Stroke Prediction Using XGboost and a Fusion of XGboost with Random Forest

Thella Ajay Kumar[1], Govada Eswar[2], Veeragoni Laxman Sai[3], Mr B Maria Joseph[4]

[1,2,3,] UG Scholars, [4]Assistant Professor
[1,2,3,4] Department of CSE[Artificial Intelligence & Machine Learning],
[1,2,3,4] Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

---***---

Abstract - Stroke is a life-threatening medical condition caused by disrupted blood flow to the brain, representing a major global health concern with significant health and economic consequences. Researchers are working to tackle this challenge by developing automated stroke prediction algorithms, which can enable timely interventions and potentially save lives. As the global population ages, the risk of stroke increases, making the need for accurate and reliable prediction systems more critical. In this study, we evaluate the performance of an advanced machine learning (ML) approach, focusing on XGBoost and a hybrid model combining XGBoost with Random Forest, by comparing it against six established classifiers. We assess the models based on their generalization ability and prediction accuracy. The results show that more complex models outperform simpler ones, with the best-performing model achieving an accuracy of 96%, while other models range from 84% to 96%. Additionally, the proposed framework integrates both global and local explainability techniques, providing a standardized method for interpreting complex models. This approach enhances the understanding of decision-making processes, which is essential for improving stroke care and treatment. Finally, we suggest expanding the model to a web-based platform for stroke detection, extending its potential impact on public health..*[2]*

Key Words: *Stroke , Machine Learning , Prediction , XG Boost , Accuracy*

1 INTRODUCTION

Stroke is increasingly regarded as one of the top causes of death and disability due to its rising occurrence worldwide. In order to reduce stroke-related mortality and long-term impairment, early intervention is essential. However, conventional techniques for estimating the risk of stroke are frequently laborious and prone to mistakes. Furthermore, machine learning models in the healthcare industry are increasingly required to be transparent and explainable. Using an interpretable machine learning model can help doctors make better treatment decisions by giving them important information about the variables that affect a patient's risk of stroke. According to statistics from the World Stroke Organization, 13 million individuals worldwide suffer a stroke every year, which results in 5.5 million deaths. Stroke is one of the leading causes of death and disability worldwide, impacting a patient's entire life, including their family, social circle, and place of employment. There is a widespread misperception that stroke only affects specific demographics, such as the elderly or those with underlying medical conditions.Based on a number of clinical risk variables, machine learning algorithms have recently demonstrated significant promise in properly predicting the risk of stroke. Clinicians might potentially lower the incidence of stroke-related complications and improve patient outcomes by using these algorithms to identify high-risk patients and take early intervention measures.

2 LITERATURE SURVEY

The prediction and early detection of stroke has become a focal point of medical research due to the devastating and often irreversible impact strokes have on individuals and their families. With the global burden of stroke-related morbidity and mortality on the rise, researchers and clinicians have been motivated to seek out more accurate, data-driven approaches that go beyond traditional clinical assessment and risk scoring.

Historically, stroke risk assessment relied heavily on clinical expertise, patient history, and basic statistical models. While these methods have provided a foundation for understanding risk, they are often limited by their inability to capture the complex interplay of genetic, lifestyle, and physiological factors that contribute to stroke. This has led to a growing interest in the application of machine learning (ML) techniques, which are uniquely equipped to analyze large, multifaceted datasets and uncover hidden patterns that may elude conventional analysis.

Among the various ML algorithms explored, XGBoost has emerged as a standout performer in the realm of stroke prediction. XGBoost's strength lies in its gradient boosting framework, which builds an ensemble of decision trees in a sequential manner, each tree correcting the errors of its predecessor. This approach not only enhances predictive accuracy but also mitigates overfitting-a common challenge in medical data analysis. Multiple studies have reported that XGBoost consistently outperforms traditional models such as Logistic Regression and Support Vector Machines (SVM), particularly when dealing with high-dimensional clinical data. For example, XGBoost has achieved accuracy rates exceeding 90% in several stroke prediction tasks, demonstrating its potential as a reliable tool for early intervention.

