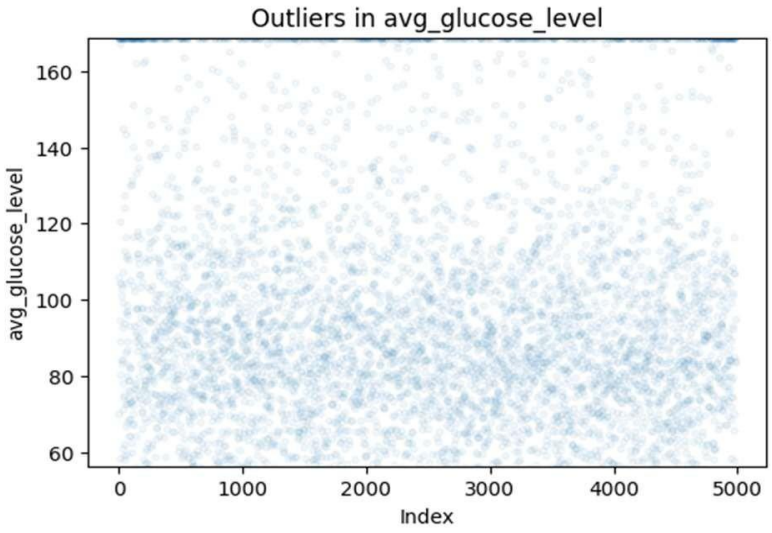


Fig. 2. Avg. Glucose Level Outliers - Before Handling



International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

Fig. 3. Avg. Glucose Level Outliers -After Handling

C) Exploratory Data Analysis

From the correlation heatmap (Fig. 4) it is obvious that age is the dominant positive correlation with stroke, comingbefore avg_glucose_level and hypertension. A class distribution is also provided in Fig. 5, which confirms severe imbalance. Fig. 6. shows that the number of cases of stroke is concentrated at higher ages.

