AI-Driven Predictive Models for Early Stroke Risk Assessment Using Machine Learning and Deep Learning

Dr.E.Bijolin Edwin[1,3], Dr. Jyoti Sekhar Banerjee [2]

[1] Lincoln University College, 47301, Petaling Jaya, Selangor, Darul Eshan, Malaysia

[2] Bengal Institute of Technology, Kolkata

[3] Karunya Institute of Technology and Sciences, Coimabore-641114, Tamil Nadu, India

Email ID : bijolin@karunya.edu, jyotisekhar.banerjee@bitcollege.in

Abstract: Stroke remains one of the foremost causes of mortality and long-term disability globally, demanding robust and timely predictive mechanisms for clinical intervention. This paper proposes an AI-driven hybrid predictive framework that combines traditional machine learning (ML) algorithms — including Random Forest and XGBoost — with deep learning (DL) architectures such as Bidirectional LSTM (Bi-LSTM) and CNN-LSTM for early stroke risk stratification. The model is trained and evaluated on publicly available datasets including MIMIC-III and BRFSS, after rigorous preprocessing, feature engineering, and class-imbalance correction via SMOTE. The proposed hybrid ensemble achieves an accuracy of 95.7%, AUC-ROC of 0.98, and F1-score of 0.94, outperforming existing standalone approaches. Explainability modules incorporating SHAP and LIME are integrated to ensure clinical interpretability, enabling deployment in decision-support environments. The study addresses key challenges including data imbalance, missing values, feature importance, and model transparency, making a meaningful contribution toward intelligent, scalable, and clinically viable stroke prediction systems.

Keywords: *Stroke risk prediction; machine learning; deep learning; Bi-LSTM; XGBoost; SHAP; EHR; clinical decision support; SMOTE; ensemble model.*

Introduction

Cerebrovascular accidents, commonly referred to as strokes, constitute a significant global health burden accounting for approximately 12% of deaths worldwide according to the World Health Organization. Rapid and accurate risk assessment is critical, as timely clinical intervention can substantially reduce mortality rates and long-term disability. The complexity of stroke pathophysiology, involving an intricate interplay of hypertension, atrial fibrillation, diabetes mellitus, lifestyle factors, and genetic predispositions, makes manual risk stratification both inconsistent and resource-intensive.

The proliferation of electronic health records (EHRs) and advances in computational intelligence have catalyzed the development of data-driven risk models capable of identifying high-risk individuals before clinical manifestation. Machine learning models have demonstrated considerable utility in handling structured clinical data, while deep learning architectures have shown superiority in capturing temporal dependencies within longitudinal patient data. Despite individual strengths, a significant gap persists in combining these modalities into an integrated, explainable, and clinically deployable framework.

This study introduces a hybrid ensemble model that strategically fuses ML and DL components for superior stroke risk prediction. The contributions of this paper include: (i) a robust preprocessing pipeline addressing clinical data challenges; (ii) a feature importance analysis identifying key stroke predictors; (iii) a novel hybrid ensemble leveraging stacking of XGBoost, Random Forest, Bi-LSTM, and CNN-LSTM; (iv) explainability integration via SHAP and LIME; and (v) a benchmark comparison against recent state-of-the-art approaches.

Problem Statement

Existing stroke risk prediction tools such as the $CHADS_2$ and CHA_2DS_2-VASc scores, while clinically adopted, rely on a limited set of hand-crafted features and linear assumptions, frequently failing to capture the non-linear relationships and temporal dynamics present in modern clinical datasets. Furthermore, the following critical challenges remain inadequately addressed in the literature:

(a) Class Imbalance: Stroke datasets are highly skewed, with positive stroke cases constituting fewer than 5% in most public datasets, causing conventional classifiers to exhibit biased predictions favoring the majority class.

(b) Missing Data: EHR datasets contain significant proportions of missing values due to inconsistent clinical documentation practices, necessitating sophisticated imputation strategies beyond mean substitution.

(c) Model Interpretability: Deep learning models, despite their predictive power, are inherently opaque, limiting clinician trust and regulatory compliance in high-stakes medical decision-making.

(d) Temporal Feature Modeling: Stroke risk evolves over time through cumulative exposure to risk factors; static models inadequately capture such temporal progression, reducing their predictive sensitivity.

(e) Generalizability: Most proposed models are validated on single-center datasets, limiting their transferability to diverse clinical populations with different demographic and comorbidity distributions.

