WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

STROKE RISK PREDICTION USING MACHINE LEARNING: A COMPARATIVE STUDY AND PROTOTYPE SYSTEM

FRANCIS DANIEL VARNEY[1],MR. SRIKANT SINGH[2]

School of Engineering P P Savani University Surat, India
24IC03CS011

ABSTRACT—

Stroke is a leading cause of death and long-term disability, and early identification of high-risk people supports prevention. This paper compares four machine learning classifiers (Logistic Regression, Random Forest, Support Vector Machine and XGBoost) for stroke risk prediction on the public Kaggle Stroke Prediction Dataset (5,110 patients, 4.9% with stroke) and presents a prototype web application. The pipeline uses median imputation, one-hot encoding, standardization and SMOTE applied only inside cross-validation folds, with a held-out test set of 1,022 patients. All models reached a ROC-AUC between 0.81 and 0.84. Logistic Regression (0.844) and SVM (0.843) achieved the highest AUC and recall (0.80 and 0.82), while XGBoost achieved the highest precision and F1-score (0.174 and 0.255) but the lowest recall (0.48). SHAP analysis showed that age was by far the dominant predictor, followed by body mass index and hypertension. A Streamlit prototype that returns a colour-coded risk category, a probability, a recommendation and a PDF report was built and checked with 11 automated tests; the model inside the prototype is trained on simulated data to demonstrate the system pipeline. The results show moderate discrimination, so the models are suitable only as an awareness aid and not for diagnosis.

Keywords— *Stroke risk prediction, machine learning, class imbalance, SMOTE, SHAP, XGBoost, Streamlit.*

I. Introduction

Stroke is one of the leading causes of death and long-term disability worldwide, and many cases are linked to risk factors that can be identified and managed, such as hypertension, heart disease, high blood glucose and smoking [1]. Identifying people at high risk before a stroke occurs allows earlier lifestyle changes and medical follow-up. Conventional risk scores rely on a small number of variables and assumed linear relationships, and they may miss more complex interactions between risk factors.

Machine learning can combine many demographic, clinical and lifestyle variables and learn such interactions from data. Three practical difficulties remain. First, strokes are rare: in the dataset used here only 4.9% of patients had a stroke, so a model that always predicts “no stroke” is 95.1% accurate and accuracy alone is misleading. Second, clinicians need to understand why a model gives a particular risk. Third, a model is only useful if it can be turned into a tool that non-technical users can operate.

This paper addresses these points. The main contributions are: (1) a comparison of four classifiers using a protocol that avoids data leakage, in which oversampling is applied only inside the training folds and the final results come from an untouched test set; (2) an interpretation of the XGBoost model with SHAP; (3) a Streamlit prototype that returns a risk category, a probability, a recommendation and a downloadable report, together with automated unit and system tests; and (4) an explicit account of the limitations of the study. Section II reviews related work, Section III describes the method, Section IV presents the results, Section V discusses limitations and Section VI concludes.

II. Literature Study

Machine learning has been applied to stroke prediction using both large clinical cohorts and small public datasets. Chun *et al.* [4] used a prospective cohort of about 0.5 million Chinese adults to compare machine learning models with conventional statistical models. Vu *et al.* [2] evaluated five models, including logistic regression, random forest, support vector machine, XGBoost and LightGBM, on 7,389 participants of the Suita Study in Japan; the models reached AUC values of only 0.68 to 0.73, with random forest the highest, which shows that discrimination on real clinical data is often modest. Tang *et al.* [5] compared mainstream models on a public stroke dataset and used SHAP to explain the predictions, reflecting the growing demand for interpretable models in healthcare.

On the Kaggle stroke dataset used in this work, Dritsas and Trigka [3] reported a stacking ensemble with an AUC of 0.989 after balancing the classes with SMOTE. Very high figures of this kind are common for this dataset, and they should be read with care, because they can partly reflect evaluation choices: when oversampling is applied before the data are split, synthetic copies of the same patients can appear in both the training and the test set, which inflates the scores. The present study therefore applies SMOTE only inside the training folds, and it reports precision, recall and the precision-recall curve in addition to accuracy and AUC.

III. Methodology

A. Dataset used

The public Stroke Prediction Dataset from Kaggle [6] was used. It contains 5,110 patient records with 12 columns: an identifier, ten predictor variables (gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI and smoking status) and the binary target "stroke". In total 249 patients (4.9%) had a stroke. BMI was missing for 201 patients (3.9%); no other variable had missing values, and no duplicate records were found.

B. Preprocessing

The identifier was dropped and marital status was coded as 0/1. Missing BMI values were replaced with the median of the training data. Gender, work type, residence type and smoking status were one-hot encoded, and age, average glucose level and BMI were standardized. All transformations were fitted on the training data only, inside a pipeline, so that no information from the test set was used.