However, the complexity of stroke risk factors has also inspired researchers to explore ensemble methods that combine the strengths of multiple algorithms. The Random Forest algorithm, known for its robustness and interpretability, has been widely used in medical diagnostics. By constructing a multitude of decision trees and aggregating their predictions, Random Forests are able to handle noisy data and complex feature interactions effectively. Notably, Random Forests also provide valuable insights into feature importance, helping clinicians identify which variables-such as age, blood pressure, and glucose levels-are most influential in predicting stroke risk.

Building on these successes, recent research has focused on fusing XGBoost with Random Forest to create hybrid models that capitalize on the unique advantages of both algorithms. These fusion models, often implemented through stacking or weighted averaging, have demonstrated even greater predictive power and generalizability. For instance, studies have shown that combining XGBoost and Random Forest can increase sensitivity and specificity, reducing the likelihood of both false positives and false negatives-a critical consideration in clinical settings where early detection can save lives.

Feature selection remains a cornerstone of effective stroke prediction models. Techniques such as Recursive Feature Elimination (RFE), LASSO regularization, and tree-based importance ranking are frequently employed to distill the most relevant predictors from vast arrays of clinical data. This not only improves model accuracy but also enhances interpretability, making it easier for healthcare professionals to trust and act upon the model's recommendations.

The use of publicly available datasets, such as those from the Framingham Heart Study and various hospital repositories, has played a pivotal role in advancing stroke prediction research. These datasets provide a standardized benchmark for evaluating different models and facilitate the reproducibility of results across studies.

In summary, the literature strongly supports the integration of advanced machine learning techniques-particularly XGBoost, Random Forest, and their fusion-for the prediction of stroke. These approaches offer a compelling blend of accuracy, robustness, and interpretability, paving the way for more personalized and effective preventive care. As the field continues to evolve, it is clear that the thoughtful application of ensemble learning and optimal feature selection will remain at the heart of innovation in stroke risk prediction.

3 PROBLEM STATEMENT

One kind of deep learning model made specifically for processing structured grid data, like pictures, is called a convolutional neural network (CNN). CNNs are particularly well-suited for applications such as speech recognition, object detection, and image classification. CNNs' primary function is to automatically and hierarchically extract features from unprocessed input data. A standard CNN is made up of the following essential parts:

Activation functions, usually ReLU, introduce non-linearity; pooling layers reduce spatial dimensions to make the model computationally efficient; fully connected layers interpret the extracted features to perform tasks like classification; and convolutional layers apply filters to detect low-level features like edges or textures. As the input data moves through the network's layers, CNNs may efficiently identify ever-more-complex features by utilizing these layers.

The disadvantages of the current system include: 1. The computational cost of training and deploying CNNs; 2. The need for a large dataset; and 3. Memory and storage requirements.

4 PROPOSED METHODOLOGY

To address the pressing issue of stroke, our proposed method automates prediction and intervention using state-of-the-art machine learning techniques. Our system adopts a firm stance to serve the increasing number of individuals who are at risk due to aging populations, with a focus on lowering the global impacts of stroke.
The primary focus is on creating precise and effective prediction algorithms to enable timely and potentially life-saving responses.

1.Providing precise stroke prediction results, which are necessary for timely care and intervention.
2. Assisting in the identification of critical health and self-reported status markers that predict stroke and facilitating targeted treatment. Enhancing the ability to forecast stroke in general and extrapolate outcomes from health and personal status data.

Unlike deep learning approaches like CNNs, which require large-scale datasets and significant computational power, our system is lightweight and effective for structured, tabular medical data. The methodology emphasizes both prediction accuracy and interpretability, enabling health professionals to understand and act upon model outputs with confidence.

The proposed model works by integrating two powerful ensemble algorithms:

- **XGBoost**, known for its gradient boosting framework and high predictive performance.
- **Random Forest**, recognized for its robustness and generalization ability.

