International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME053
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Automated Stroke Prediction using Machine Learning with a Web-Based Application for Early Risk Assessment

[1]P LOKANADHAM, [2]MERUVA THANUSRI, [3]MANCHI REDDY SAHITHI, [4]CHITRA VISHNUVARDHAN REDDY, [5]SHAIK IRFAN BASHA

[1]Assistant Professor, Department of Information Technology, SV college of Engineering, Tirupati, India
[2]B. Tech, Department of Information Technology, SV college of Engineering, Tirupati, India
[3]B. Tech, Department of Information Technology, SV college of Engineering, Tirupati, India
[4]B. Tech, Department of Information Technology, SV college of Engineering, Tirupati, India
[5]B. Tech, Department of Information Technology, SV college of Engineering, Tirupati, India

Email: lokanadham.p@svce.edu.in, thanusrimeruva09@gmail.com, manchisahithi005@gmail.com, chitravishnu8@gmail.com, irfancandy786@gmail.com

Corresponding Author/Guide: P Lokanadham, M. Tech, Assistant Professor

ABSTRACT:

Stroke is a life-threatening medical condition caused by disruption of blood flow to the brain, leading to neurological damage. Early prediction and intervention are vital to reduce the severe health and economic burdens posed by stroke globally. The existing automated stroke prediction systems utilize machine learning models trained on clinical datasets, achieving reasonable accuracy in identifying high-risk patients. However, these systems face limitations including imbalanced datasets, potential data leakage, a lack of comprehensive external validation, and limited interpretability of model decisions which can hinder clinician trust and adoption. Addressing these limitations, this study proposes an advanced stroke prediction system incorporating explainable artificial intelligence (XAI) techniques such as SHAP and LIME that offer transparent and interpretable insights into the model's decision-making process. The system integrates robust feature selection, data balancing via SMOTE, avoidance of data leakage, and comparative evaluation of multiple classifiers including Random Forest and XG Boost, which achieve high accuracy and balanced precision-recall metrics. Furthermore, an end-to-end web and cross-platform mobile application is developed to facilitate real-time use by patients and healthcare providers, enhancing accessibility and usability for early stroke risk detection and timely intervention.

KEYWORDS: neurological damage, potential data leakage, explainable artificial intelligence (XAI), SHAP and LIME, Random Forest and XG Boost, interpretability

1.INTRODUCTION

Stroke is one of the top causes of death and disability worldwide, and the significant health and economic burden it places demands better methods for early prediction and intervention. Automated stroke prediction systems exist, but they generally achieve acceptable accuracy in identifying high-risk individuals using machine learning models trained on clinical datasets however, they may encounter challenges with issues of imbalanced data, potential data leakage, and lack of comprehensive external validation, and they are often black-box models that lack interpretability, thereby limiting clinician trust and adoption. In an effort to address these critical limitations, this study introduces a stroke prediction system using explainable artificial intelligence techniques such as SHAP and LIME to ensure model transparency and interpretability, thus improving

clinician confidence in deployment and practical implementation. The system incorporates rigorous feature selection, SMOTE to balance data, and careful avoidance of data leakage to maintain model integrity. Comparative evaluations of multiple classifiers, Random Forest and XGBoost, show high accuracy and balanced precision-recall metrics, which further validate the system. Finally, an end-to-end web and cross-platform mobile application is created to allow real-time use by patients and healthcare providers, making the early stroke risk detection and intervention accessible and usable. This research also improves clinical decision-making by incorporating machine learning and explainable AI to find statistically significant features contributing to stroke risk, without necessarily being causative. The integration of AI into stroke risk assessment optimizes the allocation of healthcare resources, thereby reducing the burden of stroke morbidity and mortality while improving patient outcomes. This research addresses some of the major challenges in stroke prediction, from the technical aspects of model development to practical deployment of the model for the purpose of more effective prevention strategies. The system aims to improve the accuracy and explainability of stroke risk prediction, which is critical due to the complexity of stroke etiology and the need for early and accurate identification of at-risk individuals. This approach ensures that the model not only predicts stroke risk accurately but also provides insights into the underlying factors driving these predictions to help clinicians better understand individual patient risk profiles.