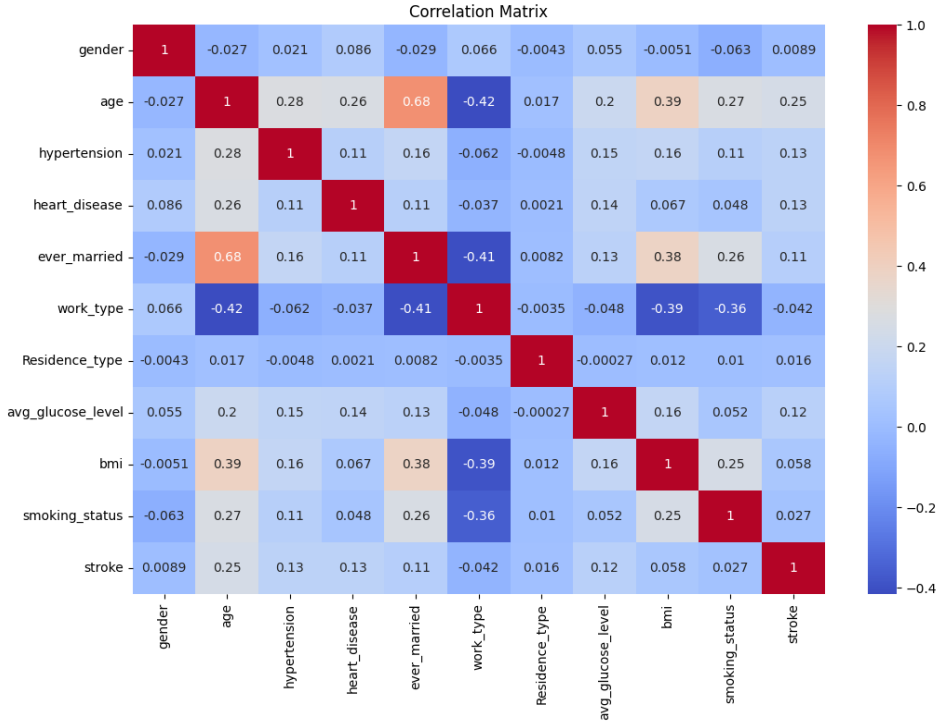


Fig. 4. Full Correlation Heatmap of All Features

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

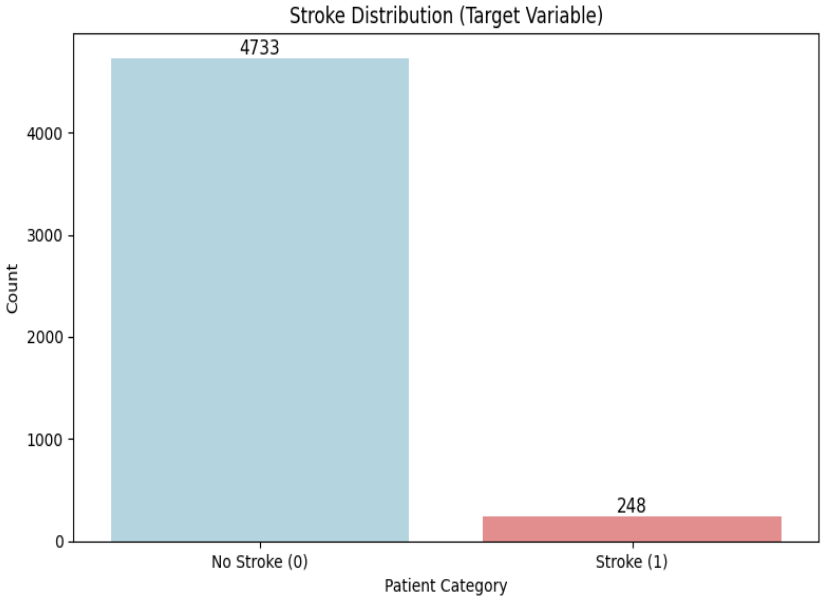


Fig. 5. Stroke Class Distribution -Original Dataset (95:5 imbalance)

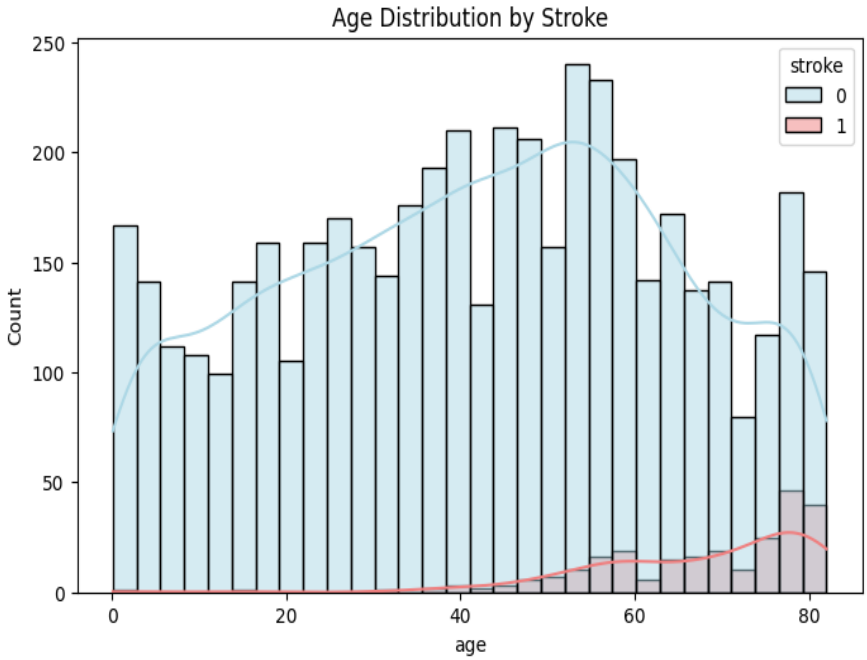


Fig. 6. Age Distribution by Stroke Outcome

D) Train-Test Split and Scaling

An 80:20 stratified split yielded Train: (3984, 10) and Test: (997, 10) as confirmed by results. To avoid leakage, StandardScaler was applied to both splits with the same parameters excercised on the training data.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

E) Feature Selection

The selectKBest algorithm (f_classif, k = 10) was used. All 10 of the available features have predictive information that is meaningful.

