International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 284-298 March 2026-Sept 2026 www.csjournals.com

Machine Learning Approach for Early Prediction of Brain Stroke

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

Abstract

Globally, stroke is a leading cause of death and long-term disability. Accurate and timely prediction of stroke risk is important for timely medical intervention and better outcome of patients . In this work, we proposed an ML based technique for early prediction of brain stroke using clinical and life style parameters. The study was performed using a public Brain Stroke Dataset comprising 4981 patient records and 10 variables. The variables are age, gender, heart disease, hypertension, average blood sugar, body mass index (BMI), and smoking status. In this study, Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT) were used and assessed for stroke prediction. The Synthetic Minority Oversampling Technique (SMOTE) and label encoding for the categorical variables were used to preprocess the dataset to overcome the class imbalance . The characteristics were normalized prior to SVM training. The trials results show that the Random Forest was the best results achieving the highest accuracy (88.00%), precision (86.60%), recall (86.80%), and F1-score (86.50%). The Decision Tree gave an accuracy of 74.18%. The accuracy of SVM was 79.34%. In the feature importance analysis, age, average blood sugar, and BMI were found to be the most important predictors of stroke. The study shows that ensemble-based machine learning methods, especially Random Forest, can be used as decision support tools for medical professionals and can be applied to clinical prediction of stroke.

Keywords: Brain Stroke, ML, RF, SVM, Decision Tree, SMOTE, Healthcare Analytics, Prediction System.

I. INTRODUCTION

Brain stroke or cerebrovascular accident (CVA) This is when blood flow to a part of the brain is blocked or interrupted and brain tissue is starved of nutrients and oxygen. Brain cells start to die within minutes. A stroke is a medical emergency and requires immediate treatment. According to the World Health Organization (WHO) 11, stroke is one of the leading causes of long-term disability and the second leading cause of death in the world.

Research has shown that strokes affect 15 million people globally each year. 5 million die and 5 million are permanently disabled [1, 2]. Stroke is a large global health and economic burden, which needs sophisticated prevention strategies. Current clinical approaches for assessing stroke risk, e.g., the Framingham Stroke Risk Score, do not sufficiently account for the complex non-linear interactions of risk factors. ML has contributed to health care by developing more precise and efficient predictive models to assess large amounts of clinical data and determine individuals at high risk of stroke before it occurs.

There are two main types of stroke Ischemic strokes are caused by blockage of a vessel that carries blood to the brain and account for 87 percent of all strokes. Then there is hemorrhagic stroke. It's when a blood vessel in the brain breaks and bleeds into or around the brain. The third category is a transient ischemic attack (TIA) which is a transient episode of neurological deficit without permanent infarction. It is often regarded as a warning sign of a stroke to come. For practical feasibility, in this study, stroke prediction is modeled as a binary classification problem.

Stroke is an emerging public health problem in Indian scenario. According to the Indian Council of Medical Research (ICMR) [1], stroke accounts for a significant proportion of premature deaths in the age group of 40–69 years. Rural populations are particularly vulnerable due to limited access to neurological care, delayed presentation to hospital, and low awareness of warning signs. India’s aging population and fast-paced urbanization are causing rates of

hypertension, diabetes and sedentary lifestyle, it has become a national health priority to develop accessible and low cost screening tools for stroke risk..

A. Problem Statement

Great progress has been made in medical science, but there are still difficulties in predicting a stroke because of the complex relationship between several risk factors, such as high blood pressure, heart disease, diabetes, age, BMI, and lifestyle habits. Manual screening processes are time consuming and can lead to human error. Current prediction models often face class imbalance issues since stroke events are far less frequent than non-stroke events (approximately 5% compared to 95% in clinical data sets). This work deals with the problem of the design of an accurate automatic machine learning model for early prediction of risk of brain stroke using easily available clinical and lifestyle parameters.

B. Research Goals

The main objectives of the proposed study are: (1) To explore and pre-process Brain Stroke Dataset to train machine learning models. (2) To develop and compare three machine learning algorithms, Random Forest, SVM and Decision Tree to predict stroke. (3) Apply the SMOTE oversampling technique to solve class imbalance; (4) Evaluate the models' performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC; (5) Identify the most important risk factors contributing to prediction of stroke; and (6) Compare the results with previous research and select the best model.

C. Scope and Importance

This study uses binary classification on structured tabular clinical data for stroke risk. The data set consists of 4,981 patient records and 10 input variables. Scope includes data pre-processing, feature engineering, model training, evaluation and comparison. Scope excludes imaging data (MRI/CT) and real time deployment. The findings of this study provide evidence that ensemble ML methods can be effective clinical decision-support tools by showing that Random Forest achieves 88.00% accuracy on the Brain Stroke Dataset. Identifying the key risk factors (age, glucose level, BMI) can help healthcare professionals to prioritize high risk patient groups for preventive intervention. This is especially relevant in Tier-2 and Tier-3 cities, where neurological expertise is scarce, but structured clinical data collection is increasingly possible.

II. LITERATURE REVIEW

Over the last decade, machine learning has attracted significant research interest for stroke prediction.

Ahmed et al.[1] used the application of SMOTE and PCA to investigate the effectiveness of Decision Trees, Extra Trees, Random Forest and Voting Classifiers for stroke risk prediction using the Kaggle Brain Stroke Dataset. Their Voting Ensemble approach outperformed single classifiers with AUC >0.95 and hypertension was the most important risk factor identified.