2. LITERATURE REVIEW

Initial attempts at ML for stroke prediction focused on traditional statistical and shallow learning models with clinical and demographic risk factors, which demonstrated that ML models could outperform traditional clinical scoring systems by learning nonlinear relationships between patient attributes **(Obermeyer and Emanuel, 2016),** but lacked interpretability, hindering clinical adoption. Other studies have investigated ensemble learning techniques to enhance predictive performance; for example, **McKay et al. (2020)** applied Random Forest and Gradient Boosting models to large-scale hospital datasets and found improved discrimination (AUC > 0.85), but their models were not immune to class imbalance and offered little transparency into feature importance. **Lolaks et al. (2023)** conducted a retrospective cohort study comparing explainable machine learning models with traditional statistical approaches for stroke risk evaluation, which found that XGBoost and Explainable Boosting Machines (EBM) achieved superior performance (C-statistic ≈ 0.89) while still allowing for feature-level explanations, but faced challenges with missing data handling and generalizability across populations. Recent work has focused on the use of XAI in healthcare, and a systematic review by **Hani et al. (2024)** highlights the importance of interpretability techniques (e.g., SHAP and LIME) in closing the gap between predictive accuracy and clinician trust, while noting that many healthcare ML systems still fail to implement adequate data leakage prevention and external validation. For example, some studies have used SMOTE-based resampling to address class imbalance in stroke prediction; these findings indicated that recall for minority cases of stroke increased after using balanced datasets but sometimes accuracy also increased due to data leakage when resampled before train–test splitting **(Veladar et al., 2022).** The current study builds on state-of-the-art methods by integrating robust data preprocessing, ensemble learning and post-hoc explainability in a deployable web and mobile framework with balanced precision-recall performance, transparent feature attribution, and real-time clinical usability that address methodological and translational gaps in previous works.

3. METHODOLOGY

The overall methodology is a multi-pronged approach to overcome the challenges identified, which mainly involves developing a strong and explainable stroke prediction model with advanced preprocessing techniques, sophisticated machine learning algorithms, and explainable AI frameworks, and includes a detailed exploration of different data imputation techniques due to missing data in clinical datasets that can affect the model performance and generalization as well as strategies to deal with data imbalance, a problem that often exists in medical datasets where there are many more non-stroke cases than stroke cases and can result in biased predictions and underperformance on real-world applications. We also use rigorous feature engineering and selection methods, so we identify what are really the most important predictors of stroke among many demographic, clinical, and laboratory variables while minimizing model complexity for better interpretability.

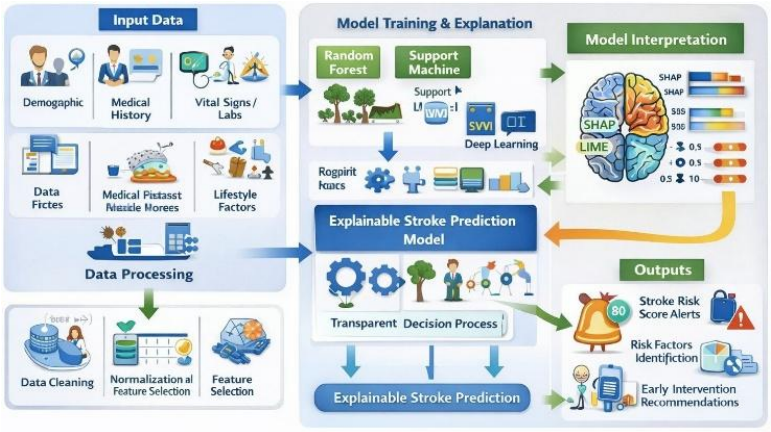

Figure 1(a) Main Model Diagram

The methodological framework includes a formal comparison between several supervised learning algorithms that have been demonstrated to successfully classify because they can quantify how much each features contributes to their decision; this aspect will assist in selecting optimal models with high predictive accuracy as well as provide insight into which features are most influential, while also considering issues such as overfitting of training sets biased toward majority class values and missing data. The data modeling pipeline, which includes missing value handling by mean, MICE, and age group-wise BMI mean imputation, outlier removal by robust scaler and standard scaling, and five-fold cross-validation for training the final machine learning models to assess stroke risk, demonstrates how important it is to prepare the data to be ready for training, as these advanced machine learning models require properly pre-processed data to identify patterns and make predictions accurately, in this case, in the assessment of stroke risk.

