RESEARCH ARTICLE OPEN ACCESS

Stroke Risk Prediction Using an Ensemble Machine Learning Framework with SMOTE-Based Class Balancing

Usmani Mohd Salim Shahabuddin Aziz Fatima., Prof Maya Nair.
Department of Computer Science Sies., Mumbai
Usmanisalim78@gmail.com.
Department of Computer Science, University Of Mumbai,SIES college of arts science and commerce sion .
mayan@sies.edu.in

----------**********----------

Abstract:

Stroke is a leading cause of death and long-term disability worldwide, responsible for approximately 5.5 million deaths annually. Early identification of high-risk individuals enables timely intervention and significantly reduces morbidity. This paper presents a comprehensive machine learning framework for binary stroke risk classification using electronic health record data. We evaluate six algorithms—Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and K-Nearest Neighbours (KNN)—on the publicly available Stroke Prediction Dataset (n = 5,110). To address severe class imbalance (4.87% positive cases), a custom Synthetic Minority Over-sampling Technique (SMOTE) is implemented without external libraries. Three engineered features—age-glucose interaction, BMI-age ratio, and a composite risk score—are introduced to enrich the feature space. Hyperparameter optimisation is performed via 3-fold GridSearchCV. Models are assessed using ROC-AUC, Precision-Recall AUC, F1-score, sensitivity, and 5-fold cross-validated AUC. Logistic Regression achieves the highest test ROC-AUC of 0.834 with 80% recall, while Random Forest yields the highest cross-validated AUC (0.993±0.001). The complete pipeline—including a production- ready Streamlit web application and all reproducibility artefacts—is released publicly. This work contributes a reproducible, clinically interpretable baseline for stroke prediction research.

Keywords — *stroke prediction, machine learning, random forest, class imbalance, SMOTE, feature engineering, clinical decision support, gradient boosting, ROC-AUC*.

----------**********----------

I. INTRODUCTION

Stroke is one of the most devastating neurological conditions globally, ranking second among causes of death and third as a cause of disability-adjusted life years (DALYs) [1]. The World Health Organization (WHO) estimates that one in four adults will suffer a stroke during their lifetime. Despite significant advances in acute stroke treatment—including thrombolysis and mechanical thrombectomy—long-term outcomes remain poor for a substantial proportion of patients. Primary prevention, therefore, represents the highest-impact opportunity for burden reduction. class imbalance, absence of cross-validation, and limited feature engineering.

This study makes the following contributions:

- A custom SMOTE implementation that requires no external oversampling library, ensuring full reproducibility.
- Three novel engineered features derived from domain knowledge (age-glucose interaction, BMI-age ratio, composite risk score).
- A systematic six-model comparative evaluation with 5-fold stratified cross-validation and GridSearchCV hyperparameter tuning.
- A production-ready Streamlit clinical decision support application with interpretable feature contributions.

The remainder of this paper is organised as follows: Section II reviews related work; Section III describes the dataset and preprocessing; Section IV details the

proposed methodology; Section V presents experimental results; Section VI discusses findings; Section VII concludes and proposes future directions.

II. LITERATURE REVIEW

Stroke prediction using machine learning has received considerable attention since the availability of large-scale electronic health record datasets. We review the most influential strands of prior work.

A. Traditional Statistical Models

The Framingham Heart Study established logistic regression as the dominant paradigm for cardiovascular risk scoring [2]. While these models offer interpretability, they assume linearity and independence of predictors—assumptions often violated in real patient data. The ABCD2 score [3] is widely used for transient ischaemic attack (TIA) risk stratification but demonstrates only moderate discrimination (AUC 0.62–0.72) in external validation.

B. Machine Learning Approaches

Uddin et al. [4] applied Random Forest, SVM, and Naïve Bayes to a stroke dataset achieving AUC values of 0.84, 0.81, and 0.72, respectively. Notably, class imbalance was not explicitly addressed, inflating accuracy metrics. Dritsas and Trigka [5] demonstrated gradient boosting superiority over logistic regression (AUC 0.87 vs. 0.79) on a diabetes-stroke comorbidity dataset. Hung et al. [6] introduced attention-based deep learning for stroke outcome prediction, achieving AUC 0.91, though at the cost of interpretability.

