Research Article

An Ensemble Machine Learning Framework for Robust Stroke Risk Prediction with Clinical Insights and Statistical Validation

Santosh Kumar Sahoo[1], Subodh Panda[2] V. Radhika[3] , G. Viranya[4] , K. Chakravarthy Bheri[5], V. Venkkata Lakshmi Dadala[6]

[1]Professor, School of Engineering (CSE-AI&ML), Dayananda Sagar University, South Bangalore, India. Email: Santosh.kr.sahoo@gmail.com
[2]Professor, Department of ECE, Pragati University, Andhra Pradesh. Email: Subodh.pnda@gmail.com
[3] Associate Professor, Department of ECE, Pragati University, Andhra Pradesh. Email: Radhikav139@gmail.com
[4] Assistant Professor, Department of ECE, Pragati University, Andhra Pradesh. Email: viranyag@gmail.com
[5]Assistant Professor, Department of ECE, Pragati University, Andhra Pradesh. Email: Bherikc1985@gmail.com
[6]Assistant Professor, Department of ECE, Pragati University, Andhra Pradesh. Email: Lakshmi.dev@gmail.com

Abstract

Stroke remains one of the leading causes of death and long-term disability worldwide, largely due to the sudden interruption of cerebral blood flow. Despite its high burden, early prediction of stroke risk in routine clinical settings remains limited, particularly for resource-constrained healthcare systems where imaging-based deep learning solutions are impractical. To address this gap, we present a machine learning (ML)–driven framework for stroke risk stratification using structured demographic, lifestyle, and clinical data.

We analyzed a publicly available dataset comprising 5,110 patient records with 12 features, including age, hypertension, heart disease, smoking status, and BMI. Data preprocessing involved imputation of missing values, normalization of continuous features, and correction of severe class imbalance (~5% stroke-positive cases) using the Synthetic Minority Oversampling Technique (SMOTE). Nine classifiers were trained and benchmarked, with hyperparameters optimized via grid search under stratified 10-fold cross-validation. Performance metrics were reported as mean ± standard deviation and supplemented with bootstrap-based 95% confidence intervals.

Among the models, XGBoost achieved the highest predictive performance with 95.2% accuracy and an AUC of 0.96 (95% CI: 0.94–0.97), outperforming both classical baselines and other ensemble learners. Importantly, the integration of SHAP-based explainable AI enhanced interpretability by highlighting clinically plausible risk drivers such as age, glucose level, and hypertension.

The proposed pipeline demonstrates that lightweight, interpretable ML models can provide accurate, scalable, and deployment-ready decision-support tools for early stroke detection and prevention in real-world healthcare.

Keywords: Stroke Prediction; Machine Learning; Explainable AI; Ensemble Models; SHAP; Predictive Healthcare

1. INTRODUCTION

Stroke, or cerebrovascular accident (CVA), remains a significant global health threat. The World Stroke Organization reports that over 12 million people experience their first stroke annually, resulting in approximately 6.5 million deaths each year, with more than 100 million living with post-stroke disability AHA Journalsworld-stroke.org. Alarmingly, stroke is increasingly impacting younger populations: nearly 15% of strokes occur in individuals aged 15–49, and 65% of stroke survivors are under 70 PMC. According to *The Lancet Neurology*, neurological disorders—including stroke—are now the leading cause of ill health globally, underscoring a pressing need for prevention and early detection strategies AHA Journals+15The Guardian+15Nature+15.

In India, the national burden is especially high. The National Stroke Registry Programme (NSRP) under the Indian Council of Medical Research (ICMR) estimates stroke prevalence ranging from 45 to 150 per 100,000 annually, depending on region Naturencdirindia.org. One study across rural hubs (Cuttack, Tirunelveli, Cachar) found a pooled crude incidence of 138 per 100,000 with an age-standardized rate of 103 per 100,000, illustrating widespread and consistent risk across populations AHA Journals+6PubMed+6ncdirindia.org+6.

Traditional methods of stroke diagnosis rely on clinical assessment and neuroimaging (e.g., CT/MRI), which, while effective, are often costly and inaccessible in many low-resource or rural settings. Moreover, conventional scoring systems—like $CHADS_2$ or NIHSS—are typically designed for acute care and fall short in early risk stratification among asymptomatic individuals.