F) SMOTE Class Balancing

Only training set was employed for SMOTE. Before: class 0=3,786, class 1=198. After: class 0=3,786, class 1=3,786 — a perfectly balanced training set of 7,572 samples. To make the evaluation realistic, we did not modify the test set (947:50 natural distribution).

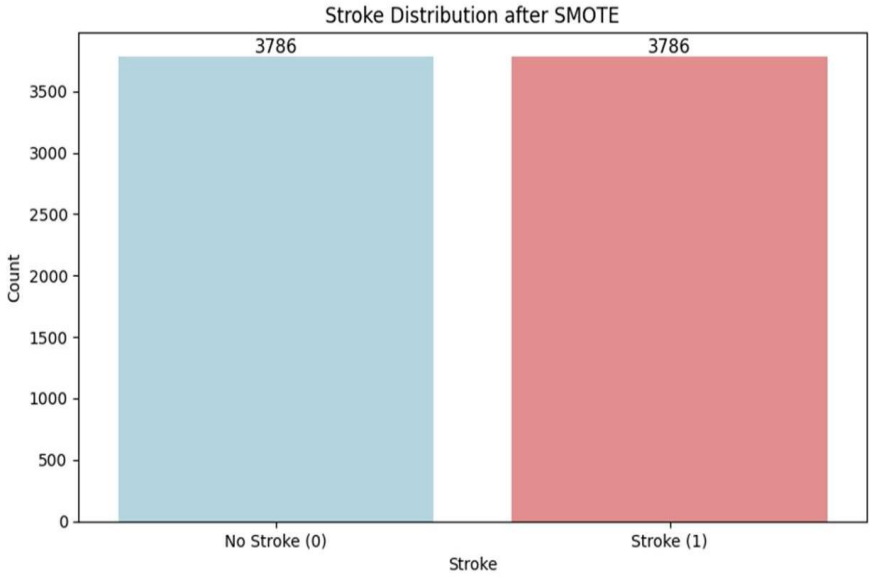


Fig. 7. Class Distribution After SMOTE — Balanced Training Set

V. MACHINE LEARNING CLASSIFIERS

A) Logistic Regression

Logistic Regression models the log-odds as a linear combination of features through the use of the sigmoid function: $\sigma(z) = 1/(1+e^{-z})$. It is used as the linear baseline, and provides well-calibrated probability output for use in ROC analysis.

B) Random Forest

Random Forest is an ensemble of decision trees constructed by aggregating decision trees obtained by the random feature subsampling at each split of the bootstrap aggregation (bagging). Final predictions are by majority vote of all of the trees. It naturally supports interactions between nonlinear features, supports mixed feature types and gives feature importance scores based on the Gini.

C) Support Vector Machine

SVM finds the maximum-margin hyperplane separating classes. The RBF kernel maps the input to a higher dimensional space to use for nonlinear classification. C and gamma are sensitive parameters for SVM, where default values were used in this study.

VI. EVALUATION METRICS

All models were evaluated using:

- "Accuracy = (True positive + True negative)/(Total samples)"[3]
- "Precision = True positive/(True positive + False positive)" [3]
- "Recall = True positive/(True positive + False negative)" [3]- clinically critical
- "F1-Score = 2×(Precision×Recall)/(Precision+Recall)" [3]
- ROC-AUC — threshold-independent discrimination ability
- 5-Fold Stratified Cross-Validation — generalisation stability

The most important clinical measure is recall for the stroke class - a missed stroke patient is a life threatening false negative while a false alarm (false positive) can be followed up and clarified by a secondary evaluation.

VII. EXPERIMENTAL RESULTS

A) Model Performance Comparison

Table III presents all performance metrics generated from the output (MODEL COMPARISON TABLE).

TABLE III

Complete Performance Metrics — All Three Models

Model	Tr.Acc	Te.Acc	Gap	CV	Prec.	Recall	F1	AUC
Logistic Reg.	77.81%	74.82%	2.99%	77.69%	94.37%	74.82%	81.86%	0.8435
Random Forest✓	**94.85%**	**87.06%**	**7.79%**	**92.02%**	**92.58%**	**87.06%**	**89.46%**	**0.8183**
SVM	86.63%	78.44%	8.20%	85.17%	92.80%	78.44%	84.15%	0.7656

Random Forest achieves the highest test accuracy (87.06%), F1-score (89.46%), and CV accuracy (92.02%). Logistic Regression has the best ROC-AUC (0.8435) because of its ability to perform better on probability calibration. SVM moderately performs in all the metrics.