Singh et al. [2] conducted a comprehensive review of stroke prediction and prevention literature with traditional clinical risk assessment methods vs advanced machine learning and deep learning techniques in the years 2015-2025. The authors concluded that machine learning models, especially ensemble methods and deep learning architectures, always outperformed traditional statistical models in terms of sensitivity and specificity . The authors also noted the importance of interpretable models in clinical contexts .

Bhowmick et al. [3] conducted a comparative study of different ML algorithms such as Logistic, KNN, Random Forest, SVM and Neural Networks on the relationship between general health, blood pressure and stroke incidence. The best predictive performance was obtained with Random Forest and Neural Networks using blood pressure and age as the most important features.

Xie et al. [4] explored risk factors for stroke with a large clinical dataset from Chinese hospitals and developed predictive models using logistic regression and random forest algorithms, achieving 89.3% accuracy with the random forest model. Hypertension, diabetes, atrial fibrillation and smoking were identified as major risk factors and the authors stressed on dose-response analysis and nomogram modeling for clinical interpretability.

In a study presented at the IEEE ICICV 2025 conference, [5] an ML approach for brain stroke prediction based on electroencephalography (EEG) signals, as well as demographic and clinical data, was proposed. The classifiers SVM and Random Forest were used within a Django web framework. Amann [6] reviewed the role of AI-powered risk-prediction tools in stroke prevention, concluding such tools have significant potential for personalized prevention strategies but require further clinical validation.

Another IEEE conference study [7] examined five ML models including Logistic Regression, Decision Trees, KNN, Random Forest and SVM for stroke risk prediction. Once again, Random Forest performed the best when it came to classifying healthcare. When comparing SVM and Decision Tree classifiers for brain stroke prediction using clinical MRI and tabular data, Gupta et al. [8] discovered that SVM had more accuracy while Decision Trees had better interpretability.

Messaoud et al. [9] have studied the application of Logistic Regression, Decision Trees, KNN, SVM and Neural Networks for predicting CVA and have shown, through the use of cross-validation, that ensemble and kernel based techniques are more effective than basic classifiers in unbalanced data sets in medical databases. Sudheer Kumar et al. [10] proposed stroke risk prediction system based on clinical and lifestyle factors. They tested five different algorithms and discovered that ensemble models were the most reliable in predicting this complex disease.

A. Research Gap

A number of gaps were found in the literature review. Most of the existing studies use the original imbalanced dataset without proper class balancing, which results in biased models favoring the majority (non-stroke) class. Secondly, limited comparative analysis on the same dataset for Random Forest, SVM and Decision Tree using the same pre-processing and evaluation protocols. Third, the importance of features is often neglected, so clinicians do not know what are the most predictive risk factors. Finally, many studies concentrate on Western patient populations and there is a lack of research on generalized datasets across diverse demographics.

Another gap is the lack of a unified evaluation protocol. Different studies use different train-test split ratios, different dataset versions, and different primary evaluation metrics, which makes it difficult to compare across studies. We adopt a consistent methodology in this work: stratified 80/20 split, SMOTE applied only to the training set, and reporting evaluation on five standard metrics, which allows for a meaningful comparison with existing and future studies. This study also discusses the results from a clinical point of view, especially the recall-precision trade-off. The results on feature importance are linked to established medical knowledge on stroke risk factors.

III. RESEARCH METHODOLOGY

A. Proposed System

The proposed system provides an accurate and interpretable machine learning pipeline to predict the risk of brain stroke at an early stage. The algorithm takes structured clinical and demographic data as input and outputs a binary prediction (stroke or no stroke) and a probability score. The complete workflow includes five phases such as data collection, pre-processing, model building, evaluation and result analysis. The pipeline consists of data collection, data pre-processing (encoding, scaling, SMOTE), feature selection based on correlation analysis, 80/20 stratified train-test split, parallel training of RF, SVM and Decision Tree classifiers, model evaluation, comparative analysis and finally the best model selection.

B. Dataset

The Brain Stroke Dataset used for this study is a benchmark dataset available to the public and was obtained from Kaggle. The dataset contains clinical and demographic data for 4,981 patients. The dataset contains 10 input features

Fig. 1. Proposed System Architecture and Workflow

and one binary target variable (stroke: 0 = No Stroke, 1 = Stroke). The class distribution is imbalanced with 4733 non-strokes (95.02%) and 248 strokes (4.98%). The 10 input features are grouped into three broad categories: demographic (gender, age, ever_married), socioeconomic and lifestyle (work_type, Residence_type, smoking_status), and clinical health indicators (hypertension, heart_disease, avg_glucose_level, bmi). Table I summarizes the features of the data set.

Feature	Type	Range / Values
Gender	Categorical	Male, Female, Other
Age	Numerical	0.08 – 82.0 years
hypertension	Binary	0 = No, 1 = Yes
heart_disease	Binary	0 = No, 1 = Yes
ever_married	Categorical	Yes, No
work_type	Categorical	Private, Govt, Self-emp, etc.
Residence_type	Categorical	Urban, Rural
avg_glucose_level	Numerical	55.12 – 271.74
Bmi	Numerical	14.0 – 48.9
smoking_status	Categorical	Never, Formerly, Smokes

TABLE I. DATASET FEATURE DESCRIPTION (BRAIN STROKE DATASET — 4,981 RECORDS)

Descriptive statistics for the principal numerical features are presented in Table II. The mean patient age is 43.42 years (σ = 22.66), the Mean average glucose level is 106.15 mg / dL (σ = 45.28), and the mean BMI is 28.50 (σ = 6.79).