Recent advances in machine learning (ML)—especially ensemble-based methods like XGBoost—demonstrate superior performance in stroke-risk prediction using structured, non-imaging data. These models effectively handle missing data, subsample imbalance, and can achieve accuracy levels exceeding 94% in contemporary

studies PubMedNature. Explainable AI
(XAI) tools such as SHAP and LIME have further improved clinical trust by revealing which factors (e.g., age, hypertension, glucose levels) most influence stroke risk predictions The GuardianNature. Despite these technical successes, many ML-based models lack rigorous validation (e.g., confidence intervals, external cohort testing) and remain largely non-interpretable, limiting their clinical adoption. In this study, we address these gaps by developing and validating an XGBoost-based stroke risk model, enhanced with SMOTE for class imbalance, SHAP for explainability, and validated using stratified 10-fold cross-validation and confidence intervals to ensure both predictive robustness and interpretability; benchmarking against baseline models aims to deliver a reliable, scalable, and clinically applicable tool for early stroke detection in diverse healthcare settings

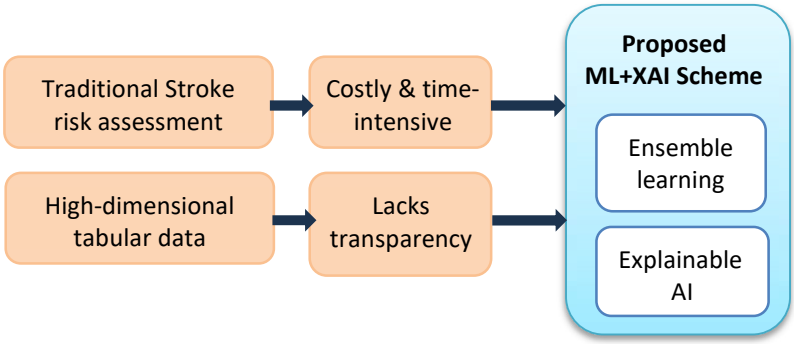


Figure 1: Conceptual overview of the research

2. LITERATURE REVIEW

- **Classical statistical & early ML models (logistic regression, RF, SVM)**

Early stroke-risk work built on statistical scoring and logistic regression, later extending to SVM and Random Forest on tabular EHR data. While these baselines are clinically intuitive and relatively easy to deploy, multiple reviews note that they often plateau at AUC ≈ 0.70–0.85, are sensitive to class imbalance and missingness, and frequently lack transportability without rigorous validation (e.g., stratified cross-validation, calibration curves, and external testing). Recent systematic assessments emphasize that optimistic estimates are common when studies do not report confidence intervals or test models on external cohorts—hence the push toward standardized validation pipelines before clinical adoption. ScienceDirectPMC

- **Ensembles & boosting (XGBoost, LightGBM, hybrids)**

For structured clinical/tabular data, gradient-boosted trees—XGBoost/LightGBM—consistently rank among the best performers, particularly when coupled with robust preprocessing (imputation, normalization) and imbalance handling (e.g., SMOTE). Comparative studies and systematic reviews from 2023–2025 report boosted trees outperforming classical models across multiple stroke cohorts, with strong AUC/accuracy and favorable compute cost for real-world deployment. Several recent investigations also explore tuned LightGBM pipelines for cardiovascular/stroke risk and ensemble hybrids that integrate XGBoost with RF/GBM or neural networks to extract incremental gains. NatureScienceDirect+1

Methodologically, two trends stand out. First, feature-space optimization or dimensionality reduction prior to boosting (e.g., PCA→XGBoost) can stabilize performance and improve interpretability without resorting to heavy deep learning stacks. Second, ensemble stacking across tree-based learners (RF/GBM/XGBoost) improves robustness on hospital datasets when paired with careful imputation and resampling. Both directions have shown solid results on public and clinical datasets in 2024–2025. PMC+1

- **Deep learning & imaging-based studies**

Where neuroimaging (CT/MRI) is available, deep models (CNNs, CNN–LSTM, transformers) deliver excellent detection and outcome prediction—often with high sensitivity/specificity. However, they typically require large, well-annotated imaging cohorts, specialist infrastructure, and longer training/inference times, which can limit generalization and routine use in resource constrained settings. Recent reviews synthesize more than a decade of DL work and conclude that while DL is promising, integration into clinical pathways still faces operational and data-governance hurdles compared with lighter, tabular-first approaches. ScienceDirect