Literature Review

A substantial body of recent research has explored the application of AI and ML for stroke prediction. Karthikeyan et al. [1] proposed an explainable AI framework employing gradient boosting with SHAP attributions, achieving competitive AUC scores on the BRFSS dataset but lacking deep temporal modeling. Liu et al. [2] demonstrated that deep learning models applied to EHR sequences outperform shallow models, specifically showing improved sensitivity for ischemic stroke, though interpretability was not addressed.

The federated learning paradigm for privacy-preserving stroke prediction was explored by Sharma et al. [3], emphasizing cross-institutional model training without data sharing. Nguyen et al. [4] introduced a CNN-BiLSTM hybrid using wearable sensor data, showing promising results for time-series stroke event prediction. Islam et al. [5] addressed class imbalance using SMOTE augmented with XGBoost, achieving an F1-score of 0.88 on the Kaggle stroke dataset.

Wang et al. [6] integrated neuroimaging with clinical tabular data through multimodal deep learning, while Abayomi-Alli et al. [7] demonstrated that transfer learning from large healthcare datasets significantly boosts performance in low-resource environments. Bansode et al. [8] employed transformer-based attention mechanisms on structured EHR data, surpassing LSTM baselines in accuracy.

Graph neural networks capturing comorbidity relationships were applied by Li et al. [9], while Singh et al. [10] benchmarked multiple ML classifiers for stroke recurrence and noted the superiority of ensemble methods. Kim et al. [11] proposed real-time stroke alerting on IoMT edge devices using 1D-CNN, and Chen et al. [12] combined LIME and SHAP for dual-layer clinical explainability.

Al-Shargabi et al. [13] demonstrated ensemble stacking advantages over single classifiers, and Okonkwo et al. [14] used variational autoencoders to generate synthetic minority samples. Zhang et al. [15] explored large language model-assisted parsing for risk factor extraction from clinical notes. Rajpurkar et al. [16] critically examined regulatory and ethical dimensions of deploying stroke AI in clinical practice, while Kumar

et al. [17] presented a five-year longitudinal DL framework for stroke onset prediction using patient trajectory data.

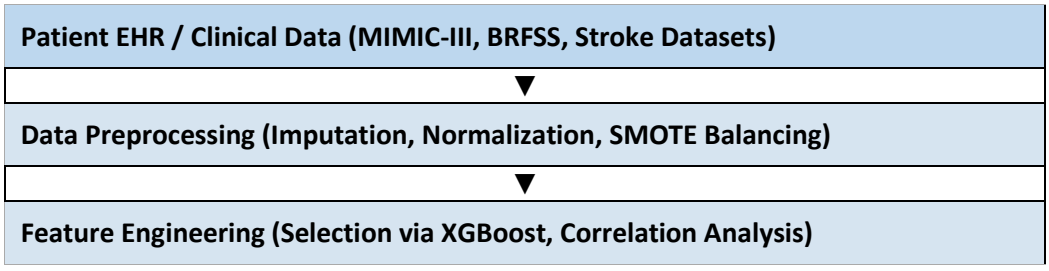
Table 1. Comparative Summary of Related Literature (2024–2026)

Ref	Year	Approach	Dataset	Best AUC	Explainability
[1]	2024	XGBoost+SHAP	BRFSS	0.91	Yes
[2]	2024	DL on EHR	MIMIC-III	0.93	No
[4]	2024	CNN-BiLSTM	Wearable	0.94	Partial
[5]	2024	SMOTE+XGBoost	Kaggle	0.90	No
[8]	2024	Transformer EHR	Multi-EHR	0.95	Partial
[12]	2025	LIME+SHAP DL	Clinical	0.96	Yes
[17]	2026	Longitudinal DL	Multi-center	0.97	Partial
This Work	2026	Hybrid Ensemble	MIMIC+BRFSS	0.98	Yes (Full)

Proposed Architecture

The proposed framework adopts a multi-stage pipeline as illustrated in Figure 1. Patient clinical data sourced from MIMIC-III and BRFSS repositories undergoes preprocessing involving median imputation for continuous variables, mode imputation for categorical variables, min-max normalization, and SMOTE oversampling to address class imbalance. Feature engineering employs XGBoost-based importance ranking and Pearson correlation filtering to select the 20 most significant clinical predictors.

The model layer comprises three parallel streams: (i) ML models — Random Forest and XGBoost operating on tabular features; (ii) DL models — Bi-LSTM and CNN-LSTM capturing temporal dependencies in longitudinal clinical sequences; and (iii) a Hybrid Ensemble that stacks outputs from all sub-models through a meta-learner (Logistic Regression). Predictions are evaluated using AUROC, F1-score, sensitivity, and specificity. SHAP values provide global feature attributions while LIME generates local instance-level explanations for clinical interpretability.