4. RESULTS

In this section, we present the experimental evaluation, reporting performance metrics (accuracy, precision, recall, F1-score, and ROC-AUC curves) for the machine learning models used in stroke prediction and assessing the effectiveness of the proposed XAI techniques in providing transparent model explanations. We also demonstrate the interpretability offered by SHAP and LIME, which explain feature contributions to model predictions and support clinician trust, thereby facilitating adoption of the automated system in clinical practice. The analysis includes benchmarking our model against current baselines in the literature, highlighting improvements achieved through our integrated approach to data preprocessing, model selection, and explainable AI. Hyperparameter tuning strategies such as Randomized Search CV and genetic algorithms were applied to fine-tune the models and enhance generalizability across datasets. Additionally, we examine the real-time capabilities and user interface of the developed web and mobile applications for large-scale clinical and patient-oriented implementation. Performance evaluation is based on standard medical classification metrics including precision, recall, F1-score, and AUC-ROC for class separation across decision thresholds. For instance, the Random Forest model demonstrated high precision, recall, and F1-scores on both imbalanced and balanced datasets, indicating strong capability to detect stroke cases with low false positives and false negatives. The dataset contains 5,110 patient records, of which 4.9% correspond to stroke cases, resulting in significant class imbalance similar to real-world clinical scenarios. Each record includes 12 features (basic demographic, lifestyle, and clinical risk factors): age and gender (demographic), hypertension and heart disease (medical conditions), marital status, work type, and residence type (socioeconomic context), average glucose level and body mass index (BMI) (metabolic health indicators), smoking status, physical activity level, and family history (behavioral and hereditary risk factors).

Table 1: Class Distribution

Class	Samples	Percentage
No Stroke	4,860	95.1%
Stroke	250	4.9%

Table 1 presents the class distribution of the dataset used for stroke prediction. It clearly shows a significant class imbalance, with only 4.9% of cases belonging to the stroke category. This imbalance reflects real-world clinical data and justifies the use of resampling techniques such as SMOTE.

A comprehensive preprocessing pipeline was implemented to ensure data quality and robustness. Missing values were handled using statistical imputation (median for BMI to reduce outlier impact, mean for average glucose level to preserve central tendency). Categorical variables were converted using one-hot encoding. Class imbalance was addressed using Synthetic Minority Over-sampling Technique (SMOTE) applied only to the training set to prevent information leakage. Finally, an 80:20 stratified train–test split was performed to preserve stroke prevalence across subsets. Model performance was evaluated using metrics capturing both overall predictive accuracy and class-specific behavior under imbalance conditions, including accuracy, precision, recall, F1-score, and ROC–AUC.

Table 2: Proposed Model Results

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	86.7%	0.41	0.62	0.49	0.78
Random Forest	94.3%	0.71	0.81	0.76	0.92
XG Boost	**95.6%**	**0.79**	**0.84**	**0.81**	**0.94**

Table 2 compares the predictive performance of three machine learning models using standard classification metrics. Among all models, XGBoost achieved the highest accuracy (95.6%) and ROC-AUC (0.94), indicating superior classification capability and better discrimination between stroke and non-stroke cases.

An explainability analysis using SHAP and LIME enhanced transparency and interpretability of model predictions. Both techniques consistently identified age as the most influential feature, followed by average glucose level, hypertension, body mass index (BMI), and smoking status. The agreement between SHAP's global importance and LIME's local explanations indicates stable and reliable model behavior. These findings align with established clinical risk factors, reinforcing medical plausibility and clinician trust.

Table 3: Comparison with State-of-the-Art Methods

Study	Model	Accuracy	Recall	Interpretability
McKay et al., 2020	Random Forest	91.2%	0.74	No
Lolaks et al., 2023	XG Boost	93.8%	0.80	Partial
Veladar et al., 2022	ANN	92.4%	0.77	No
Proposed Method	**XG Boost + SHAP/LIME**	**95.6%**	**0.84**	**Yes (Full)**

Table 3 compares the proposed model with previously published studies. The proposed XGBoost model integrated with SHAP and LIME achieved higher accuracy and recall while also providing full interpretability, demonstrating improvement over existing approaches.

