IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141259

Brain Stroke Prediction

Mrs. R S Geethanjali[1], M Sowmya[2], M Meghana[3], and R Prudvi Ganesh[4]

Assistant Prof., Department of CSE, K. S. School of Engineering and Management, Bangalore[1]

Student, Department of CSE, K. S. School of Engineering and Management, Bangalore[2-4]

Abstract: Brain Stroke prediction introduces a robust, multi-modal machine learning system engineered to precisely forecast brain stroke risk by integrating two fundamentally different data sources: conventional, structured clinical data and complex, unstructured CT/MRI neuroimaging. The system is built upon a dual-stream architecture: the gradient boosting algorithm XGBoost is deployed to analyze the patient's record features (e.g., demographics and history), and a deep convolutional network, EfficientNet-B0, is dedicated to extracting visual pathological indicators from the brain scans. A core objective is to ensure system trustworthiness through the application of Explainable AI (XAI), specifically SHAP (Shapley Additive Explanations), which guarantees clarity and interpretability for medical professionals. This scalable solution marks a significant advancement in early stroke detection and enables evidence-based clinical decision support.

Keywords: Explainable AI (XAI), SHAP (Shapley Additive Explanations) , EfficientNet-B0, XGBoost, Convolutional Neural Network (CNN).

I. INTRODUCTION

The project details the creation of an advanced multi modal machine learning platform for stroke risk prediction, which addresses the inherent deficiencies of diagnostic models relying on singular data inputs, particularly in a critical public health domain. The system employs a two-pathway architecture to synthesize information from both structured clinical records and unprocessed CT/MRI scans: the XGBoost technique is used to develop a model based on patient history and demographics, while the EfficientNet-B0 deep learning network is utilized to identify and classify features within the image data. A definitive, confident prediction is achieved by synthesizing the outputs of these two specialized models through a weighted averaging ensemble method. This methodology aims to deliver exceptionally accurate early stroke detection results, and its utility is further enhanced by implementing XAI tools like SHAP, which ensures model transparency and builds user confidence for practical medical implementation.

II. RELEVANT LITERATURE

Early Prediction of Stroke using Machine Learning:
This research focused on developing and comparing various conventional machine learning models for predicting brain stroke risk, using only patient clinical records. The authors emphasized that timely prediction is crucial for mitigating severe health consequences. The study analyzed a dataset of 5,110 patient records with 12 attributes, including age, sex, and pre-existing conditions. After standard preprocessing, several algorithms were evaluated using metrics like accuracy and F1-score. The Random Forest and XGBoost models were identified as the top performers, both achieving a high prediction accuracy of 96%. Key risk factors isolated by the analysis included smoking habits, residential location, and pre-existing heart conditions. The authors suggested that this predictive model should be integrated into accessible web-based tools for widespread early risk assessment.

Brain Stroke Prediction Using Machine Learning Techniques
Ibrahim Almubark's work utilized a public Kaggle dataset containing 5,110 clinical and demographic records to predict stroke risk. A critical part of the data preparation involved handling missing BMI values using the k-Nearest Neighbors imputation method, followed by scaling the features with Min-Max-Scaler and transforming categorical features with One-Hot-Encoder, resulting in 17 features. The author evaluated five supervised algorithms, including Random Forest and an Artificial Neural Network (ANN), and fine-tuned them using Bayesian hyper-parameter optimization. The performance evaluation placed particular emphasis on average precision, noting its suitability for severely imbalanced datasets. The concluding recommendation was for future work to expand the dataset size and incorporate complex sources like medical imaging and genomic information.

III. SYSTEM DESIGN AND METHODOLOGY

The system utilizes a Dual-Model Architecture connected by a final ensemble layer. This design ensures that the prediction leverages both structured and unstructured patient information:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141259