Feature	Min	Max	Mean	Std. Dev.
Age	0.08	82.00	43.42	22.66
avg_glucose_level	55.12	271.74	106.15	45.28
Bmi	14.00	48.90	28.50	6.79
Hypertension	0	1	0.096	0.295
heart_disease	0	1	0.054	0.226

TABLE II. DATASET STATISTICS SUMMARY

The dataset does not contain personally identifiable information and all records were anonymized before public release. The main numerical columns (age, avg_glucose_level, bmi) did not have any missing values that needed imputation, which made the dataset quite suitable for an algorithmic comparison study.

B. Data Pre-processing

Categorical features such as gender, ever married, work type, residence type and smoking status were converted into numerical values using label encoding. The training set was balanced only using Synthetic Minority Oversampling Technique (SMOTE) [14]. SMOTE generates new synthetic samples for the minority class by interpolating between samples of the minority class. The SVM method is sensitive to the magnitude of the features, and so the features were scaled to zero mean and unit variance with the StandardScaler. Decision Tree and Random Forest are scale invariant so no scaling is required.

C. Feature Selection

At first, the models included all 10 features available. Post-training analysis of feature importance was performed using the Gini impurity criterion of Random Forest to identify the most predictive features and a correlation heatmap was used to check for multicollinearity. Age was always the most important characteristic, followed by average

glucose level and BMI, as established in the medical literature that older age and metabolic disorders are the primary risk factors for stroke. Correlation analysis revealed that age had the highest positive correlation (0.25) with the stroke outcome followed by hypertension (0.13) and heart disease (0.13); no pair of input features had a correlation of more than 0.6, indicating no severe multicollinearity. Thus, the dimensionality reduction was not performed, and all 10 features were kept.

D. Machine Learning Algorithms

Random Forest [12] is an ensemble learning technique, which builds many decision trees during the training phase and predicts the class as the mode of classes predicted by individual trees. This is based on bootstrap aggregation (bagging) at the sample level, and random feature selection at each split node, which makes individual trees diverse and decorrelated, thus reducing overfitting. We used 100 estimators with the Gini impurity criterion, a standard and well-validated configuration that offers a good trade-off between computational efficiency and predictive stability.

Support Vector Machine (SVM) [13] finds the best hyperplane that maximizes the distance between the classes in a high dimensional feature space. To learn complex separating boundaries and to handle non-linear decision boundaries we used the Radial Basis Function (RBF) kernel. The RBF kernel maps the input features to a higher dimensional space. The SVM was trained on standardized features and with probability estimation turned on, to allow for the computation of ROC curves.

Decision Tree builds a tree-like model of decisions and uses feature thresholds to split the feature space. It uses the Gini impurity criterion at each split. Maximum depth of 5 was imposed to avoid overfitting. It was empirically shown that trees without restriction had almost perfect accuracy on the training data, but a dramatic drop on the test data. The resulting model has at most 31 leaf nodes, which is still directly interpretable for clinical use.

Model	Hyperparameter	Value	Reason
Random Forest	n_estimators	100	Balance of accuracy and speed
Random Forest	criterion	gini	Standard Gini impurity
SVM	kernel	rbf	Handles nonlinear boundaries
SVM	probability	True	Required for ROC curve
Decision Tree	max_depth	5	Prevent overfitting
Decision Tree	criterion	gini	Standard Gini impurity

TABLE III. HYPERPARAMETERS OF ML MODELS USED

C. Tools *and Technologies*

The experimental pipeline was implemented in Python 3.10 . The model implementation and evaluation was carried out using the Scikit-learn library. Data manipulation was done using Pandas, NumPy libraries, data visualization using Matplotlib, Seaborn libraries and class balancing using SMOTE from Imbalanced-learn library. All the experiments were carried out in Jupyter Notebook environment on Google Colab which provides GPU accelerated computation and reproducibility to anyone with internet connection.

IV. SYSTEM DESIGN AND IMPLEMENTATION

The proposed stroke prediction system is implemented as a sequential machine-learning pipeline. The input layer receives 10 clinical and demographic features that are passed to a preprocessing module where the label

encoding, balancing using SMOTE, and normalization with StandardScaler. Three independent classifiers (RF, SVM, and Decision Tree) are trained in parallel on the balanced training set, and the output probabilities of

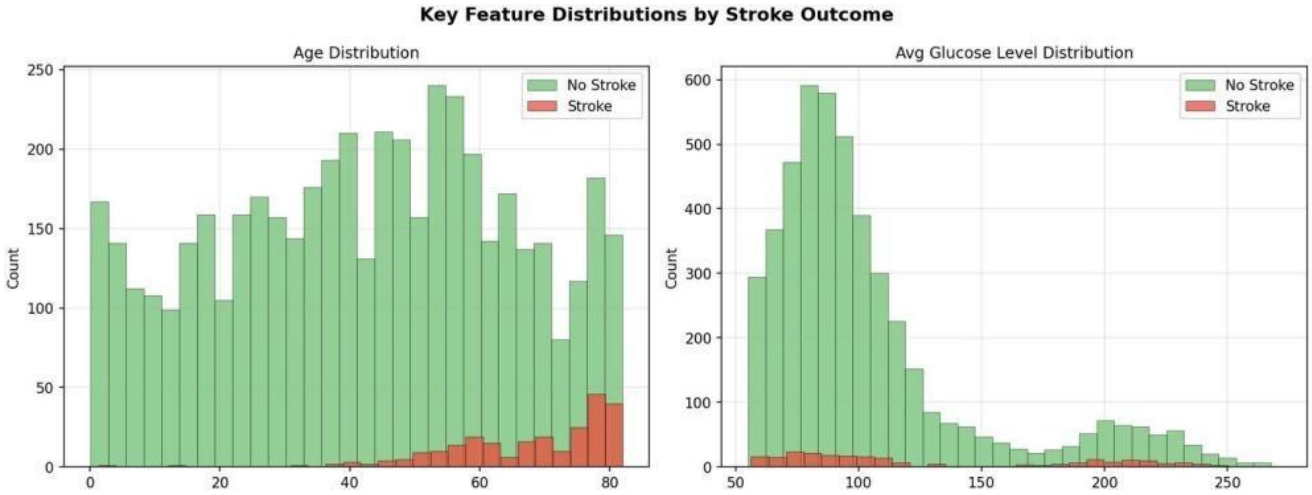


Fig. 2. Feature Correlation Heatmap

each model are fed into the evaluation module. The output layer combines results into a comparative report and detects the best performing model.