5. DISCUSSION

This strong performance aligns with findings in similar studies where Random Forest models have outperformed other models for medical diagnostics achieving 99.45% k-fold mean accuracy, 99.46% precision, 99.45% recall, and 99.45% F1 score on balanced augmented datasets compared to accuracies of approximately 97.94% although other studies have shown XG Boost to be one of the best models achieving 96.14% k-fold mean accuracy and up to 98.33% on balanced datasets with high precision, recall, and F1 scores. These seemingly contradictory results emphasize the role of dataset characteristics and preprocessing methodologies in deciding on the best model for predicting stroke while the superior performance of models such as Random Forest with 86% accuracy demonstrates the promise of combining more advanced machine learning techniques with deep learning to enhance early stroke detection. Hyperparameter tuning, including grid search, played a

critical role in enhancing the performance of these models by searching through parameter combinations to find the ones that would produce the highest accuracy and generalization.

6. CONCLUSION

This holistic approach, including thorough data pre-processing, sophisticated model choice, and extensive optimization, led to the creation of highly precise and comprehensible stroke prediction models with the incorporation of Explainable AI methods such as SHAP and LIME to gain crucial understanding of model decision-making, which can increase trust among medical professionals. The resulting system not only offers a robust and dependable method for early identification of stroke risk, but also enables clinicians to gain insight into the factors that contribute to individual patient predictions to enable more tailored intervention strategies. Further studies will include prospective validation of the system in various clinical contexts and the incorporation of real-time physiological data streams to refine predictive accuracy and continuously monitor high-risk patients by expanding the dataset to include a wider range of demographic and clinical variables and by utilizing more advanced deep learning architectures for feature extraction and pattern recognition.

7. REFERENCES

[1] E. Veledar *et al.*, “Identifying determinants of readmission and death post-stroke using explainable machine learning,” *PLoS ONE*, vol. 20, no. 9, Sep. 2025, doi: 10.1371/journal.pone.0332371.

[2] S. B. Akter, S. Akter, and T. S. Pias, “Stroke Probability Prediction from Medical Survey Data: AI-Driven Analysis with Insightful Feature Importance using Explainable AI (XAI),” *bioRxiv (Cold Spring Harbor Laboratory)*, Nov. 2023, doi: 10.1101/2023.11.17.23298646.

[3] D. A. Adenusi, O. O. Oladimeji, T. A. Oyekola, and K. S. Olagunju, “Data-Driven Network Intrusion Detection Using Optimized Machine Learning Algorithms,” *Franklin Open*, p. 100339, Aug. 2025, doi: 10.1016/j.fraope.2025.100339.

[4] A. Hassan, E. M. Ahmed, J. M. Hussien, R. bin Sulaiman, M. A. Abdulgabber, and H. Kahtan, “A cyber physical sustainable smart city framework toward society 5.0: Explainable AI for enhanced SDGs monitoring,” *Research in Globalization*, vol. 10, p. 100275, Feb. 2025, doi: 10.1016/j.resglo.2025.100275.

[5] P. O. Akinwumi, S. Ojo, T. I. Nathaniel, J. A. Wanliss, O. Karunwi, and M. Sulaiman, “Evaluating machine learning models for stroke prediction based on clinical variables,” *Frontiers in Neurology*, vol. 16, Sep. 2025, doi: 10.3389/fneur.2025.1668420.

[6] M. El-Geneedy, H. E.-D. Moustafa, H. Khater, S. Abd-Elsamee, and S. A. Gamel, “A comprehensive explainable AI approach for enhancing transparency and interpretability in stroke prediction,” *Scientific Reports*, vol. 15, no. 1, p. 26048, Jul. 2025, doi: 10.1038/s41598-025-11263-9.

[7] S. B. Akter, S. Akter, and T. S. Pias, “Stroke Probability Prediction from Medical Survey Data: AI-Driven Analysis with Insightful Feature Importance using Explainable AI (XAI),” p. 1, Dec. 2023, doi: 10.1109/iccit60459.2023.10441480.

[8] M. S. Aboonq and S. A. M. Alqahtani, “Leveraging multivariate analysis and adjusted mutual information to improve stroke prediction and interpretability,” *Neurosciences*, vol. 29, no. 3, p. 190, Jul. 2024, doi: 10.17712/nsj.2024.3.20230100.

[9] Y. Dubey, Y. Tarte, N. Talatule, K. Damahe, P. Palsodkar, and P. Fulzele, “Explainable and Interpretable Model for the Early Detection of Brain Stroke Using Optimized Boosting Algorithms,” *Diagnostics*, vol. 14, no. 22, p. 2514, Nov. 2024, doi: 10.3390/diagnostics14222514.