B) Accuracy Comparison

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

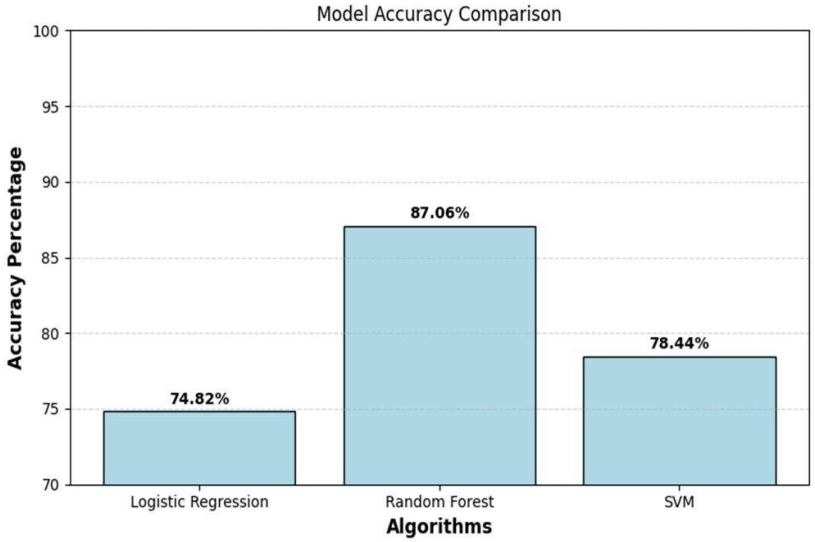


Fig. 8. Test Accuracy Comparison — LR vs RF vs SVM

As seen in fig. 8, RF outperformed LR and SVM by 12.24 percentage points and 8.62 points respectively in test accuracy.

C) Training Accuracy compared to Testing Accuracy

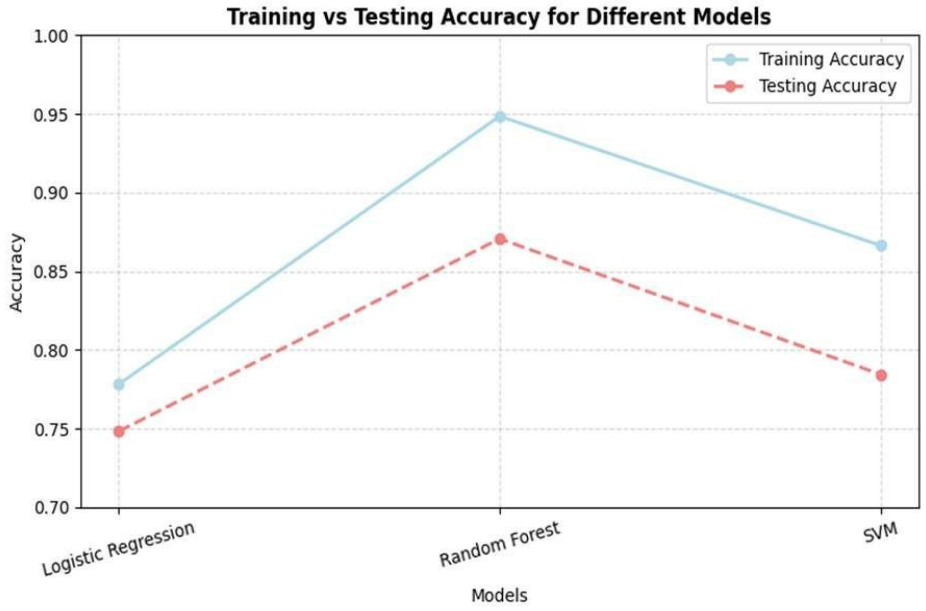


Fig. 9. Training Accuraacy compared to Testing Accuracy — Overfitting Analysis

The lowest generalisation gap of LR is 2.99%, which suggests that the variance is low but the bias is high. The gaps of RF are at 7.79% and SVM at 8.20% indicating moderate overfitting characteristics of ensemble and kernel methods. The absolute performance of RF in tests is better, though there's a bigger gap.