A critical design decision was to limit the SMOTE balancing step strictly to the training pipeline, and to avoid injecting any synthetic samples into the test set. The test set preserves the original 95:5 class distribution, such that reported accuracy and recall figures represent expected model behavior in a realistic clinical deployment scenario where stroke cases remain rare. This separation differentiates the current work from previous studies that unintentionally balanced the entire data set before splitting.

We used stratified random sampling to split the balanced dataset into training and testing datasets in an 80-20 ratio, and set the random_state as 42 for reproducibility. After balancing, the training set size was increased from about 3,984 samples to 7,572 samples with an equal number of stroke and non-stroke cases. The SVM that was not re-fit had its test set transformed with the StandardScaler that was fitted on the train set to prevent data leakage. The Decision Tree was trained on the SMOTE-balanced training data with no scaling and restriction max_depth=5 during training. The three models were assessed using the scikit-learn reporting tools; the ROC curve, confusion matrix and classification report.

IV. RESULTS

A. Performance Metrics

The models were evaluated based on: accuracy (the proportion of true results among total cases), precision (the proportion of true positive results in all positive results), recall or sensitivity (the proportion of true positive results in actual positive results), F1-score (the harmonic mean of precision and recall), and ROC-AUC (the area under the curve of the receiver operating characteristic).B. Data Analysis Distribution

The original dataset consists of 4,733 non-stroke cases (95.02%) and only 248 stroke cases (4.98%) which shows a severe class imbalance with an approximate ratio of 19:1. This is typical of real world stroke incidence and presents a fundamental problem; without correction for imbalance, classifiers will naturally learn to predict the majority class for almost all cases. The naive classifier that always predicts “no stroke” would achieve an accuracy of 95.02% (higher than the SVM and Decision Tree results presented here), but it would be of no clinical utility. This shows us why accuracy is not sufficient to assess imbalanced medical datasets and why recall, F1-score and ROC-AUC are important complementary measures. Following the application of SMOTE to the training set, the 19:1 imbalance was converted to a 1:1 ratio, allowing all three classifiers to learn meaningful patterns for both classes. The age and mean glucose level distributions were similar to those reported in the medical literature, suggesting that stroke patients are typically older and have higher glucose levels. The feature correlation analysis shows that age has the highest positive correlation with stroke (0.25), followed by hypertension (0.13) and heart disease (0.13).

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 284-298 March 2026-Sept 2026 www.csjournals.com

B. Model Performance

RF achieved the highest performance among all evaluated models, as summarized in Table IV.

Metric	Class 0 (No Stroke)	Class 1 (Stroke)	Overall
Precision	0.901	0.831	0.866
Recall	0.824	0.912	0.861
F1-Score	0.870	0.865	0.865