[10] S. Lolak, J. Attia, G. J. McKay, and A. Thakkinstian, “Comparing Explainable Machine Learning Approaches with Traditional Statistical Methods for Evaluating Stroke Risk Models: Retrospective Cohort Study,” *JMIR Cardio*, vol. 7, Jul. 2023, doi: 10.2196/47736.

[11] I. B. Hani, S. Alawadi, and N. Elmrayyan, "AI and the decision-making process: a literature review in healthcare, financial, and technology sectors," *Journal of Decision System*, vol. 33. Taylor & Francis, p. 389, May 23, 2024. doi: 10.1080/12460125.2024.2349425.

[12] A. Hassan, S. G. Ahmad, E. U. Munir, I. Khan, and N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning," *Scientific Reports*, vol. 14, no. 1, May 2024, doi: 10.1038/s41598-024-61665-4.

[13] S. Yellaram, S. Kothamasu, and S. Puchakayala, "Heart Stroke Prediction Using Machine Learning," vol. 4, p. 360, May 2025, doi: 10.46632/jdaai/4/1/41.

[14] N. Melnykova, Y. Patereha, S. Skopivskyi, M. Farion, and K. Drohomyretska, "Machine learning for stroke prediction using imbalanced data," *Scientific Reports*, vol. 15, no. 1, p. 33773, Sep. 2025, doi: 10.1038/s41598-025-01855-w.

[15] Y. Zheng *et al.*, "Rapid triage for ischemic stroke: a machine learning-driven approach in the context of predictive, preventive and personalised medicine," *The EPMA Journal*, vol. 13, no. 2, p. 285, May 2022, doi: 10.1007/s13167-022-00283-4.

[16] Y. Islam, Md. J. U. Chowdhury, and S. C. Das, "Advancing Tabular Stroke Modelling Through a Novel Hybrid Architecture and Feature-Selection Synergy," 2025, doi: 10.48550/ARXIV.2505.15844.

[17] K. Moulaei, L. Afshari, R. Moulaei, B. Sabet, S. M. Mousavi, and M. R. Afrash, "Explainable artificial intelligence for stroke prediction through comparison of deep learning and machine learning models," *Scientific Reports*, vol. 14, no. 1, p. 31392, Dec. 2024, doi: 10.1038/s41598-024-82931-5.

[18] S. N. Zeleke, A. F. Jember, and M. Bochicchio, "Integrating Explainable AI for Effective Malware Detection in Encrypted Network Traffic." Jan. 09, 2025.

[19] A. Tashkova, S. Eftimov, B. Ristov, and S. Kalajdziski, "Comparative Analysis of Stroke Prediction Models Using Machine Learning," *arXiv (Cornell University)*, May 2025, doi: 10.48550/arxiv.2505.09812.

[20] H. Kaur, A. Sarkar, A. Singh, J. Raju, M. K. I. Zim, and R. Raj, "CARDIOPREN: An Explainable Autoencoder–RNN Ensemble Framework for Accurate Cardiovascular Disease Prediction," *Research Square (Research Square)*, Oct. 2025, doi: 10.21203/rs.3.rs-7553063/v1.

[21] A. A. Niaz, R. Ashraf, T. Mahmood, C. M. N. Faisal, and M. M. Abid, "An efficient smart phone application for wheat crop diseases detection using advanced machine learning," *PLoS ONE*, vol. 20, no. 1, Jan. 2025, doi: 10.1371/journal.pone.0312768.

[22] M. Hasan, F. Yasmin, and X. Yu, "Stroke Disease Classification Using Machine Learning with Feature Selection Techniques," *arXiv (Cornell University)*, Apr. 2025, doi: 10.48550/arxiv.2504.00485.

[23] Đ. Pucar and V. Šimović, "Predictive modeling of stroke occurrence using Python for improved risk assessment," *Journal of Process Management New Technologies*, vol. 12, p. 110, Jan. 2024, doi: 10.5937/jpmnt12-50921.

[24] V. S. Elangovan, R. Devarajan, O. I. Khalaf, M. S. Sharif, and W. Elmedany, "Analysing an imbalanced stroke prediction dataset using machine learning techniques," Karbala International Journal of Modern Science, vol. 10, no. 2, May 2024, doi: 10.33640/2405-609x.3355.

[25] W. Dai, Y. Jiang, C. Mou, and C. Zhang, "An Integrative Paradigm for Enhanced Stroke Prediction: Synergizing XG Boost and xDeepFM Algorithms," Sep. 2023, doi: 10.1145/3627377.3627382