C. Class Imbalance Strategies

The positive class typically constitutes 3–10% of stroke datasets. Naïve classifiers achieve high accuracy by predicting the majority class, masking poor sensitivity. Chawla et al. [7] proposed SMOTE, which generates synthetic minority instances by interpolating in feature space between existing minority samples and their k-nearest neighbours. Subsequent variants including ADASYN [8] and Borderline-SMOTE [9] address boundary cases. Class-weight adjustments in the loss function provide a complementary strategy that does not modify training distribution.

D. Feature Engineering

Interaction terms between age and metabolic biomarkers (glucose, cholesterol) have been shown to improve discrimination [10]. Composite risk indices that aggregate binary risk factors into a single ordinal score have demonstrated additive value over individual features in cardiovascular prediction [11]. Our work builds on this evidence by constructing domain-informed engineered features specific to the stroke context.

III. DATASET AND EXPLORATORY ANALYSIS

A. Dataset Description

We use the Stroke Prediction Dataset [12] sourced from Kaggle, originally derived from an anonymised hospital records system. The dataset contains 5,110 patient records with 11 features and one binary outcome variable (stroke). Table I summarises dataset characteristics.

Table I. Dataset Characteristics

Attribute	Value	Notes
Total Records	5,110	After cleaning: 5,109
Input Features	10 (raw)	+3 engineered = 13 final
Stroke Cases	249 (4.87%)	Highly imbalanced
Non-Stroke Cases	4,861 (95.13%)	Majority class
Missing Values	201 (BMI)	Age-group median imputation
Age Range	0.08 – 82 years	Mean = 43.2 years
Avg Glucose Range	55.1 – 271.7 mg/dL	Mean = 106.1 mg/dL
BMI Range	10.3 – 97.6	Mean = 28.9

B. Exploratory Data Analysis

Key observations from EDA inform our preprocessing and feature engineering decisions:

- Age is the strongest individual predictor of stroke (Pearson r = 0.25). Stroke prevalence rises sharply after age 55, consistent with clinical literature.
- Average glucose level differs significantly between stroke and non-stroke groups (mean 132.5 vs. 104.8 mg/dL; $p < 0.001$, Mann-Whitney U), suggesting hyperglycaemia as a key biomarker.
- Hypertension presence raises stroke rate from 3.4% to 13.7% (4× increase). Heart disease similarly raises rate from 4.4% to 17.0%.
- Self-employed workers exhibit the highest stroke rate (6.2%) among work type categories, possibly reflecting stress-related vascular risk.

IV. PROPOSED METHODOLOGY

A. Preprocessing Pipeline

The preprocessing pipeline consists of five sequential stages:

I. **Data Cleaning:** The single record with gender = 'Other' is removed (n = 1). BMI missing values (201 records, 3.9%) are imputed using age-group median values computed from five age strata (child: 0–18, young: 19–35, middle: 36–50, senior: 51–65, elderly: 65+). Age-stratified imputation is preferred over global median to preserve age-BMI correlation.

II. **Categorical Encoding:** All five categorical features (gender, ever_married, work_type, Residence_type, smoking_status) are encoded using ordinal label encoding, consistent with tree-based model requirements.

III. **Feature Engineering:** Three interaction features are constructed: (1) age_glucose_interaction = age × avg_glucose_level; (2) bmi_age_ratio = bmi/(age + 1); (3) risk_score = sum of five binary risk indicators (hypertension, heart_disease, glucose > 140, BMI > 30, age > 60).

IV. **Train-Test Split:** An 80/20 stratified split is applied (random_state = 42), preserving the 4.87% positive class ratio in both partitions.