C. Train-test split and class balancing

The data were split with stratification into 80% for training (4,088 records, 199 strokes) and 20% for testing (1,022 records, 50 strokes), using a fixed random seed (42). To handle the class imbalance, SMOTE [9], [11] was applied inside each cross-validation fold of the training data and never touched the validation folds or the test set. On the complete training set, SMOTE raises the stroke class from 199 to 3,889 records, the size of the majority class.

D. Models and hyperparameter tuning

Four classifiers were compared: Logistic Regression, Random Forest, Support Vector Machine (RBF kernel) and XGBoost [8], implemented with scikit-learn [7] and the XGBoost library. Hyperparameters were tuned by randomized search with five-fold stratified cross-validation, using ROC-AUC as the score. The searched settings were C for Logistic Regression; number of trees, maximum depth and minimum samples per leaf for Random Forest; C and gamma for the SVM; and number of estimators, maximum depth, learning rate and subsample ratio for XGBoost. The selected values are listed in Table II.

E. Evaluation and interpretation

Each model was evaluated once on the held-out test set. Because strokes are rare, accuracy is reported but not used to rank the models. The main measures are recall (the share of strokes detected), precision, F1-score, ROC-AUC and the area under the precision-recall curve (PR-AUC), together with confusion matrices. The XGBoost model was interpreted with SHAP [10] using the TreeExplainer.

F. Prototype system

The prototype was developed with Streamlit. The user enters nine inputs (age, gender, hypertension, heart disease, average glucose level, BMI, smoking status, residence type and work type) or generates a synthetic patient. The inputs are encoded in the same way as in training, and the model returns the stroke probability, which is converted into a risk category: Low (below 30%), Moderate (30% to below 70%) or High (70% or above). The application shows the category, the probability with a progress bar and a short recommendation, and it can export a one-page PDF report. For demonstration, the model inside the prototype is an XGBoost classifier trained on 2,000 simulated patient records generated by a documented labelling rule, so the prototype demonstrates the system and does not report clinical performance. Fig. 1 summarises the overall method.

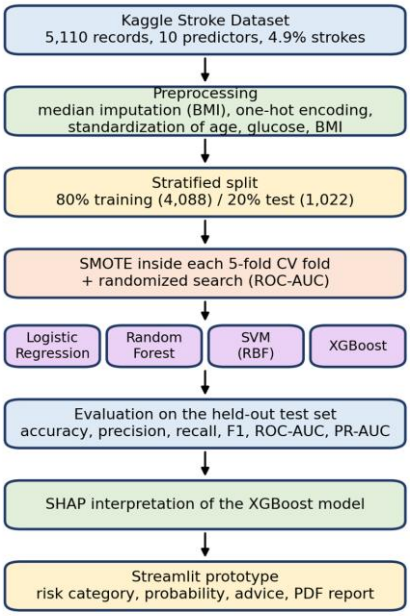


Fig. 1. Overall methodology of the study.

IV. Results and Analysis

A. Exploratory analysis

Stroke was strongly related to age (Fig. 2). The median age was 71 years in the stroke group and 43 years in the non-stroke group, and 72.7% of all

strokes occurred at age 60 or older; the stroke rate was 0.4% below age 40 and 13.2% from age 60 upward. Stroke rates were 13.3% with hypertension against 4.0% without, and 17.0% with heart disease against 4.2% without. Patients with missing BMI had a stroke rate of 19.9% against 4.9% overall, so the missingness was informative. BMI differed little between the groups (median 29.7 against 28.0).

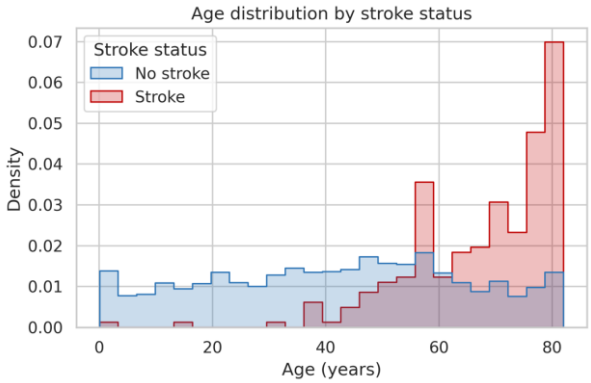


Fig. 2. Age distribution by stroke status.

B. Model comparison

Table I and Fig. 3 show the test-set results. The four models reached a ROC-AUC between 0.81 and 0.84. Logistic Regression had the highest AUC (0.844) with a recall of 0.80, and the SVM was almost identical (AUC 0.843, recall 0.82, PR-AUC 0.285). XGBoost had the highest accuracy (0.863), precision (0.174) and F1-score (0.255), but the lowest recall (0.48). Random Forest had an AUC of 0.809 and a recall of 0.70. Cross-validated AUC values on the training data (0.838, 0.837, 0.812 and 0.809 for Logistic Regression, SVM, Random Forest and XGBoost) gave the same ranking.