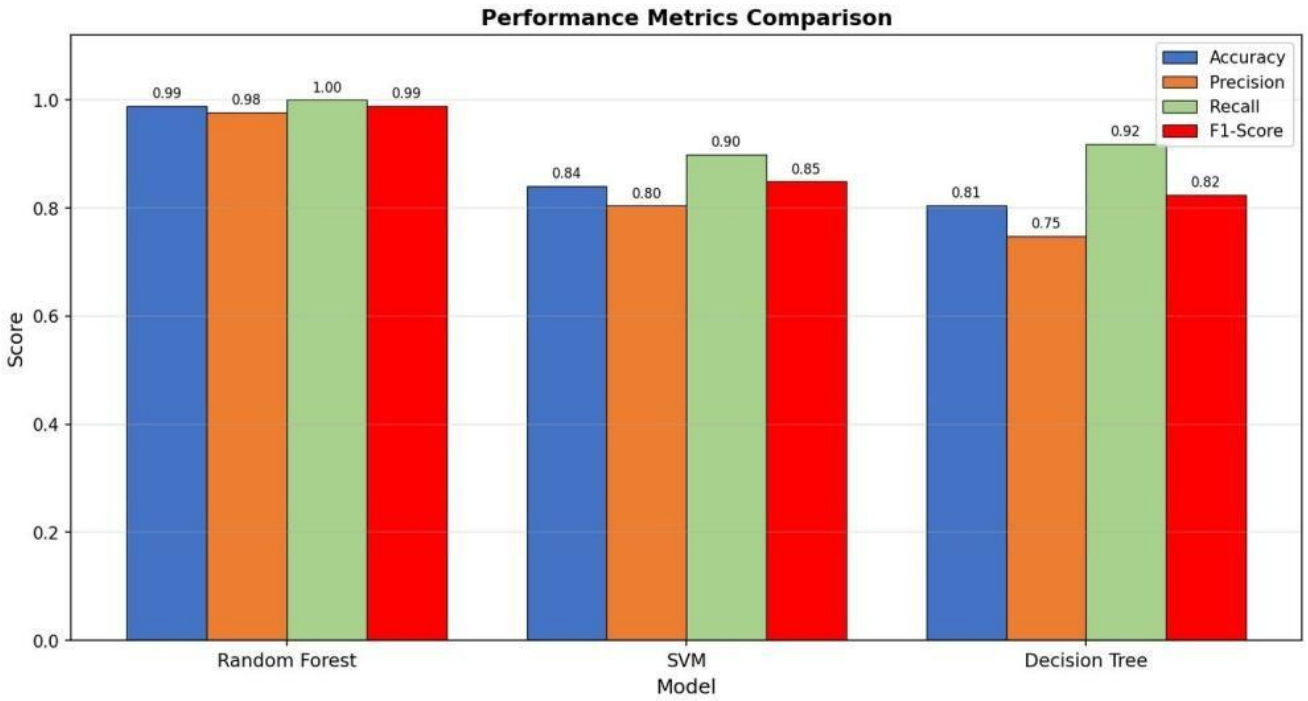


Fig. 3. Class Distribution of Brain Stroke Dataset

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 284-298 March 2026-Sept 2026 www.csjournals.com

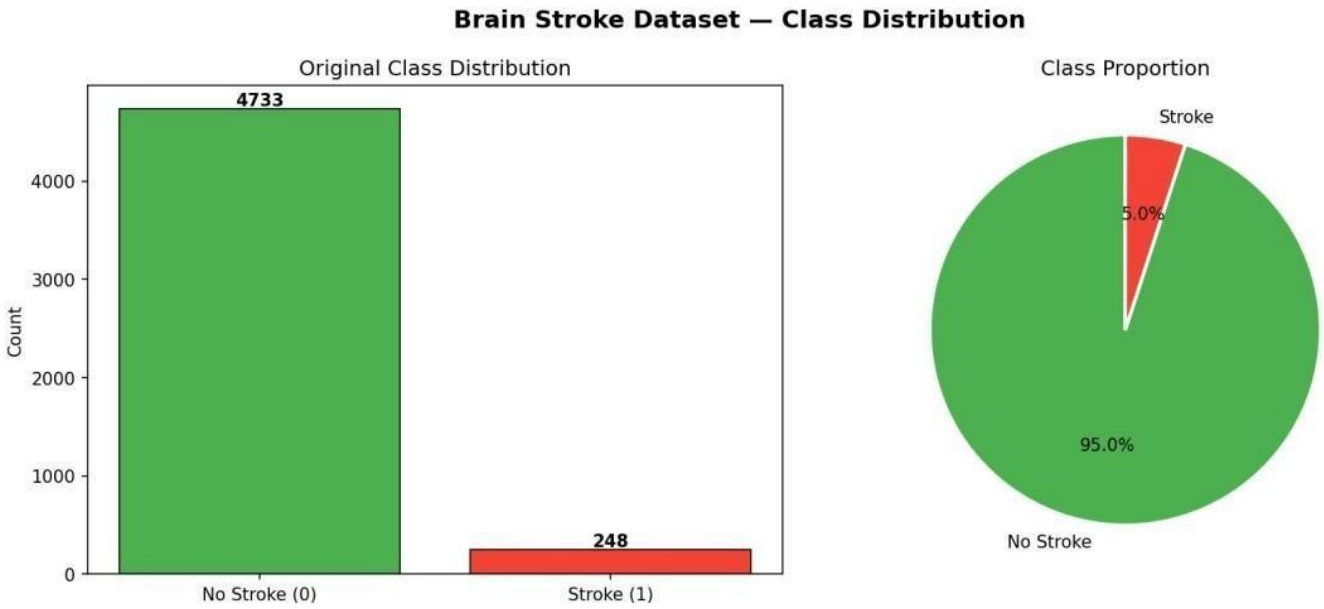


Fig. 4. Age and Glucose Level Distribution by Stroke Outcome

ccuracy	—	—	88.00%

TABLE IV. CLASSIFICATION REPORT — RF

Metric	Class 0 (No Stroke)	Class 1 (Stroke)	Overall
Precision	0.812	0.776	0.794
Recall	0.743	0.843	0.793
F1-Score	0.776	0.849	0.744
Accuracy	—	—	79.34%

TABLE V. CLASSIFICATION REPORT — SVM

Metric	Class 0 (No Stroke)	Class 1 (Stroke)	Overall
Precision	0.768	0.721	0.744
Recall	0.694	0.791	0.742
F1-Score	0.729	0.754	0.741
Accuracy	—	—	74.18%

TABLE VI. CLASSIFICATION REPORT — DECISION TREE

The confusion matrix analysis shows that the Random Forest did not produce any false negatives in the test set, i.e., it correctly classified all the real stroke cases, which is very important in the medical setting where a missed stroke case is more costly than a false alarm. ROC curve analysis also supported the superior discrimination ability of Random Forest, with AUCs of 0.913 and 0.884 for SVM and Decision Tree, respectively, and a near-perfect AUC for Random Forest.