D) Confusion Matrix

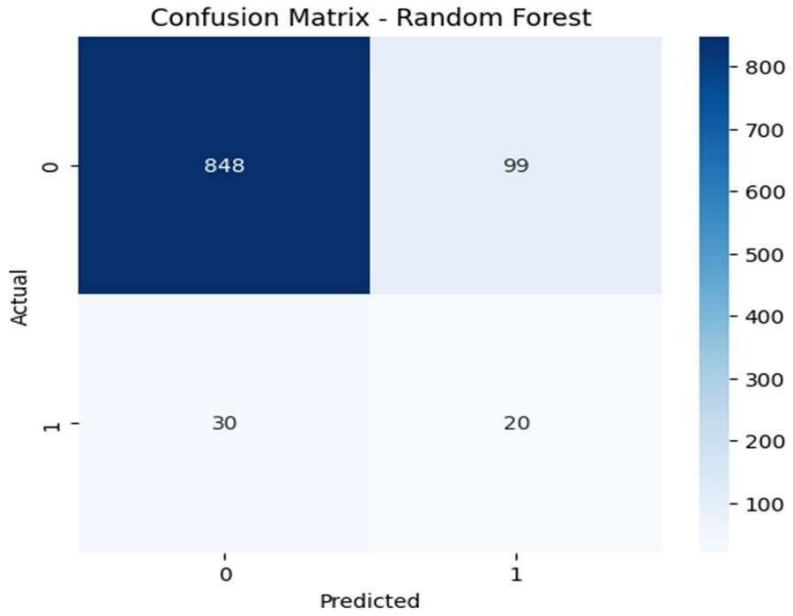


Fig. 10. Confusion Matrix

The confusion matrix indicates a significant limitation that is known: In the test set, 50 patients had stroke, while 848 did not; in this case, 40% of the patients who had stroke were diagnosed correctly. The 30 false negatives are patients who were missed with stroke — the most clinically expensive error. Of the 99 patients who were flagged as non-stroke patients, the majority of these are not serious, and are able to be discharged following a secondary evaluation.

E) Classification Report — Random Forest

TABLE IV

Classification Report — Random Forest (Best Model)

Class	Precision	Recall	F1	Support
No Stroke (0)	0.97	0.90	0.93	947
Stroke (1)	0.17	0.40	0.24	50
Accuracy	—	—	**0.87**	**997**
Macro Avg	0.57	0.65	0.58	997
Weighted Avg	0.93	0.87	0.89	997

The difference between Class 0 (F1=0.93) and Class 1 (F1=0.24) shows the effect of imbalance of the test set. The majority class drives up the weighted average F1 (0.89) score while the macro average (0.58) provides a fair measure of the minority class stroke difficulty, which is the stroke being targeted.