By fusing these two models, our approach benefits from:

- The precision and speed of XGBoost,
- The stability and ensemble strength of Random Forest.

This hybrid setup is designed to extract critical patterns from patient data—such as age, glucose levels, hypertension, BMI, smoking habits, and heart disease history—allowing for accurate stroke predictions. The system also incorporates both **global and local model explanation techniques**, ensuring transparency in decision-making and helping clinicians understand the key factors contributing to a high stroke risk prediction.

4.1 Advantages of Using XGBoost in Stroke Prediction

1. **High Accuracy and Efficiency** XGBoost is known for its exceptional predictive performance. In your project, it consistently delivers high accuracy (up to 96%) in detecting stroke cases, outperforming many traditional classifiers.
2. **Built-in Regularization**XGBoostincorporates L1 (Lasso) and L2 (Ridge) regularization, which helps **prevent overfitting**, especially important in medical datasets where the number of features may be large.
3. **Handles Missing Values Automatically** Medical data often has gaps. XGBoost can automatically **learn the best way to handle missing values**, reducing the need for extensive preprocessing.
4. **Feature Importance Ranking** The model provides **inherent feature importance scores**, helping you identify which health indicators (e.g., BMI, glucose level, age) are most influential in stroke prediction. This is vital for explainability in healthcare.
5. **Scalability and Speed** XGBoost is optimized for speed and memory efficiency. It can handle **large datasets with low training time**, making it ideal for real-time or batch-based hospital systems.
6. **Handles Non-linear Relationships** Stroke risk factors often interact in non-obvious ways. XGBoost captures **complex, non-linear feature interactions** better than linear models.
7. **Parallel Processing and GPU Support** Training can be accelerated using **multi-core CPUs or GPUs**, beneficial when scaling your model to national or regional healthcare datasets.
8. **Cross-validation Support Built-in** XGBoost supports **in-model k-fold cross-validation**, simplifying evaluation and improving generalizability without requiring external loops.
9. **Compatibility with Ensemble Fusion** It integrates well with other models like Random Forest in your hybrid setup, allowing **flexible ensemble strategies** to boost prediction performance.
10. **Open Source and Python-Compatible** Being widely used and open-source, XGBoost has excellent community support, rich

documentation, and seamless integration with **Python, Flask, and Scikit-learn**, which you're using for this project.

4.2 PROPOSED TECHNIQUE USED OR ALGORITHM USED

1) Dataset Preparation

The foundation of this study was a publicly available dataset, obtained in CSV format under the name *"stroke-dataset.csv."* The dataset consists of **5110 records**,each capturing individual-level demographic, health, and lifestyle information. Key features include age, gender, hypertension status, heart disease, average glucose level, body mass index (BMI), and smoking status. The target variable for prediction is the binary attribute "stroke," where a value of **1** indicates the occurrence of a stroke, and **0** indicates otherwise.

2)Library Imports and Programming Environment

All experiments were conducted using the Python programming language due to its robust ecosystem of machine learning libraries. Essential packages included:

- Pandas and NumPy for structured data manipulation,
- Matplotlib for basic visualization tasks,
- Scikit-learn for preprocessing, model evaluation, and Random Forest implementation,
- XGBoost for gradient boosting modeling,
- and Pickle for saving trained models.

These tools collectively supported the implementation of data-driven modeling and performance assessment.

3) Data Preprocessing

Upon initial examination, it was observed that the *"bmi"* feature contained missing values. These were imputed using the K-Nearest Neighbors (KNN) imputation technique to maintain the integrity of the dataset. Continuous variables were standardized using StandardScaler to ensure uniform scaling across features. Furthermore, categorical variables were encoded into numerical representations suitable for machine learning algorithms. Additional engineered features were introduced by segmenting continuous attributes like age, BMI, and average glucose levels into categorical bins to enhance model interpretability and performance.