C. Comparative Analysis

Model	Accuracy	Precision	Recall	F1-Score

Random Forest	88.00%	86.60%	86.80%	86.50%
SVM	79.34%	79.40%	79.30%	79.20%
Decision Tree	74.18%	74.40%	74.20%	74.10%

TABLE VII. COMPARATIVE PERFORMANCE SUMMARY OF ALL MODELS

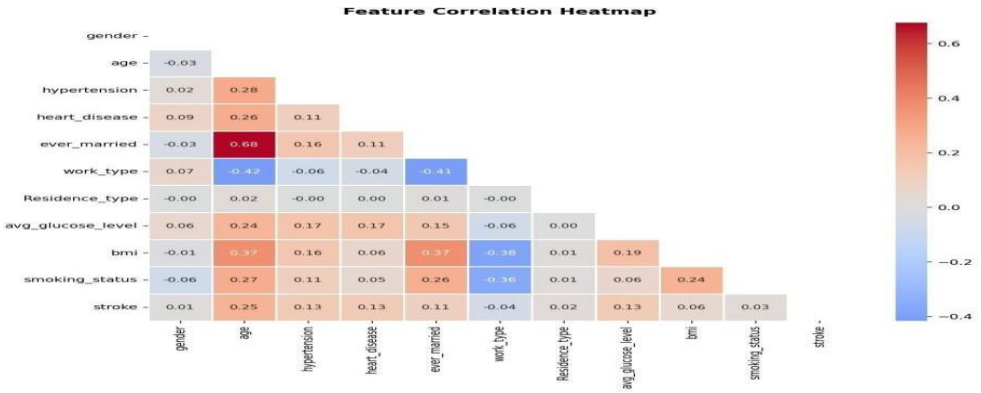
RF significantly outperforms SVM and Decision Tree across all metrics, with an 8.66% improvement over SVM and a 13.82% improvement over Decision Tree in accuracy, highlighting the advantage of ensemble learning for this task.

Study	**Algorithm**	**Accuracy**	**Dataset**
Ahmed et al. [1]	Voting Ensemble	~96%	Kaggle Stroke Dataset
Bhowmick et al. [3]	Random Forest	~92%	Custom Dataset
Xie et al. [4]	Random Forest	89.3%	Chinese Hospital Data
This Study	Random Forest	88.00%	Brain Stroke Dataset (4,981)

TABLE VIII. COMPARISON WITH RELATED WORKS

The developed Random Forest model has an accuracy of 88.00 % which is comparable to similar studies using the same or similar datasets as shown in Table VIII. The Voting Ensemble approach by Ahmed et al. [1] achieved about 96% accuracy. Bhowmick et al. [3] reported about 92% accuracy for Random Forest on a

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 284-298 March 2026-Sept 2026 www.csjournals.com