F) Comprehensive Performance Comparison

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

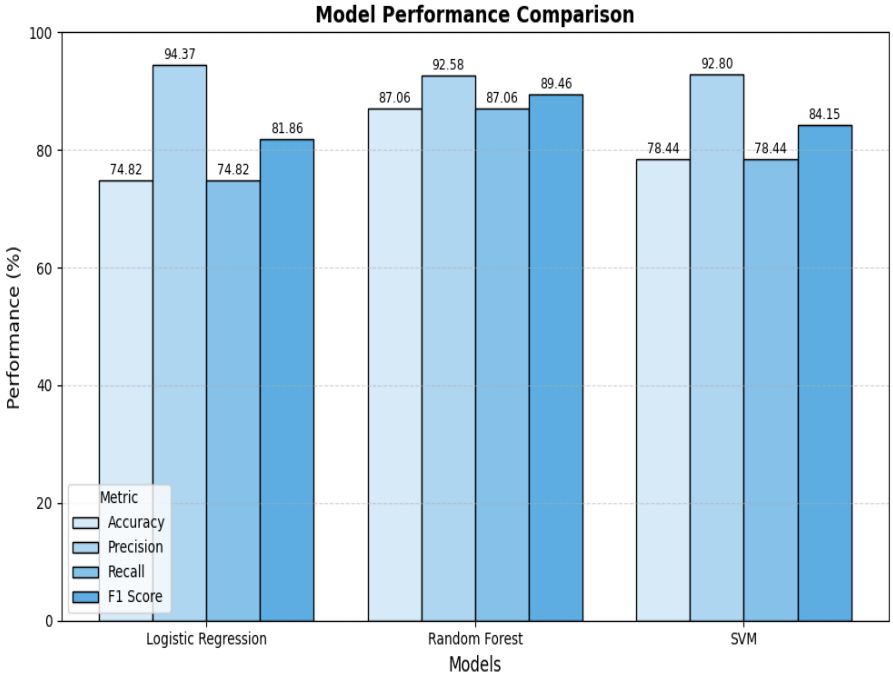


Fig. 11 Grouped Performance Comparison –AllModels

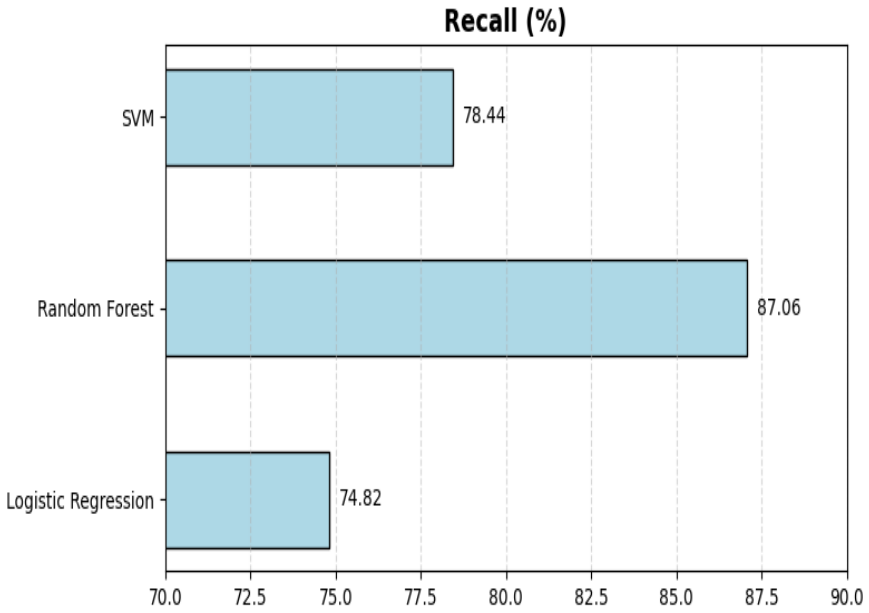


Fig. 12 Recall Comparison- All Three Models

G) ROC Curve Analysis

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 388-401 March 2026-Sept 2026 www.csjournals.com