4) Dataset Splitting and Imbalance Handling

A significant class imbalance was present in the dataset, with a considerably smaller number of positive stroke cases. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE creates synthetic examples of the minority class, leading to a more balanced training set. The final dataset was split into training and test sets, maintaining an 80:20 ratio, with the oversampling applied exclusively to the training subset to prevent data leakage.

5) Model Selection and Evaluation

Three machine learning models were developed and evaluated:

- **Random Forest (RF):** An ensemble learning method based on bagging decision trees.
- **XGBoost (XGB):** A scalable and efficient implementation of gradient boosting**.**
- **Hybrid Model (XGB_RF):** A novel ensemble that combines XGBoost and Random Forest outputs.

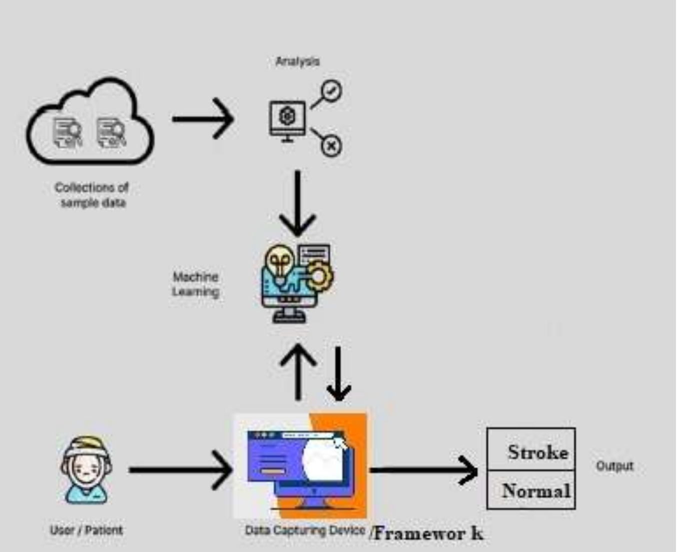
All models were trained on the SMOTE-balanced training dataset. Evaluation metrics included accuracy, F1-score, recall, and ROC-AUC, providing a comprehensive view of predictive performance. Among the three, the XGBoost model achieved the highest accuracy of 95%, outperforming both the Random Forest model (81%) and the hybrid XGB_RF model (84%). These results affirm the superiority of gradient boosting in handling complex patterns within structured health data.

6) Model Persistence and Deployment Readiness

To facilitate future usage in a production environment, the trained models were serialized using the Pickle module. The final XGBoost model was saved as ***"XGBstroke.pkl"*, and the hybrid** model as *"XGB_RFstroke.pkl"*. These persisted models can be loaded into user interface applications or API-based services, enabling real-time predictions and integration into health decision support systems.

4.2 Results

This project confirms that advanced ensemble machine learning approaches, particularly the fusion of XGBoost and Random Forest, provide a powerful and reliable method for stroke risk prediction. The models not only achieve high accuracy but also offer clinical interpretability, making them suitable for integration into healthcare decision support systems.



5.FUTURE ENHANCEMENT & CONCLUSION

Finally, our machine learning-based Stroke Prediction system serves as a useful backup to traditional diagnostic tools in clinical settings for Computer-Assisted Diagnosis (CAD). Although AI models are quite good at predicting the future, the use of Explainable Artificial Intelligence (XAI) techniques addresses the critical issue of decision-making transparency. Our strategy allows doctors to understand and trust the system's outputs by providing interpretable explanations for predictions, such as emphasizing the impact of specific clinical parameters on confidence levels. This not only provides a collaborative environment in which medical specialists can critically evaluate and enhance the model's performance, but it also boosts trust in the decision support system.

Explainable AI is also an important consideration for future work. As predictive models become more complex, it is crucial to ensure that their recommendations are transparent and understandable to clinicians and patients alike. Developing user-friendly interfaces and visualization tools that clearly explain how a prediction was made will help build trust and support real-world adoption.

Lastly, there is a need for broader validation across diverse populations and healthcare environments. By testing these models in different hospitals, regions, and demographic groups, we can ensure their reliability and fairness, minimizing the risk of bias and maximizing their benefit to all patients Furthermore, our diagnosis framework perturbation-based explanation method shows promise for wider applications in a variety of medical domains, highlighting the potential contribution of explainability to the advancement of healthcare AI.