TABLE I

Test-set performance of the four models (n = 1,022; 50 strokes)

Model	Acc.	Prec.	Recall	F1	ROC-AUC	PR-AUC
Logistic Regression	0.749	0.139	0.80	0.237	0.844	0.266
Random Forest	0.750	0.127	0.70	0.215	0.809	0.205
SVM (RBF)	0.725	0.131	0.82	0.226	0.843	0.285
XGBoost	0.863	0.174	0.48	0.255	0.816	0.192

Accuracy is not informative here: always predicting “no stroke” would score 95.1%. Precision is low for all models (0.13 to 0.17) because the stroke rate is only about 5%, so most patients flagged as at risk will not have a stroke. The differences between the models are small compared with the uncertainty, since the test set holds only 50 stroke cases and each stroke detected changes recall by two percentage points. That a linear model matched the more complex ones suggests that risk in this dataset increases mostly in a smooth, additive way with age and the main clinical risk factors.

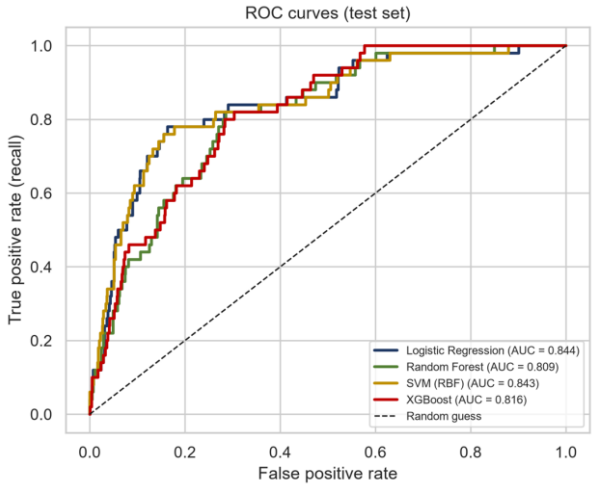


Fig. 3. ROC curves of the four models on the test set.

TABLE II

Selected hyperparameters

Model	Selected values
Logistic Regression	C = 0.01
Random Forest	trees = 400, max depth = 6, min samples per leaf = 1
SVM (RBF)	C = 0.5, gamma = 0.01
XGBoost	estimators = 200, max depth = 3, learning rate = 0.05, subsample = 1.0

C. Confusion matrices

Fig. 4 shows the confusion matrices. The SVM detected 41 of the 50 strokes and Logistic Regression 40, but they raised 272 and 247 false alarms. XGBoost raised only 114 false alarms and flagged 138 patients in total, at the cost of missing 26 of the 50 strokes. The models with the highest recall therefore flagged 28% to 31% of all patients as at risk. Which trade-off is preferable depends on how a screening programme weighs a missed stroke against an unnecessary check-up.

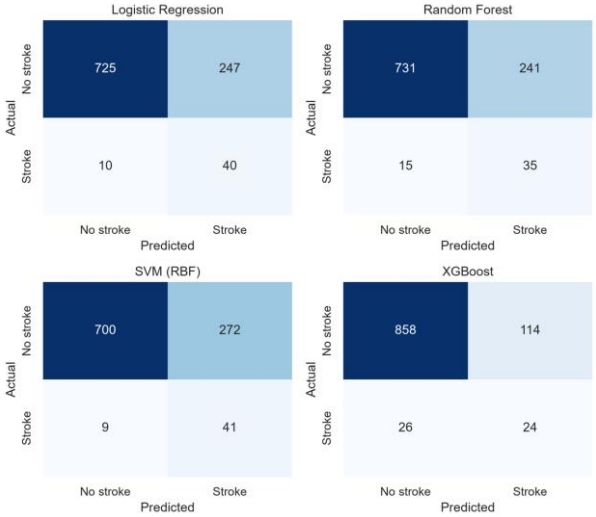


Fig. 4. Confusion matrices of the four models on the test set.

D. Feature importance

SHAP analysis of the XGBoost model (Fig. 5, Table III) showed that age was by far the most influential feature: its mean absolute SHAP value (2.18) was almost six times that of the next feature, BMI (0.37). High ages pushed the predicted risk up and low ages pushed it strongly down. Hypertension ranked sixth (0.165) and pushed predictions upward when present, while average glucose ranked eighth (0.123) and heart disease did not reach the top ten. Several features without a known causal link to stroke, such as gender, residence type and marital status, also carried weight. Marital status is closely tied to age (children are never married) and probably acts as a proxy, and the others probably reflect noise in a dataset with only 249 strokes. The model therefore agrees with clinical knowledge on age and hypertension, but it does not learn only genuine risk signals.