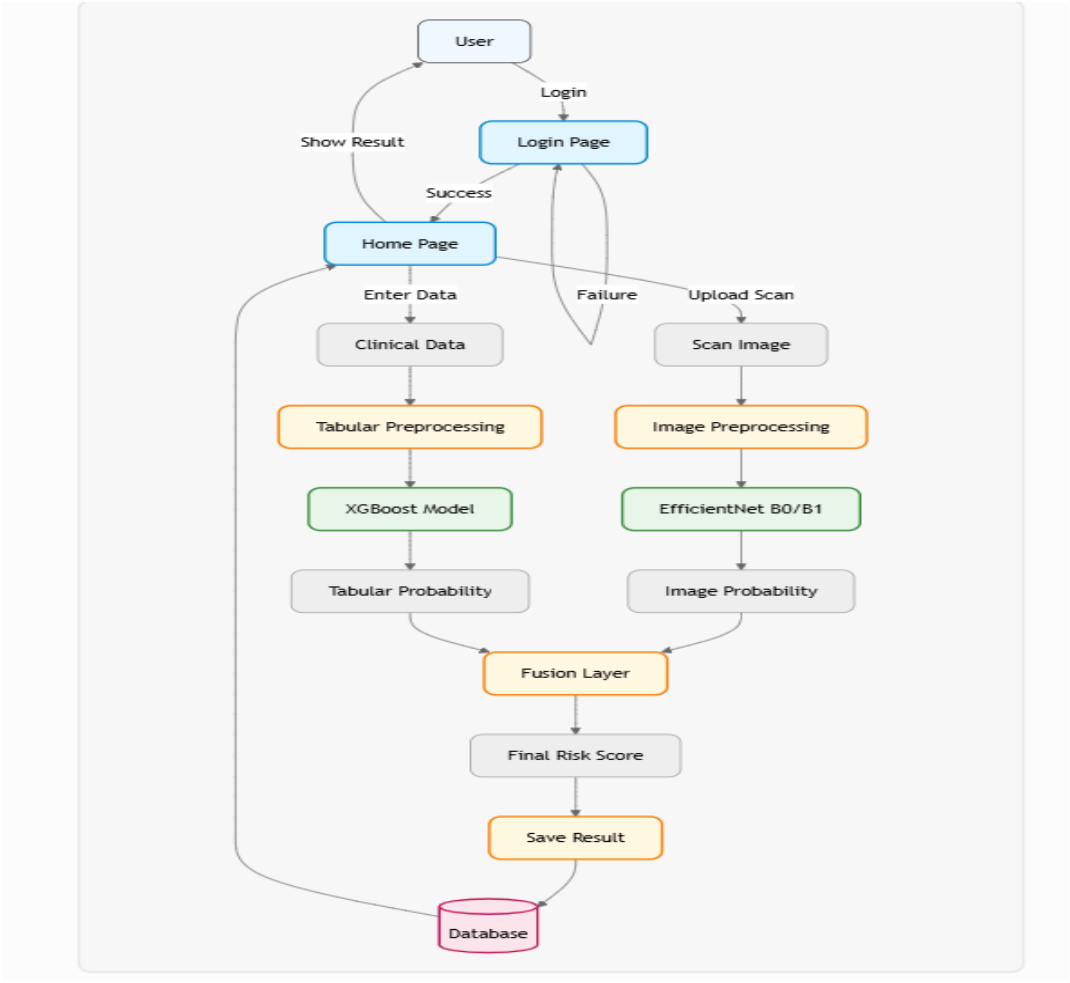

Fig 1. Work Flow Diagram

A. System Design

1. Tabular Processing Stream: Handles all structured patient data, including demographics, vital signs, and medical history.
2. Image Processing Stream: Handles unstructured data from medical scans, specifically CT or MRI images.
3. Fusion Layer: Combines the probability outputs from the two streams using a Weighted Ensemble Averaging method to derive the final, comprehensive risk score.
4. Explainability Component (XAI): The system is designed to integrate SHAP analysis to provide clear interpretations of the prediction drivers, enhancing transparency for clinicians.
5. Deployment Architecture: The model is integrated into a web-based platform with a Python backend

(Flask/Django) and an HTML/CSS/JavaScript frontend, ensuring a user-friendly interface for clinicians.

B. Methodology:

The project methodology is divided into data preprocessing, model implementation for each stream, and final output integration:

1. Input Preparation and Segregation: Clinical and image data undergo highly customized preparation procedures, including feature normalization, categorical encoding (for structured data), and data augmentation (for visual data). This process yields two specialized, clean inputs.
2. Dual-Model training: The tabular input is trained Independent Model Training: The structured data input is processed and trained using the XGBoost algorithm, employing a Regularized Objective Function and optimizing parameters through a randomized search approach. Concurrently, the image input is trained using EfficientNet-B0, which relies on Transfer Learning and features a fine tuned, bespoke classification head.
3. cores derived from both prediction models are fed into a Fusion Layer utilizing a Weighted Ensemble Averaging technique to generate the definitive stroke risk score. The system's dependable performance is validated

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141259

extensively using metrics like Accuracy, ROC-AUC, F1-score, and PR-AUC. Furthermore, model decision clarity is ensured by applying SHAP analysis.