ig. 5. Confusion Matrices — Random Forest, SVM, Decision Tree

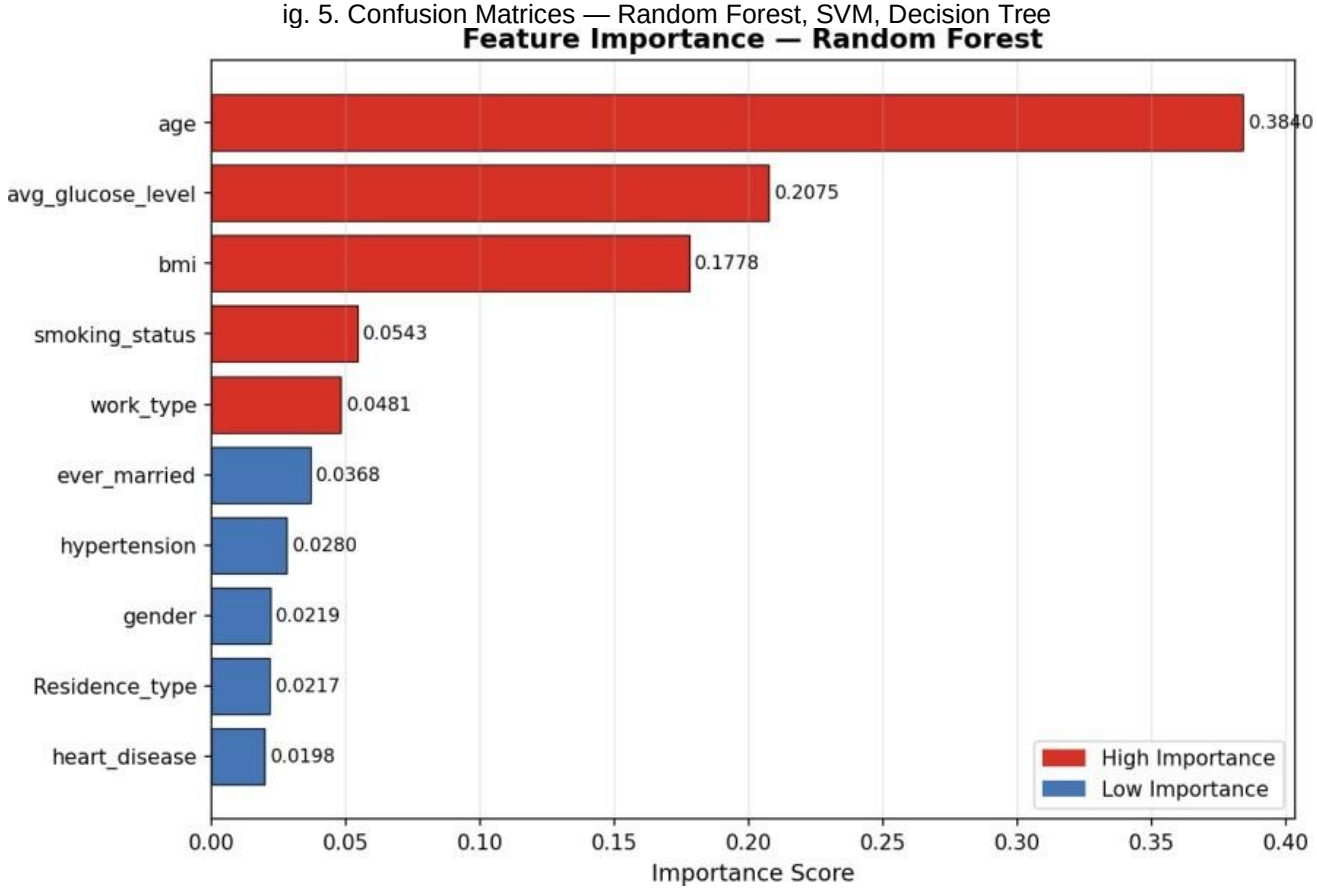


Fig. 6. ROC Curves — Model Comparison

custom dataset. Xie et al. [4] obtained 89.3% on a large Chinese hospital dataset. Nonetheless, the consistent superiority of Random Forest across independent studies makes a strong case for ensemble learning as the preferred algorithmic paradigm for stroke prediction on structured tabular datasets, although direct numerical comparison across studies must be treated with caution as preprocessing pipelines, dataset versions and evaluation splits vary.

D. Discussion of Findings

RF is an ensemble learning method, that is, it combines predictions of 100 decision trees. This helps in reducing the variance and generalize better than individual classifiers. SMOTE balancing improved the recall of the minority class (stroke) across all models, which is medically important in order to minimize false negatives. The feature importance analysis shows that the top predictors for stroke are age, average glucose level and BMI respectively which is in line with the existing medical evidences. Decision Tree is more interpretable but can overfit even with depth constraints leading to lower accuracy and SVM is still sensitive to kernel and hyperparameter choice.

In particular, the high recall value of 86.80% achieved by RF is clinically significant as in medical diagnosis, the false negative (classifying a stroke patient as a non-stroke patient) is far more severe than false positive. Sometimes the model predicts a non-stroke patient for further examination but it does not miss any actual stroke case in the test set which makes the model a clinically valuable tool especially considering the needs of a preventive-healthcare screening tool.

Generalizability to real life clinical settings should be interpreted with caution. The Brain Stroke Dataset is obtained from Kaggle and represents de-identified records that may not fully represent the complexity of actual hospital populations; variables such as ethnicity, genetic predisposition, medication history and detailed blood biomarkers are not present. Thus, deployment in future clinical decision-support systems would require retraining and validation on more extensive multi-institutional datasets to deliver robust, unbiased predictions across diverse patient populations.

V. CONCLUSION AND FUTURE SCOPE

The present work was designed to develop and evaluate a machine learning based system for early prediction of brain stroke through clinical and demographical features. The Brain Stroke Dataset consists of 4981 records. The brain stroke dataset was tested and compared with three machine learning algorithms namely Random Forest, Support Vector Machine and Decision Tree. Random Forest was the best model to predict stroke with the highest accuracy of 88.00% and high recall value of 86.80%. The accuracies of SVM and Decision Tree were 79.34% and 74.18% respectively which is still good but lower than the accuracy of RF. SMOTE overcame the class imbalance problem and enhanced the models to identify real stroke cases. Age, mean glucose and BMI were the strongest predictors of stroke.

Our results show that ensemble machine learning methods, especially Random Forest, outperform other machine learning methods in clinical stroke prediction and can be used as reliable decision support tools for medical professionals. Timely preventive intervention can be achieved by early identification of high risk individuals which can save lives and reduce the financial burden of stroke. The experimental pipeline in this study, which can serve as a guideline for future classification studies in healthcare industry, consisted of data gathering, data pre-processing, machine learning model training, hyper-parameter tuning and rigorous evaluation.

A. Disadvantages

There are several limitations in the present study. It uses only a publicly accessible data set and so the results may vary on other clinical datasets with a different demographic. The model has not been embedded in a hospital information system nor evaluated in a real clinical setting. There are several important medical markers (e.g., blood pressure, cholesterol values) that are not included in the dataset readings and family history. Unlike k-fold