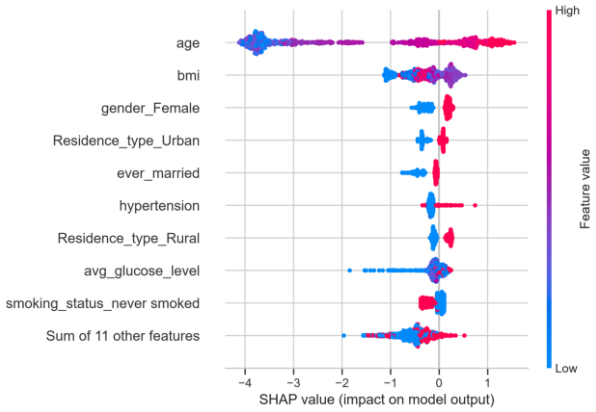


Fig. 5. SHAP summary plot of the XGBoost model on the test set.

TABLE III

Mean absolute SHAP values (top eight features)

Rank	Feature	Mean |SHAP|
1	Age	2.183
2	BMI	0.373
3	Gender (female)	0.229
4	Residence type (urban)	0.202
5	Marital status	0.196
6	Hypertension	0.165
7	Residence type (rural)	0.162
8	Average glucose level	0.123

E. Prototype and testing

Fig. 6 shows the prototype for a simulated patient aged 79 with hypertension and heart disease, which received a High Risk result of 91.7%. The prototype was checked with five unit tests (feature encoding, simulator ranges, threshold boundaries and report generation) and six system tests, in which complete patient profiles were passed through the whole pipeline and the resulting risk level was compared with the level expected from the simulator's labelling rule. All 11 tests passed.

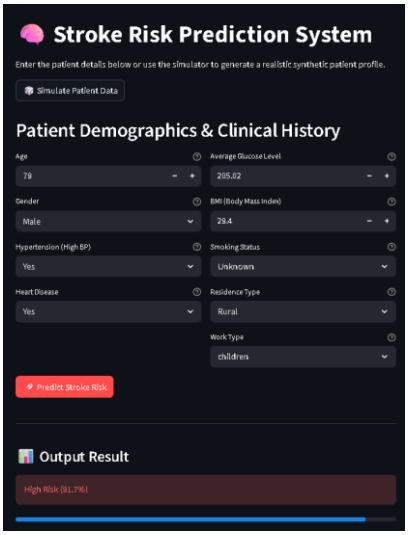


Fig. 6. The prototype application showing a High Risk result.

V. Discussion and Limitations

Four limitations should be kept in mind. First, the study uses one public dataset with only 249 strokes, and the test set contains 50, so every metric has a wide uncertainty; the results come from a single train-test split and were not validated externally. Second, SMOTE creates synthetic patients that do not exist in reality. Third, the dataset has no information on medication, family history, atrial fibrillation or measured blood pressure, which are important in clinical practice. Fourth, the model inside the prototype is trained on simulated data. Before any similar tool could be used with patients, it would need validation on independent clinical data, a check for bias across patient groups, calibrated probabilities and review by clinicians.

VI. Conclusion

This paper compared four machine learning models for stroke risk prediction on a public dataset using a protocol without data leakage, and it presented a prototype application. All models showed moderate discrimination (ROC-AUC 0.81 to 0.84): Logistic Regression and the SVM detected the most strokes, XGBoost produced the fewest false alarms, and none of them is accurate enough for diagnosis. Age was the dominant predictor, followed by BMI and hypertension. The prototype shows how such a model can be offered as an awareness tool with a risk category, a recommendation and a report. Future work includes validation on larger and independent clinical data, probability calibration, decision thresholds chosen for a target recall, SHAP-based personalised advice in the application, and a user study with clinicians.

REFERENCES

[1] American Heart Association/American Stroke Association, "2024 Guideline for the primary prevention of stroke: A guideline from the American Heart Association/American Stroke Association," *Stroke*, 2024, doi: 10.1161/STR.0000000000000475.

[2] T. Vu *et al.*, "Machine learning approaches for stroke risk prediction: Findings from the Suita Study," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 7, Art. no. 207, 2024, doi: 10.3390/jcdd11070207.

[3] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, Art. no. 4670, 2022, doi: 10.3390/s22134670.

[4] M. Chun *et al.*, "Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults," *J. Am. Med. Inform. Assoc.*, vol. 28, no. 8, pp. 1719–1727, 2021.

[5] X. Tang, M. Tang, W. Liu, and S. Cui, "Explainable machine learning for stroke risk prediction: A comparative study with SHAP-based interpretation," *Front. Neurol.*, vol. 16, Art. no. 1716984, doi: 10.3389/fneur.2025.1716984.

[6] fedesoriano, "Stroke Prediction Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[7] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[11] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.