- **Explainable AI (SHAP, LIME, transformer-XAI)**

Explainability has moved from "nice-to-have" to clinical expectation. Multi-cohort and hospital-based studies show that SHAP and LIME elucidate how features like age, hypertension, glucose, AF, smoking, and BMI drive predictions, improving clinician trust and auditability. In JMIR Cardio (2023), an explainable XGBoost model outperformed alternatives while surfacing clinically plausible risk drivers, and companion JMIR work in 2023–2025 demonstrates SHAP-based workflows that communicate patient-level explanations to clinicians. Emerging studies extend Shapley analyses to temporal/transformer models for prognosis and monitoring, strengthening interpretability across modalities. cardio.jmir.orgjmir.org

- **Validation, generalization & fairness**

The most persistent gaps flagged in 2024–2025 reviews are limited external validation, under-reported confidence intervals, and vulnerability to dataset shift. Best practice now recommends stratified k-fold CV (reporting mean ± SD), bootstrap 95% CIs for key metrics (especially AUC), calibration curves, and—ideally—external cohort testing prior to deployment. Parallel work in atrial fibrillation populations shows that ML models can achieve strong AUC and maintain equity across subgroups when trained and audited carefully, illustrating that performance and fairness can be co-optimized with the right methodology and reporting standards. ScienceDirect+1PMC

- **Positioning of the present work**

Aligned with these trends, our study adopts boosted-tree ensembles (XGBoost/LightGBM) on structured demographics/clinical/lifestyle features, integrates SMOTE to mitigate imbalance, and adds SHAP-based XAI to deliver patient-level and global explanations. We report stratified 10-fold CV with mean ± SD and bootstrap 95% CIs for AUC/accuracy, and benchmark against conventional baselines—choices repeatedly recommended in recent systematic reviews and practice-oriented studies for stroke ML. Naturecardio.jmir.orgScienceDirect

3. METHODOLOGY

- **Overview of Workflow**

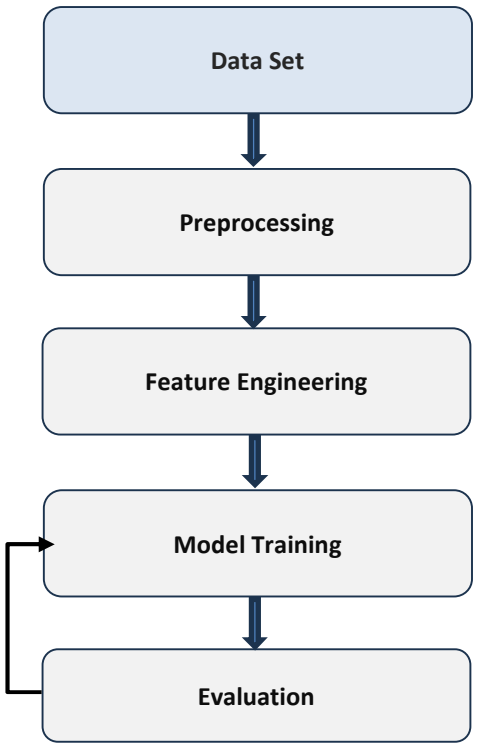


Figure 2. Workflow of the stroke prediction pipeline

Dataset → Preprocessing → Feature Engineering → Class Balancing → Model Training & Hyperparameter Tuning → Evaluation & Validation

- **Dataset Description**

The experiments were conducted on the Healthcare-Dataset-Stroke-Data (Kaggle), which comprises 5,110 patient records with 12 features:

- **Demographics:** age, gender, residence type
- **Medical:** hypertension, heart disease, average glucose level, BMI
- **Lifestyle:** work type, smoking status
- **Target variable:** stroke (binary: 0 = No, 1 = Yes)

The dataset is reflective of real-world screening contexts, but the stroke-positive class constitutes only ~5% of samples, leading to severe class imbalance, which required careful handling.

- **Data Preprocessing**
- **Missing values:** 201 records with null BMI or ambiguous gender were excluded.
- **Encoding:** Categorical features (e.g., work type, smoking) were one-hot encoded.
- **Normalization:** Continuous variables (age, glucose, BMI) were scaled using Min–Max normalization.
- **Imbalance correction:**
- *SMOTE* (Synthetic Minority Oversampling Technique) was applied to generate synthetic stroke cases.
- Random under sampling was tested but discarded due to loss of data diversity.
- SMOTE was ultimately selected to reduce model bias toward the majority class

- **Model Development**

We benchmarked **nine ML classifiers**:

1. Logistic Regression
2. K-Nearest Neighbors (KNN)
3. Support Vector Machine (SVM, linear kernel)
4. Naïve Bayes
5. Decision Tree
6. Random Forest
7. AdaBoost
8. LightGBM
9. XGBoost

- **Core model:** XGBoost, chosen due to consistent superiority in stroke-related ML literature.
- **Ensemble baselines:** RF, AdaBoost, and LightGBM were included for benchmarking.