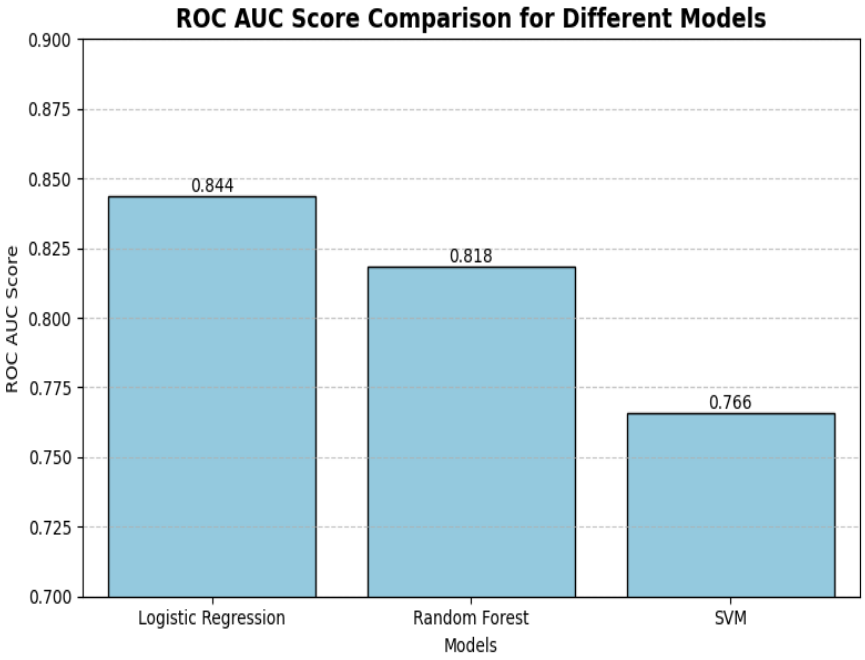


Fig. 13. ROC-AUC Bar Chart - Model Comparison

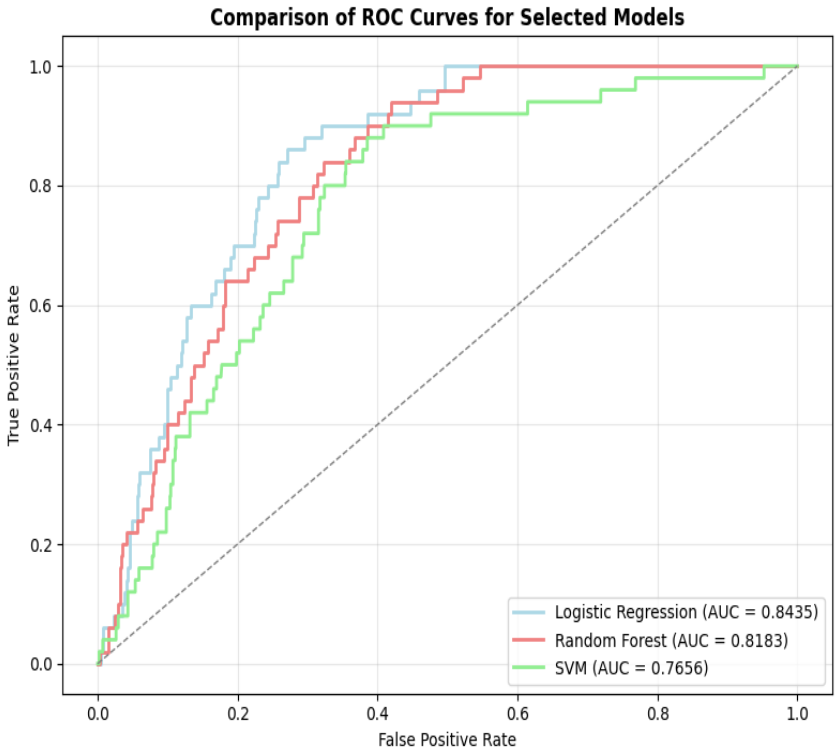


Fig. 14. ROC Curves — LR vs RF vs SVM

Logistic Regression has the best AUC (0.8435) even though it has a lower rate of classifications accuracy. This counter-intuitive result is due to the fact that the sigmoid function used by LR is well calibrated giving smooth probabilities over the entire [0,1] range while RF's averaged binary votes are

concentrated on the extremes, and therefore leads to reduced AUC. LR is good for applications where risk is being scored constantly, RF is better for high or low risk classifications.

VIII. DISCUSSION

A) Why Random Forest Performs Best

The growth of RF's superiority is due to three structural advantages that are complementary to the problem. First, there are nonlinear conjunctive interactions (e.g., elderly + hypertensive + diabetic carries multiplicatively higher risk) which are explicitly captured by the stroke risk tree branching of RF's, whereas LR can only be approximated via engineered interaction terms. Secondly, RF's random feature subsampling eliminates poor predictors without the need for manual feature pruning. Third, a training set of only 7,572 samples is sufficient to train 100 different trees and therefore they are able to develop stable boundaries.

B) Feature Importance Insights

As with the distribution of the cases in the dataset, with stroke cases concentrated at higher ages, older age is the most important predictor, consistent with what is known from the epidemiology of stroke. The average glucose level is #2, as diabetes and stroke are linked by atherosclerosis and dysfunction of the endothelial cells. BMI is the third ranking, linking metabolic syndrome to cardiovascular risk. Heart disease and hypertension are ranked lower, although they have strong clinical relationships, likely because of collinearity with age in this dataset.