REFERENCES :

[1] Learn About Stroke. Accessed: May 25, 2022. [Online].Available:https://www.worldstroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke

[2] T. Elloker and A. J. Rhoda, “The relationship between social support and participation in stroke: A systematic review,” Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[3] M. Katan and A. Luft, “Global burden of stroke,” Seminar Neurol., vol. 38, no. 2, pp. 208–211, Apr. 2018.

[4] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart, A. Simats, E. Pecharroman, O. Ventura, M. Ribó, D. Vivien, J. C. Sanchez, and J. Montaner, “Blood biomarkers to differentiate ischemic and hemor-rhagic strokes,” Neurology, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021.

[5] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li, “Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,” J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[6] A. Alloubani, A. Saleh, and I. Abdelhafiz, “Hypertension and diabetes mellitus as a predictive risk factors for stroke,” Diabetes Metabolic Syn-drome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

[7] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, “Stroke risk factors, genet-ics, and prevention,” Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[8] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, “Stroke symp-toms and the decision to call

for an ambulance," Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.

[9] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, "Response to symptoms of stroke in the UK: A systematic review," BMC Health Services Res., vol. 10, no. 1, pp. 1–9, Dec. 2010.

[10] L. Gibson and W. Whiteley, "The differential diagnosis of suspected stroke: A systematic review," J. Roy. College Physicians Edinburgh, vol. 43, no. 2, pp. 114–118, Jun. 2013.

[11] N. M. Murray, M. Unberath, G. D. Hager, and F. K. Hui, "Artificial intelligence to diagnose ischemic stroke and identify large vessel occlu-sions: A systematic review," J. NeuroInterventional Surgery, vol. 12, no. 2, pp. 156–164, Feb. 2020.

[12] Y. Zhao, S. Fu, S. J. Bielinski, P. A. Decker, A. M. Chamberlain, V. L. Roger, H. Liu, and N. B. Larson, "Natural language processing and machine learning for identifying incident stroke from electronic health records: Algorithm development and validation," J. Med. Internet Res., vol. 23, no. 3, Mar. 2021, Art. no. e22951.

[13] B. McDermott, A. Elahi, A. Santorelli, M. O'Halloran, J. Avery, and E. Porter, "Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis," Physi-ological Meas., vol. 41, no. 7, Aug. 2020, Art. no. 075010.

[14] A. Bivard, L. Churilov, and M. Parsons, "Artificial intelligence for decision support in acute stroke—Current roles and potential," Nature Rev. Neurol., vol. 16, no. 10, pp. 575–585, Oct. 2020.
[15] W. Wang, M. Kiik, N. Peek, V. Curcin, I. J. Marshall, A. G. Rudd, Y. Wang,A.Douiri, C. D. Wolfe, and B. Bray, "A systematic review of machine learning models for predicting outcomes of stroke with structured data," PLoS ONE, vol. 15, no. 6, Jun. 2020, Art. no. e0234722.

[16] M. S. Sirsat, E. Fermé, and J. Câmara, "Machine learning for brain stroke: A review," J. Stroke Cerebrovascular Diseases, vol. 29, no. 10, Oct. 2020, Art. no. 105162.

[17] A. K. Arslan, C. Colak, and M. E. Sarihan, "Different medical data mining approaches based prediction of ischemic stroke," Comput. Methods Programs Biomed., vol. 130, pp. 87–92, Jul. 2016.

[18] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, "Explainable artificial intelligence model for stroke prediction using EEG signal," Sensors, vol. 22, no. 24, p. 9859, Dec. 2022.

[19] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," Sensors, vol. 22, no. 13, p. 4670, Jun. 2022.

[20] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, "An explainable machine learning pipeline for stroke prediction on imbalanced data," Diagnostics, vol. 12, no. 10, p. 2392, Oct. 2022