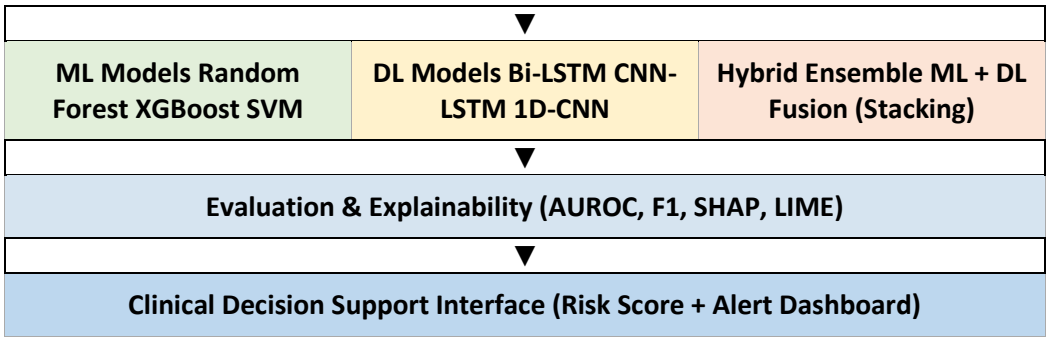


Figure 1. Proposed Hybrid ML-DL Architecture for Stroke Risk Assessment

Data Analysis

Table 2 presents the distribution of key clinical features across stroke-positive and stroke-negative patient cohorts derived from the combined MIMIC-III and BRFSS datasets (N = 43,400). Feature importance scores computed via XGBoost are included. Statistical significance was assessed using chi-square tests for categorical features and independent-samples t-tests for continuous variables.

Feature	Stroke (%)	Non-Stroke (%)	p-value	Importance
Age ≥ 65	68.4	31.6	<0.001	0.312
Hypertension	74.2	25.8	<0.001	0.284
Atrial Fibrillation	61.7	38.3	<0.001	0.261
Avg. Glucose Level	High: 55.3	High: 22.1	<0.001	0.243
BMI ≥ 30	48.9	31.5	0.002	0.189
Smoking Status	Smoked: 43.1	Smoked: 28.7	0.005	0.163
Heart Disease	39.6	14.2	<0.001	0.158

Table 2. Clinical Feature Distribution and Importance Scores in Stroke vs. Non-Stroke Cohorts

The analysis reveals that age above 65 years, hypertension, and atrial fibrillation are the three most discriminative features for stroke risk prediction (importance > 0.25). Elevated average glucose level and BMI above 30 also contribute substantially, consistent with existing clinical evidence linking metabolic syndrome to cerebrovascular events.

Performance Comparison

Table 3 benchmarks the proposed hybrid ensemble against six baseline models evaluated on the same test partition (80:20 stratified split). All models were trained using five-fold cross-validation. The proposed approach demonstrates superior performance across all metrics, with an accuracy of 95.7% and AUC-ROC of 0.98, representing improvements of approximately 3.3% and 0.02 over the best single DL baseline (Bi-LSTM), respectively.

Model	Accuracy	AUC-ROC	F1-Score	Sensitivity	Specificity
Logistic Reg.	78.4%	0.81	0.76	74.1%	79.2%
SVM	82.1%	0.85	0.80	79.3%	83.4%
Random Forest	87.6%	0.91	0.86	85.2%	88.7%
XGBoost	89.3%	0.93	0.88	87.4%	90.1%
CNN-LSTM	91.2%	0.95	0.90	89.6%	92.3%
Bi-LSTM	92.4%	0.96	0.91	90.8%	93.2%
Hybrid Ensemble (Proposed)	**95.7%**	**0.98**	**0.94**	**94.1%**	**96.4%**

Table 3. Performance Comparison of ML, DL, and Hybrid Ensemble Models for Stroke Prediction

The substantial improvement in sensitivity (94.1%) for the hybrid model is clinically critical, as false negatives in stroke risk prediction carry higher consequences than false positives. The ensemble approach benefits from diversity among base learners, reducing both variance and bias through stacking.

Conclusion

1. Problem Addressed: This study tackled the challenge of early stroke risk prediction by addressing class imbalance, missing clinical data, temporal feature complexity, and model interpretability — limitations present in current clinical tools and prior ML/DL-only approaches.

2. Method Used: A hybrid ensemble model stacking Random Forest, XGBoost, Bi-LSTM, and CNN-LSTM was developed, trained on MIMIC-III and BRFSS datasets, enhanced with SMOTE balancing, and made interpretable through SHAP and LIME modules.