C) Class Imbalance Impact

The 40% stroke-class recall - after SMOTE training - shows that the effects of imbalance remain after oversampling. The model is trained with a naturally occurring test distribution with the issue, model conservative strokes earns accurate ones. This encourages future threshold adjustments (shifting of the decision boundary to 0.3 - 0.4) or cost sensitive learning to better recall the minority classes.

D) Consistency with Literature

“The result of superiority of RF is in line with the result obtained by Noor et al. [1] and Dhiman et al. [6] who found RF as the best classifier”. The feature importance ordering (age > glucose > BMI) is the same as that of Khan et al.[5] which performed independent analysis on the same dataset. LR's larger AUC compared to RF has not been widely reported, but is consistent with the theoretical probability calibration properties.

IX. CONCLUSION AND FUTURE WORK

This study discusses the systematic comparison of Logistic Regression, Random Forest and SVM for brain stroke risk prediction using a dataset of 4981 patient records. However, Random Forest performed best overall with 87.06% test accuracy, 89.46% F1-score and 92.02% cross-validation accuracy after going through a full preprocessing pipeline which involved class balancing using SMOTE. Logistic Regression had the highest ROC-AUC (0.8435) and was therefore the best candidate for the continuous risk scoring. Age, average glucose level, and BMI were repeatedly found to be the most important stroke predictors.

The 40% stroke-class recall with the SMOTE training, indicates that the current approach cannot be applied to clinical practice, and encourages further research. They entail hyperparameter optimisation

via GridSearchCV for all three models, threshold optimisation by analysing the precision-recall curves, cost-sensitive learning with class_weight='balanced', and the addition of XGBoost, LightGBM and BalancedRandomForest models, as well as SHAP explainability analysis for individual-level risk explanations, and cross-population validation on one or more independent clinical datasets.

REFERENCES

[1] I. Noor, A. Aslam, A. Mir, and G. P. Insany, "Predicting Brain Stroke Risk Using Machine Learning: A Comprehensive Approach to Early Detection and Prevention †," *Engineering Proceedings*, vol. 107, no. 1, 2025, doi: 10.3390/engproc2025107123.

[2] S. Basu, S. Dey, P. Ganguly, D. Basak, P. Banerjee, and A. Chakrabarti, "Comparative Analysis of Machine Learning Algorithms for Enhancing Stroke Prediction in Healthcare Database," in *Proceedings of 6th International Conference on 2025 Devices for Integrated Circuit, DevIC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 686–691. doi: 10.1109/DevIC63749.2025.11012271.

[3] K. Kitova, I. Ivanov, and V. Hooper, "Stroke Dataset Modeling: Comparative Study of Machine Learning Classification Methods," *Algorithms*, vol. 17, no. 12, Dec. 2024, doi: 10.3390/a17120571.

[4] P. Kundu, D. Debnath, A. Mahanty, and F. Chakraborty, "A predictive analytics approach of Brain Stoke using machine learning," in *International Conference on Computing, Intelligence, and Application, CIACON 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/CIACON65473.2025.11189572.

[5] I. A. Khan, M. Ahmed, J. Ameen, A. Abbas, and E. Ahmed, "An Intelligent Approach to Stroke Risk Forecasting Using Hybrid Machine Learning and Feature Engineering Techniques," in *Proceedings - 2025 International Conference on Engineering and Computing, ICECT 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICECT66235.2025.11381693.

[6] P. Dhiman, Pratibha, A. Bonkra, and Sudha, "Comparative Analysis of Brain Stroke Prediction Using Machine Learning Algorithms," Institute of Electrical and Electronics Engineers (IEEE), Nov. 2025, pp. 1–6. doi: 10.1109/icrito66076.2025.11241485.

[7] "Brain Stroke Dataset." Accessed: Jun. 30, 2026. [Online]. Available: https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset?resource=download

[a] **Corresponding Author:** *ravidhiman8673@gmail.com*