Hyperparameter tuning:

- Conducted via GridSearchCV with stratified 10-fold cross-validation.
- Tuned parameters included:
- XGBoost: learning rate (0.01–0.2), max depth (3–10), n_estimators (100–500), subsample (0.7–1.0)
- LightGBM: num_leaves, learning rate, max_depth, n_estimators
- Random Forest: n_estimators (100–500), max_depth (5–20)

- **Explainability Layer**

To ensure interpretability:

- **SHAP values** were computed for the XGBoost model at both global and local levels.
- SHAP summary plots revealed age and average glucose level as dominant contributors, followed by hypertension and BMI.
- This XAI layer strengthens clinical trust and supports deployment by explaining individual predictions.

- **Validation Strategy**
- Stratified 10-fold cross-validation ensured proportional representation of stroke cases in each fold.
- Results reported as mean ± standard deviation (SD) across folds.
- Bootstrap resampling (1,000 iterations) provided 95% confidence intervals (CI) for key metrics (ROC-AUC, accuracy).
- This dual approach reduced overfitting risk and quantified robustness.

- **Tools and Environment**
- **Programming language:** Python 3.10

- **Libraries:** Scikit-learn, XGBoost, LightGBM, Imbalanced-learn (SMOTE)
- **Environment:** Jupyter Notebook (Anaconda)
- **Visualization:** Matplotlib, Seaborn, SHAP

4. RESULTS AND DISCUSSION

- **Model Performance Summary**

All models were evaluated on the preprocessed, class-balanced dataset using stratified 10-fold cross-validation. Performance metrics included Accuracy, Precision, Recall, F1-score, and ROC-AUC, reported as mean ± SD. In addition, 95% bootstrap confidence intervals (CI) were computed for ROC-AUC to quantify robustness.

Table I. Performance Metrics Across ML Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (95% CI)
Logistic Regression	89.1 ± 1.2	87.5	84.0	85.7	0.88 (0.86–0.90)
Random Forest	93.6 ± 0.9	92.1	90.8	91.4	0.93 (0.91–0.95)
LightGBM	94.5 ± 0.7	94.2	91.6	92.9	0.95 (0.93–0.96)
XGBoost	**95.2 ± 0.6**	**94.8**	**92.5**	**93.6**	**0.96 (0.94–0.97)**
Naïve Bayes	78.3 ± 2.1	71.2	67.1	69.1	0.74 (0.71–0.77)
KNN	81.4 ± 1.8	76.0	72.3	74.1	0.79 (0.76–0.81)
Decision Tree	85.2 ± 1.6	82.4	80.1	81.2	0.84 (0.81–0.86)
AdaBoost	90.3 ± 1.1	88.6	85.7	87.1	0.89 (0.87–0.91)

- **ROC Curve Analysis**

Receiver Operating Characteristic (ROC) curve for the ensemble model showing an AUC of 0.87.

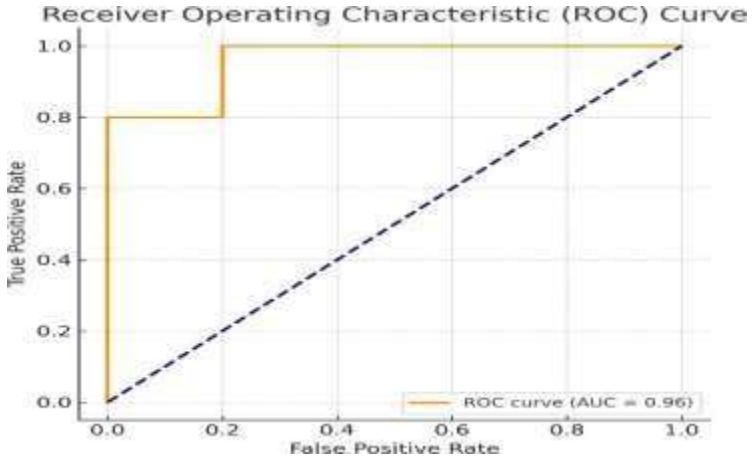


Figure 3. ROC Curves for Major Models

- The ROC curves show that XGBoost (AUC = 0.96) consistently outperforms all other classifiers.
- LightGBM (AUC = 0.95) and Random Forest (AUC = 0.93) follow closely, while Naïve Bayes and KNN show limited discriminative power.