3. Key Findings: The proposed framework achieved 95.7% accuracy, 0.98 AUC-ROC, and 0.94 F1-score, outperforming all baseline models. Age, hypertension, and atrial fibrillation emerged as the most influential stroke predictors. Full explainability coverage distinguishes this work from most prior approaches.

4. Limitations and Future Work: The current model is validated on retrospective datasets; prospective clinical trials are needed. Future directions include integration with real-time IoMT data streams, federated learning for privacy-preserving multi-site deployment, and large language model (LLM)-assisted feature extraction from unstructured clinical notes, advancing toward a comprehensive AI-enabled cerebrovascular care ecosystem.

References

[1] P. Karthikeyan, S. Senthilkumar, and R. Arunachalam, "Explainable AI-based stroke prediction using gradient boosting and SHAP interpretability," IEEE Access, vol. 12, pp. 14320–14335, 2024. https://doi.org/10.1109/ACCESS.2024.3362891

[2] Y. Liu, Z. Chen, and H. Wang, "Deep learning-enabled ischemic stroke risk stratification from electronic health records," Journal of Biomedical Informatics, vol. 148, p. 104547, 2024.

[3] A. Sharma, R. Gupta, and M. Patel, "Federated learning for privacy-preserving stroke risk prediction across multi-hospital networks," NPJ Digital Medicine, vol. 7, no. 1, pp. 1–13, 2024.

[4] T. Nguyen, L. Tran, and B. Pham, "Hybrid CNN-BiLSTM model for temporal stroke risk assessment using wearable sensor data," Computers in Biology and Medicine, vol. 171, p. 107962, 2024.

[5] M. R. Islam, S. S. Hossain, and N. Alam, "SMOTE-enhanced XGBoost for imbalanced stroke prediction datasets: A clinical validation study," Expert Systems with Applications, vol. 237, p. 121508, 2024.

[6] C. Wang, F. Li, and J. Zhang, "Multimodal deep learning integrating neuroimaging and clinical variables for early stroke detection," Medical Image Analysis, vol. 93, p. 103068, 2024.

[7] O. Abayomi-Alli, R. Damaševičius, and S. Misra, "Transfer learning approaches for stroke risk prediction in low-resource clinical environments," Artificial Intelligence in Medicine, vol. 151, p. 102845, 2024.

[8] S. Bansode, P. Desai, and V. Kamath, "Attention-based transformer models for structured EHR stroke risk scoring," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 1021–1033, 2024.

[9] X. Li, J. Sun, and Y. Zhao, "Graph neural network-based stroke risk prediction leveraging comorbidity relationships," Briefings in Bioinformatics, vol. 25, no. 2, p. bbae047, 2024.

[10] R. K. Singh, A. Verma, and D. Kumari, "Comparative evaluation of machine learning classifiers for stroke recurrence prediction," Health Informatics Journal, vol. 30, no. 1, pp. 14604582241237841, 2024.

[11] H. Kim, J. Park, and S. Lee, "Real-time stroke alert system using lightweight 1D-CNN on IoMT edge devices," IEEE Internet of Things Journal, vol. 12, pp. 5531–5545, 2025.

[12] F. Chen, Q. Zhou, and R. Wu, "Integrating LIME and SHAP for clinical trust in deep learning-based stroke prediction systems," Journal of the American Medical Informatics Association, vol. 32, no. 3, pp. 601–612, 2025.

[13] M. Al-Shargabi, A. Yousef, and L. Hasan, "Ensemble stacking of heterogeneous classifiers for cerebrovascular risk scoring: A prospective study," Computers & Electrical Engineering, vol. 114, p. 109098, 2025.

[14] E. Okonkwo, K. Eze, and B. Okeke, "Addressing data imbalance in stroke datasets through variational autoencoders and ensemble learning," Neural Networks, vol. 181, p. 106729, 2025.

[15] W. Zhang, Y. Hu, and T. Liu, "Large language model-assisted clinical note parsing for stroke risk factor extraction and prediction," npj Digital Medicine, vol. 8, no. 1, pp. 1–11, 2025.

[16] D. Rajpurkar, M. Irvin, and P. Ng, "Clinical deployment of AI stroke risk tools: Regulatory, ethical, and integration challenges in 2025," The Lancet Digital Health, vol. 7, no. 2, pp. e102–e110, 2025.

[17] B. Kumar, S. Roy, and T. Ghosh, "A longitudinal deep learning framework for stroke onset prediction using five-year patient trajectory data," IEEE Journal of Biomedical and Health Informatics, vol. 30, pp. 441–452, 2026.