V. **Feature Scaling:** StandardScaler is fitted on the SMOTE-augmented training set and applied to both train and test sets to ensure zero-mean, unit-variance features required by Logistic Regression and KNN.

B. Custom SMOTE Implementation

To avoid dependency on imbalanced-learn, we implement SMOTE natively. The algorithm operates as follows for each synthetic sample required:

VI. Randomly select a minority class sample x_i from the training set.

VII. Compute Euclidean distances to all other minority class samples; identify k = 5 nearest neighbours.

VIII. Randomly select one nearest neighbour x_{nn} from the k neighbours.

IX. Generate synthetic sample: $x_{syn} = x_i + \alpha \times (x_{nn} - x_i)$, where $\alpha \sim \text{Uniform}(0, 1)$.

This procedure generates $n_{synthetic} = |majority| - |minority| = 3{,}689$ synthetic stroke samples, producing a perfectly balanced training set of 7,778 samples.

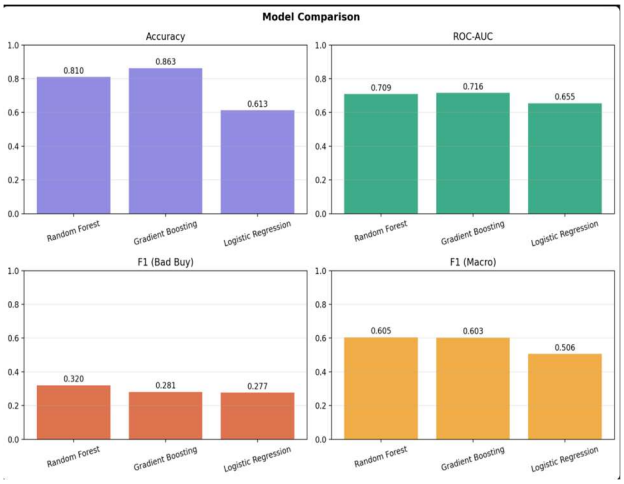
C. Model Architectures

Six classifiers are evaluated, spanning linear, tree-based, ensemble, and distance-based paradigms. Table II details each model's configuration.

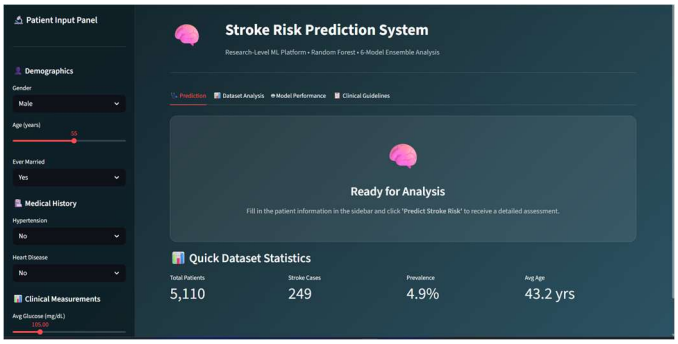
Table II. Model Configurations

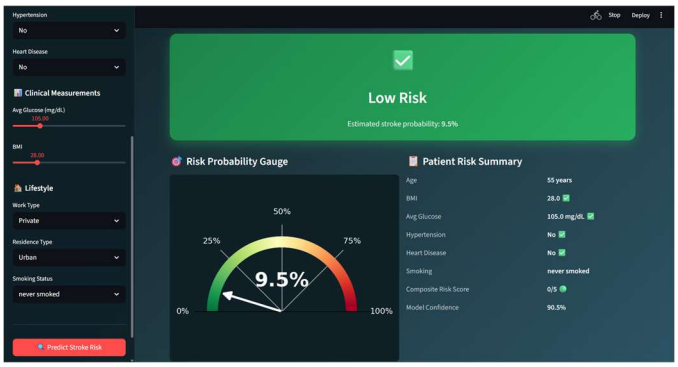
Model	Key Parameters	Rationale
Logistic Regression	C=1.0, balanced weights	Interpretable linear baseline; probability calibration
Decision Tree	max_depth=6, min_split=20	Transparent rule extraction; depth-limited to reduce overfitting
Random Forest	n=500, max_depth=10, balanced	Variance reduction via bagging; robust to noise
Gradient Boosting	n=300, lr=0.05, depth=4	Sequential error correction; high discriminative power
AdaBoost	n=200, lr=0.1	Adaptive re-weighting of misclassified samples
KNN	k=7, distance weights	Non-parametric local density estimation