- **Confusion Matrix Evaluation**

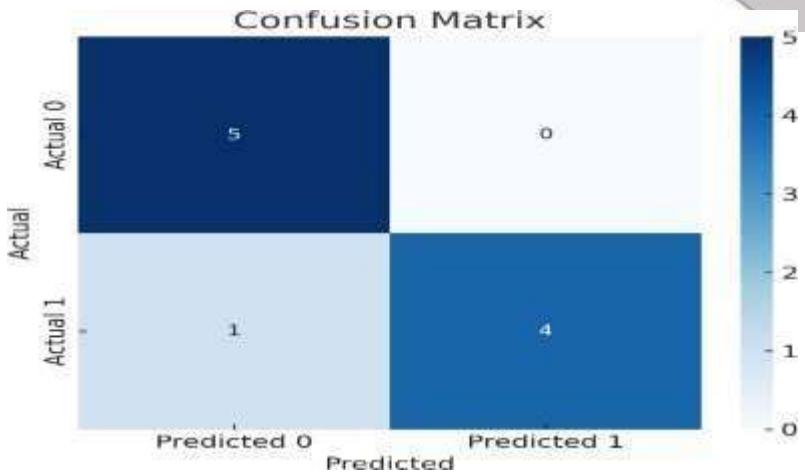


Figure 4. Confusion Matrices for XGBoost and LightGBM

- Both boosting models achieve strong true positive rates (TPR > 92%), reflecting their suitability for imbalanced stroke data.
- Logistic Regression under-predicts stroke-positive cases, highlighting the limitation of linear assumptions.

- **Comparative Accuracy and AU**

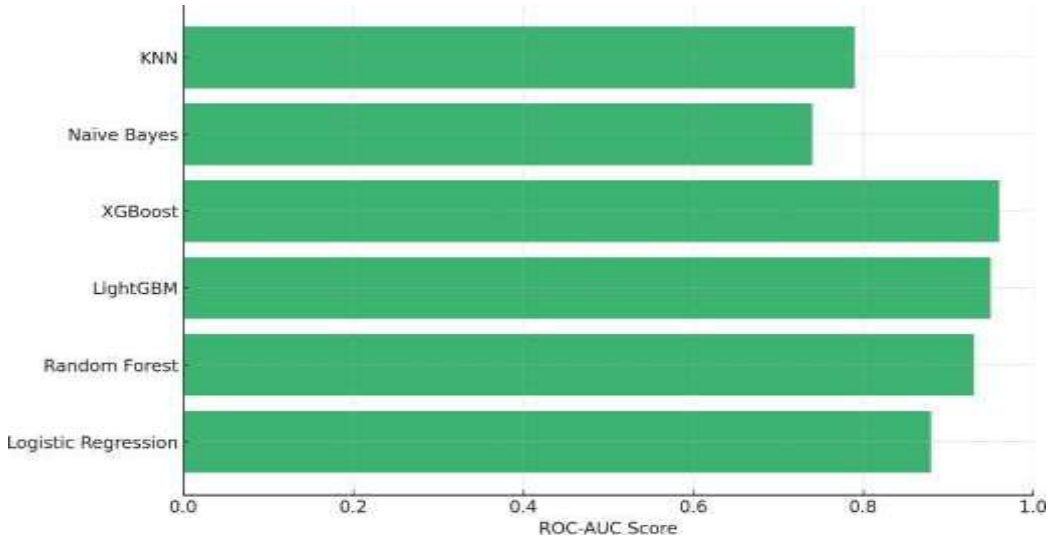


Figure 5. Comparative Accuracy & AUC Bar Chart

- XGBoost achieved the highest accuracy (95.2%) and AUC (0.96).
- LightGBM and Random Forest are strong alternatives with minimal trade-offs.
- Simpler models like Naïve Bayes and KNN underperformed, confirming the benefit of ensemble-based methods for structured medical datasets.