IV. RESULTS AND DISCUSSION

The rigorous evaluation confirmed that the multi-modal architecture successfully overcomes the drawbacks of reliance on single-source prediction models, thereby validating the central hypothesis that combining diverse data notably improves the accuracy and reliability of stroke risk assessment. The final weighted ensemble model showcased strong performance, as evidenced by high ROC-AUC and PR-AUC scores, which is crucial for achieving dependable predictions in medical datasets that are frequently imbalanced. The superior results from the weighted ensemble indicate that clinical and imaging data offer complementary diagnostic information; specifically, the XGBoost pathway establishes a fundamental risk foundation from patient history, while the EfficientNet-B0 pathway detects granular, pathological visual indicators from CT/MRI scans. Importantly, the practical clinical value of the system is enhanced by the integrated Explainable AI (XAI) feature, which uses SHAP analysis to provide transparency by pinpointing influential factors such as patient age or visual anomalies. This clarity empowers clinicians to trust the system's prediction and tailor preventative care plans. Future development will focus on transitioning from this current two-stage architecture to end to-end multi-modal fusion models to enable deeper feature interactions and further maximize predictive accuracy.

V. CONCLUSION AND FUTURE WORK

CONCLUSION

The rigorous performance assessment confirms that the multi-modal architecture effectively addresses the inherent limitations associated with single data source prediction models, validating the central concept that synthesizing diverse data substantially elevates stroke risk assessment reliability and precision. The final weighted ensemble model exhibited formidable performance, particularly achieving high scores in ROC-AUC and PR-AUC, a necessity for generating reliable predictions from often imbalanced medical cohorts. The enhanced results provided by the weighted ensemble indicate a complementary relationship between clinical and imaging diagnostics ; the XGBoost stream establishes the underlying risk profile from historical patient data, while the EfficientNet-B0 stream identifies critical, pathological visual signatures within the CT/MRI scans. The system's utility in a clinical setting is profoundly strengthened by the embedded Explainable AI (XAI) capability, which employs SHAP analysis to provide transparency by pinpointing the most influential factors, whether they are patient demographics or visual anomalies. This clarity allows practitioners to trust the model's outcome and customize preventative treatment strategies. Moving forward, development efforts must prioritize shifting the system from its current two-stage design toward end-to-end multi-modal fusion models to facilitate deeper feature interactions and further optimize predictive success.

Future Work

The completed framework presents several clear opportunities for subsequent development and refinement. Firstly, the system can be expanded to include additional impactful, multi-modal data streams, such as Electrocardiogram (ECG) readings or genetic indicators, thereby boosting its predictive capability. Secondly, to advance beyond the existing two-stage architecture, the weighted fusion layer could be superseded by an end-to-end multi-modal fusion model—potentially leveraging transformer-based designs—to allow for deeper cross-modal feature interplay and potentially elevate accuracy. Lastly, for full clinical integration, attention must be paid to optimizing the model for deployment within a real-time clinical environment, which involves seamless integration with existing Hospital Information Systems (HIS) and enhancing the XAI component to deliver personalized, actionable clinical advice.

REFERENCES

[1] G. Revathy, U. Sesadri, S. Theodore, J. J. P. Thilagavathy, S. Senthilvadivu, and V.S. Murugan, "Early Prediction of Stroke using Machine Learning," in Proc. 4th Int. Conf. Electron. Sustain. Commun. Syst. (ICESC), Coimbatore, India, Jul. 2023, pp. 1074–1075, doi: 10.1109/ICESC57686.2023.10193052.

[2] I. Almubark, "Brain Stroke Prediction Using Machine Learning Techniques," in *Proc. IEEE Int. Conf. on Big Data (BigData)*, Sorrento, Italy, 2023, pp. 6104–6108, doi: 10.1109/BigData59044.2023.10386474.

[3] R. Suryawanshi, V. Kulkarni, P. Ghule, K. Patil, H. Patil, and Y. Manala, "Brain Stroke Prediction using Logistic Regression with Logarithmic Transform," in Proc. Int. Conf. Sustainable Expert Systems (ICSES), Coimbatore, India, Mar. 2024, pp. 873–877,doi:10.1109/ICSES63445.2024.10763034.

[4] M. M. H. Bhuiyan, S. Akter, A. K. Acharya, and N. K. Ray, "Enhanced Deep Learning Hybrid Model for the Prediction of Brain Stroke," in *Proc. Int. Conf. Emerging Systems and Intelligent Computing (ESIC)*, Bhubaneswar, India, 2025, pp. 927–931, doi: 10.1109/ESIC64052.2025.10962579.