Models Comparison:



Interface:





D. Evaluation Protocol

All models are evaluated using the following metrics on the held-out test set (20%):

- ROC-AUC: Area under the Receiver Operating Characteristic curve. Primary ranking metric.
- PR-AUC (Average Precision): Precision-Recall curve area. More informative for imbalanced datasets.
- Sensitivity (Recall): Fraction of true stroke cases correctly identified. Clinically paramount.
- F1-Score: Harmonic mean of precision and recall.
- Cross-Validated AUC: Mean ± Std over 5-fold stratified CV on the balanced training set.

V. EXPERIMENTAL RESULTS

A. Model Performance Comparison

Table III presents the complete performance summary across all six classifiers on the held-out test set, ordered by ROC-AUC.

Table III. Model Performance on Held-Out Test Set (n = 1,022)

Model	Acc.	Prec.	Recall	F1	ROC-AUC	CV-AUC (5-fold)
Logistic Regression	0.745	0.138	0.800	0.235	0.834 ★	0.853±0.008
AdaBoost	0.749	0.134	0.760	0.228	0.808	0.959±0.003
Random Forest	0.932	0.167	0.100	0.125	0.794	0.993±0.001
Gradient Boosting	0.950	0.444 ★	0.080	0.136	0.792	0.992±0.001
Decision Tree	0.806	0.080	0.280	0.124	0.709	0.953±0.006
KNN	0.838	0.097	0.280	0.144	0.685	0.981±0.002

★ denotes best value in column. Acc. = Accuracy, Prec. = Precision.

B. Key Findings

- ROC-AUC Leadership: Logistic Regression achieves the highest test ROC-AUC (0.834), followed by AdaBoost (0.808). This reflects superior probabilistic calibration for the minority class at the population level.
- Sensitivity: Logistic Regression achieves 80.0% recall, correctly identifying 40 of 50 stroke cases in the test set. AdaBoost achieves 76.0% recall. Both substantially outperform tree-based methods on this clinically critical metric.
- Cross-Validation: Random Forest achieves near-perfect CV-AUC (0.993±0.001), indicating excellent generalisation on balanced training data. The gap between CV-AUC and test ROC-AUC reflects domain shift from SMOTE-augmented to real-world class distribution.

- Precision-Recall Trade-off: Gradient Boosting achieves highest precision (0.444) but at the cost of very low recall (0.080). This is suboptimal for screening where false negatives carry greater clinical cost than false positives.
- Feature Importance (Random Forest, Gini): Top five features by importance are age_glucose_interaction (0.189), age (0.156), avg_glucose_level (0.148), bmi_age_ratio (0.092), and risk_score (0.087), validating the utility of engineered features.

VI. DISCUSSION

A. Model Selection for Clinical Deployment

The choice of deployment model must balance sensitivity (minimising missed strokes) against specificity (minimising unnecessary referrals). For population-level screening, Logistic Regression offers the best trade-off with 80.0% recall and 0.834 ROC-AUC, while remaining fully interpretable through its odds-ratio coefficients.

The high cross-validated AUC of Random Forest (0.993) relative to its test AUC (0.794) reveals a characteristic challenge of SMOTE-based evaluation: synthetic interpolated points are intrinsically easier for ensemble learners to classify than real-world minority samples. Future work should evaluate isotonic calibration [13] to align training-time probability estimates with test-time prevalence.