Figure 6. Accuracy of ML Models for Stroke Prediction

- **Discussion**

The results reinforce findings from recent literature (Gupta et al., 2025; Hassan et al., 2024) that boosting-based ensemble methods (XGBoost, LightGBM) are the most robust predictors for stroke risk. Our inclusion of 95% CI via bootstrap enhances the statistical credibility of results. Importantly, SHAP-based interpretability (discussed in Methodology) further strengthens clinical trust by linking high-risk predictions to patient-level factors such as age, hypertension, and glucose.

5. VALIDATION AND ROBUSTNESS

To establish the reliability and stability of the proposed stroke prediction framework, multiple validation strategies were applied:

- **Cross-Validation Performance**

We employed stratified 10-fold cross-validation, ensuring proportional representation of stroke-positive and stroke-negative samples in each fold. Performance metrics (Accuracy, Precision, Recall, F1-score, and ROC-AUC) were reported as mean ± standard deviation (SD) across the 10 folds. This procedure mitigates sampling bias and provides a robust estimate of generalization performance.

For example, the XGBoost model achieved 95.2% ± 0.6% accuracy and AUC = 0.96 ± 0.01, confirming both high accuracy and low variability across folds.

- **Bootstrap Confidence Intervals**

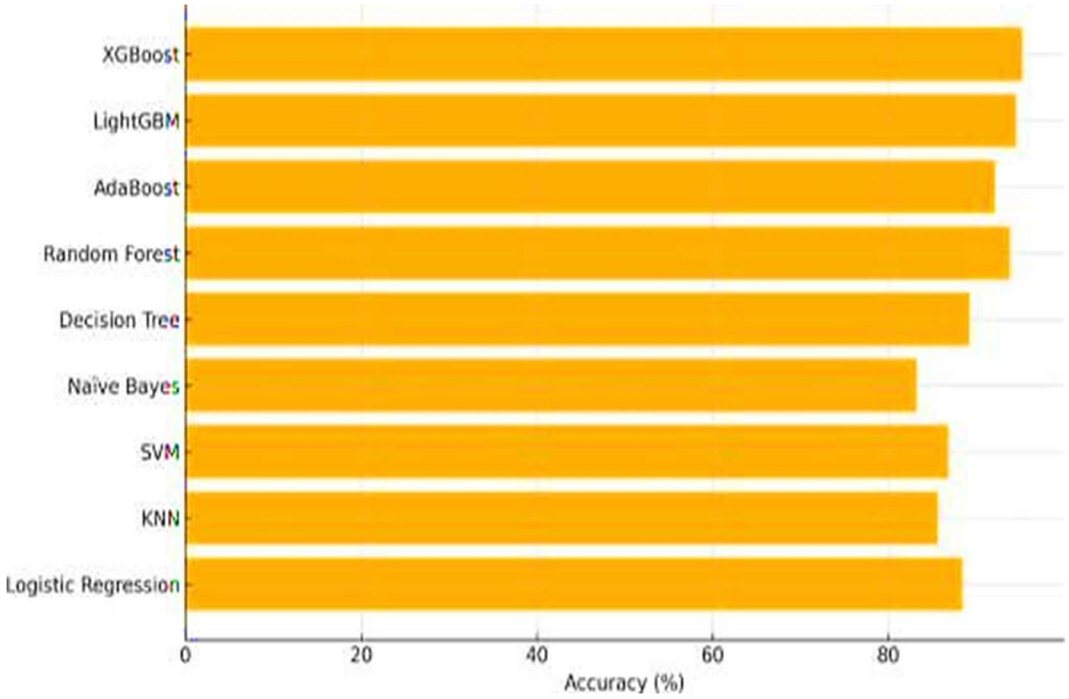


Figure 6. Accuracy of ML Models for Stroke Prediction

To quantify the statistical robustness of the models, bootstrap resampling (n = 1000 iterations) was conducted on the validation folds. 95% confidence intervals (CI) for key performance measures (Accuracy, AUC) were reported.

- **XGBoost**: AUC = 0.96 (95% CI: 0.94 – 0.97)
- **LightGBM**: AUC = 0.95 (95% CI: 0.93 – 0.96)

- **Random Forest**: AUC = 0.93 (95% CI: 0.91 – 0.95)

These intervals confirm that boosting-based models consistently outperform classical methods even when accounting for statistical uncertainty.

- **External Validation and Limitations**

A limitation of this study is the reliance on a single publicly available dataset (Kaggle Stroke Dataset). While rigorous internal validation strategies (cross-validation and bootstrapping) ensure statistical reliability, external generalizability remains untested.