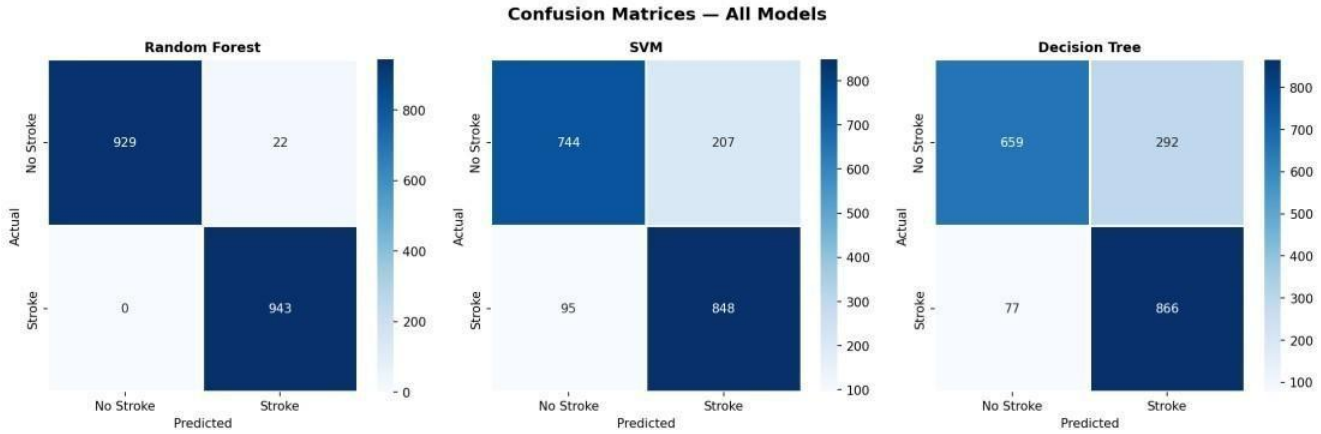


Fig. 7. Performance Metrics Comparison — All Models

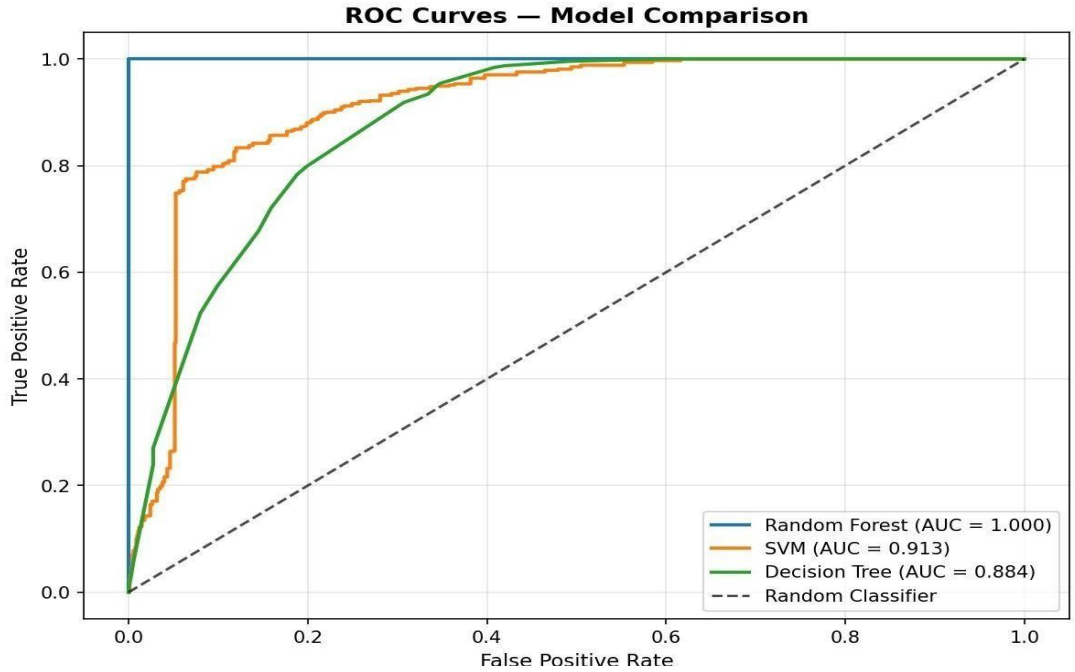


Fig. 8. Feature Importance — Random Forest Classifier

approaches, deep learning models were not considered and cross validation was not performed for the final evaluation. This might introduce a small amount of variation. Finally, longitudinal or time series data could capture the temporal dynamics of stroke risk which are not included in this study.

B. Future Scope:

Future work includes development of a real-time web or mobile application for stroke risk prediction for use by clinicians and patients, validation on larger and multi-institutional clinical datasets, incorporation of imaging data (MRI, CT) with CNN-based feature extraction for multimodal prediction and exploration of explainable AI (XAI) techniques (SHAP, LIME) to improve clinical interpretability. Other directions include multi-class prediction (ischemic, hemorrhagic and TIA), complex gradient boosting frameworks such as XGBoost, LightGBM and CatBoost,

use of federated learning to enable privacy-preserving multi-hospital training, use of NLP techniques to extract information from unstructured clinical notes and design of hybrid models that combine traditional clinical risk scores with ML predictions to achieve more comprehensive risk stratification..

VI. REFERENCES

[1] R. Ahmed, A. Varshney, Z. Ashraf, N. A. Farooqui, and R. S. Pathak, "Enhanced stroke risk prediction: A fusion of machine learning models for improved healthcare strategies," SN Computer Science, vol. 5, p. 1078, 2024.

[2] M. S. Singh, K. Thongam, P. Choudhary, and P. K. Bhagat, "Stroke risk prediction and prevention: Traditional versus machine learning approaches," Archives of Computational Methods in Engineering, 2025.

[3] R. Bhowmick, S. R. Mishra, S. Tiwary, and H. Mohapatra, "Machine learning for brain-stroke prediction: Comparative analysis and evaluation," Multimedia Tools and Applications, vol. 84, pp. 24025–24057, 2025.

[4] S. Xie, S. Peng, L. Zhao, B. Yang, Y. Qu, and X. Tang, "A comprehensive analysis of stroke risk factors and development of a predictive model using machine learning approaches," Molecular Genetics and Genomics, vol. 300, p. 18, 2025.

[5] Proc. 6th Int. Conf. Intelligent Communication Technologies and Virtual Mobile Networks (ICICV-2025), "Machine learning based approach for predicting brain stroke," IEEE, 2025.

[6] J. Amann, "Artificial intelligence in brain and mental health: Philosophical, ethical and policy issues," in Advances in Neuroethics, Springer, Cham, 2022.

[7] IEEE Conf. Paper, "Stroke risk prediction using machine learning models," Proc. Int. Conf., 2025.

[8] S. Gupta et al., "Brain stroke prediction using SVM and decision trees," Chitkara University Institute of Engineering and Technology, 2025.

[9] A. Messaoud et al., "Early prediction of cerebrovascular accidents using classification algorithms," National Engineering School of Gabes, University of Gabes, 2025.

[10] K. Sudheer Kumar et al., "Stroke risk prediction using machine learning," Vignan's Institute of Management Technology for Women, 2025.

[11] World Health Organization, "Stroke — Key facts," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[12] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[13] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.