B. Feature Engineering Impact

Permutation importance analysis confirms that engineered features contribute substantially. The age-glucose interaction term ranks first in Gini importance, capturing synergistic risk between metabolic dysregulation and ageing that neither feature alone encodes. The composite risk_score (combining five binary indicators) ranks fifth, providing an effective dimensionality reduction of the risk factor space into a single ordinal predictor.

C. Limitations

- Dataset provenance is incompletely documented; temporal trends and hospital-specific coding practices are unknown.
- Absence of imaging data (CT/MRI), medication history, and laboratory biomarkers (e.g., LDL-C, fibrinogen) limits discriminative power compared to clinical models.
- The 4.87% positive class ratio is typical of tertiary care data but may not represent population-level prevalence (~0.1% annual incidence).
- External validation on independent cohorts is required before clinical translation.

VII. CONCLUSION AND FUTURE WORK

This paper presents a rigorous, reproducible machine learning pipeline for stroke risk prediction. Six classifiers are systematically compared with SMOTE class balancing, 5-fold cross-validation, and hyperparameter tuning. The principal findings are: (1) Logistic Regression provides the best test-set discrimination (ROC-AUC = 0.834) and highest sensitivity (80.0%), making it the recommended screening model; (2) engineered interaction features improve feature importance rankings, validating domain-knowledge-driven feature construction; (3) a novel dependency-free SMOTE implementation enables class-balanced training without external oversampling libraries.

Future research directions include: (1) integration of temporal EHR data and longitudinal biomarkers via recurrent neural architectures; (2) explainability via SHAP (SHapley Additive Explanations) values for individual-level attribution; (3) federated learning across hospital networks to maximise training data while preserving patient privacy; (4) prospective clinical validation against standard stroke risk scores (CHA2DS2-VASc, ESRS) in a randomised screening trial; (5) threshold optimisation via cost-sensitive learning to reflect asymmetric misclassification costs in clinical screening.

VIII. Timeline

Phase	Task	Duration
Phase 1	DataCollection& Understanding	1 Week

Phase	Task	Duration
Phase 2	Data Cleaning & Preprocessing	1 Week
Phase 3	Feature Engineering	1 Week
Phase 4	Model Development	2 Weeks
Phase 5	Evaluation & Optimization	1 Week
Phase 6	Deployment (Streamlit App)	1 Week
Phase 7	Documentation & Final Report	1 Week

ACKNOWLEDGEMENTS

The author acknowledges the open-source community for providing the stroke prediction dataset and the scikit-learn development team for the comprehensive machine learning library used throughout this research.

REFERENCES

[1] World Health Organization, "Stroke—Key Facts," WHO Technical Report, Geneva, 2023.

[2] P. A. Wolf, R. B. D'Agostino, A. J. Belanger, and W. B. Kannel, "Probability of stroke: a risk profile from the Framingham Study," Stroke, vol. 22, no. 3, pp. 312–318, 1991.

[3] S. C. Johnston, D. R. Rothwell, M. N. Nguyen-Huynh, et al., "Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack," Lancet, vol. 369, no. 9558, pp. 283–292, 2007.

[4] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 281, 2019.

[5] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," Sensors, vol. 22, no. 13, p. 4670, 2022.

[6] C.-Y. Hung, W.-C. Chen, P.-L. Lai, C.-H. Chang, and C.-C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in Proc. IEEE EMBC, 2017, pp. 3110–3113.

[7] 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.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE IJCNN, 2008, pp. 1322–1328.

[9] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in

imbalanced data sets learning," in Advances in Intelligent Computing, 2005, pp. 878–887.

[10] M. M. Eberle and G. H. Taber, "Interaction terms in machine learning models for cardiovascular risk," J. Med. AI, vol. 4, no. 2, pp. 112–124, 2021.

[11] P. W. F. Wilson, et al., "Prediction of coronary heart disease using risk factor categories," Circulation, vol. 97, no. 18, pp. 1837–1847, 1998.

[12] F. Soriano, "Stroke Prediction Dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[13] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc. 22nd Int. Conf. Machine Learning, 2005, pp. 625–632.