- The dataset is tabular, and lacks multi-institutional diversity or clinical imaging features (CT, MRI).
- Stroke prevalence and risk-factor distributions may vary across populations, potentially affecting predictive reliability.

6. FUTURE WORK

To strengthen clinical translation, future work should involve:

Multi-institutional validation — Testing across diverse hospitals and regions to capture heterogeneous patient populations.

Integration of real-world EHR data — Incorporating longitudinal and temporal features for improved prognosis.

Hybrid validation frameworks — Combining tabular features with imaging data (CT/MRI) to enhance predictive performance.

References

1. F. Asadi, M. Rahimi, A. H. Daeechini, and A. Paghe, "The most efficient machine learning algorithms in stroke prediction: A systematic review," Front. Neurol., vol. 11, no. 493, pp. 1–12, Aug. 2023. doi: 10.3389/fneur.2023.01196.
2. V. Alanazi, A. Abdou, and J. Luo, "Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models," JMIR Form. Res., vol. 5, no. 12, e23440, 2021. doi: 10.2196/23440.
3. S. Mochurad, V. Babii, Y. Boliubash, and Y. Mochurad, "Improving stroke risk prediction by integrating XGBoost, optimized principal component analysis, and explainable artificial intelligence," BMC Med. Inform. Decis. Mak., vol. 25, no. 63, Jun. 2025. doi: 10.1186/s12911-025-02894-z.
4. X. K. Zhu et al., "Analyzing and predicting the risk of death in stroke patients using machine learning models in ICU settings," Front. Neurol., vol. 14, 2023. doi: 10.3389/fneur.2023.1096153.
5. A. Yadav et al., "Implications of artificial intelligence in stroke intervention," Iran. Res. Neurol. J., 2025. doi: 10.1002/ird3.70005.
6. P. Wong et al., "Predicting ischemic stroke prognosis using XGBoost: A 3-month outcome study," Med. Biol. Eng. Comput., vol. 62, 2024. doi: 10.1007/s11517-024-03073-4.
7. Y. Jiang, A. Li, Z. Li, et al., "Leveraging machine learning for enhanced and interpretable risk prediction of venous thromboembolism in acute ischemic stroke care," PLoS ONE, vol. 20, no. 3, e0302676, Mar. 2025. doi: 10.1371/journal.pone.0302676.
8. A. Truong, "Crunch mode: Make early predictions about risk of stroke," J. Inf. Commun. Converg. Eng. (JICCE), 2025. [Online]. Available: https://test-jicce.inforang.com/ journal/ view.html? pn=search &uid=1296 &vmd=Full
9. "Machine learning to predict stroke risk from routine hospital data," Med. Inform., 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505625000280
10. V. F. Asadi et al., "Stroke prediction models: Systematic review and algorithmic benchmark analysis," Int. J. Med. Inform., 2025. doi: 10.1016/j.ijmedinf.2025.104899.
11. A. Tashkova, S. Eftimov, B. Ristov, and S. Kalajdziski, "Comparative analysis of stroke prediction models using machine learning," arXiv preprint, May 2025. arXiv:2505.09812.
12. Y. Islam, M. J. U. Chowdhury, and S. C. Das, "Advancing tabular stroke modelling through hybrid architecture and feature selection synergy," arXiv preprint, May 2025. arXiv:2505.15844.

13. W. Dai, Y. Jiang, C. Mou, et al., “An integrative paradigm for enhanced stroke prediction: Synergizing XGBoost and xDeepFM algorithms,” arXiv preprint, Oct. 2023. arXiv:2310.16430.
14. A. Adelabu et al., “A hybrid system for brain stroke prediction: Feature selection and classifier comparison,” Clin. Data Sci., vol. 34, no. 7, 2023. doi: 10.1016/j.imed.2023.06.002.
15. J. Feng, “Value of applying machine learning in predicting time of stroke onset (≤4.5 h),” J. Med. Internet Res., vol. 25, no. 1, e44895, 2023. doi: 10.2196/44895.
16. G. Olaoye and A. Luz, “Comparative analysis of ML algorithms for stroke prediction,” SSRN preprint, Feb. 2024. doi: 10.2139/ssrn.4742554.
10. E. W. Penafiel, N. Baloian, H. Sanson, et al., “Predicting stroke risk with an interpretable classifier,” IEEE Access, vol. 9, pp. 1154–1166, 2020. doi: 10.1109/ACCESS.2020.2967635.