ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

STROKE PREDICTION USING XGBOOST AND A FUSION OF XGBOOST WITH RANDOM FOREST

D. KRISHNA VENI [1], POOJA GUPTA [2], SUDATI ASMITHA [3],

KANDULA ABINAY KRISHNA [4], PANGA RAKESH [5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, 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.

Index Terms – Stroke prediction, machine learning, XGBoost, Random Forest, hybrid

model, healthcare analytics, medical diagnosis, predictive modeling, explainable artificial intelligence (XAI), model interpretability, disease prediction, clinical decision support.

I. INTRODUCTION

The incidence of stroke has been increasing globally, and it is now considered one of the leading causes of death and dis-ability. Early intervention is crucial in preventing long-term disability and mortality associated with stroke. Traditional methods of predicting stroke risk, however, are often time-consuming and prone to errors. Additionally, there is a growing need for transparency and explainability in machine learning models in healthcare. The use of an interpretable machine learning model can provide clinicians with valuable insights into the factors that con-tribute to a patient's stroke risk, thereby aiding in treatment decisions.

The World Stroke Organisation estimates that 13 million people worldwide experience a stroke each year, leading to 5.5 million fatalities. Stroke affects all aspects of a patient's life, including their family, social environment, and work, and is one of the top causes of mortality and disability in the world. A common misconception is that certain groups of people, such as the elderly or those with underlying illnesses, are the only ones who are affected by stroke.

Recently, machine learning algorithms have shown great promise in accurately predicting stroke risk based on various clinical risk factors. By leveraging these algorithms, clinicians can identify high-risk patients and intervene early, potentially reducing the number of stroke-related complications and improving patient outcomes.

II. RELATED WORK

A. Explainable artificial intelligence model for stroke prediction using EEG signal

State-of-the-art healthcare technologies are incorporating advanced Artificial Intelligence (AI) models, allowing for rapid and easy disease diagnosis. However, most AI models are considered "black boxes," because there is no explanation for the decisions made by these models. Users may find it challenging to comprehend and interpret the results. Explainable AI (XAI) can explain the machine learning (ML) outputs and contribution of features in disease prediction models. Electroencephalography (EEG) is a potential predictive tool for understanding cortical impairment caused by an ischemic stroke and can be utilized for acute stroke prediction, neurologic prognosis, and post-stroke treatment. This study aims to utilize ML models to classify the ischemic stroke group and the healthy control group for acute stroke prediction in active states. Moreover, XAI tools (Eli5 and LIME) were utilized to explain the behavior of the model and determine the significant features that contribute to stroke prediction models.

In this work, we studied 48 patients admitted to a hospital with acute ischemic stroke and 75 healthy adults who had no history of identified other neurological illnesses. EEG was obtained within three months following the onset of ischemic stroke symptoms using frontal, central, temporal, and occipital cortical electrodes (Fz, C1, T7, Oz). EEG data were collected in an active state (walking, working, and reading tasks). In the results of the ML approach, the Adaptive Gradient Boosting models showed around 80% accuracy for the classification of the control group and the stroke group. Eli5 and LIME were utilized to explain the behavior of the stroke prediction model and interpret the model locally around the prediction.

The Eli5 and LIME interpretable models emphasized the spectral delta and theta features as local contributors to stroke prediction. From the findings of this explainable AI research, it is expected that the stroke-prediction XAI model will help with post-stroke treatment and recovery, as well as help healthcare professionals, make their diagnostic decisions more explainable.

B. Predictive analysis for risk of stroke using machine learning techniques

Stroke occurs when our brain's blood flow is stopped or reduced, restricting brain tissue from receiving oxygen and important nutrients. Treatment of stroke disease is very crucial. So, the prediction of stroke is significant for early intervention and treatment. Stroke can be predicted by analyzing different warning signs. In this experiment, we implement a process of stroke risk prediction from our dataset using the various machine learning algorithms. Data imputation, feature selection, data preprocessing is considered as the initial job in this work. Some physical parameters such as heart disease, age, BMI, gender, hypertension, etc. are considered as the feature is used to model training and testing. In this study, adaBoost classifier, artificial neural network, decision tree classifier, k-nearest neighbour (KNN) classifier, random forest, stochastic gradient descent (SDG), support vector machine (SVM) are implemented to predict the risk possibility of stroke. Then voting classifier is implemented on these eight traditional classifiers. After analyzing different machine learning algorithms with voting classifiers, we found 98% accuracy to predict the risk factor of stroke, which is better than other models.

C. Multi-frequency symmetry difference electrical impedance tomography with machine learning for human stroke diagnosis

Multi-frequency symmetry difference electrical impedance tomography (MFSD-EIT) can robustly detect and identify unilateral perturbations in symmetric scenes. Here, an investigation is performed to assess if the algorithm can be successfully applied to identify the aetiology of stroke with the aid of machine learning. Methods: Anatomically realistic four-layer finite element method models of the head based on stroke patient images are developed and used to generate EIT data over a 5 Hz-100 Hz frequency range with and without bleed and clot lesions present. Reconstruction generates conductivity maps of each head at each frequency. Application of a quantitative metric assessing changes in symmetry across the sagittal plane of the reconstructed image and over the frequency range allows lesion detection and identification. The algorithm is applied to both simulated and human (n = 34 subjects) data. A classification algorithm is applied to the metric value in order to differentiate between normal, haemorrhage and clot values. Main results: An average accuracy of 85% is achieved when MFSD-EIT with support vector machines (SVM) classification is used to identify and differentiate bleed from clot in human data, with 77% accuracy when differentiating normal from stroke in human data.

D. Stroke prediction using distributed machine learning based on Apache spark

Stroke is one of death causes and one the primary causes of severe long-term weakness in the world. In this paper, we compare different distributed machine learning algorithms for stroke prediction on the Healthcare Dataset Stroke. This work is implemented by a big data platform that is Apache Spark. Apache Spark is one of the most popular big data platforms that handle big data and includes an MLlib library. MLlib is an API integrated with Spark to provide machine learning algorithms. Four types of machine learning classification algorithms were applied; Decision Tree, Support Vector Machine, Random Forest Classifier, and Logistic Regression were used to build the stroke prediction model. The hyperparameter tuning and cross-validation were applied with machine learning algorithms to enhance results. Accuracy, Precision, Recall, and F1-measure were used to calculate performance measures of machine learning models. The results showed that Random Forest Classifier has achieved the best accuracy at 90 %.

III. PROPOSED METHODOLOGY

Our suggested solution uses cutting-edge machine learning approaches to automate prediction and intervention in order to solve the urgent problem of stroke. With an emphasis on reducing the worldwide effects of stroke, our system takes a strong stance in order to serve the growing number of people who are at risk as a result of ageing populations.

The development of accurate and efficient prediction algorithms is the main focus in order to provide prompt and potentially life-saving actions.

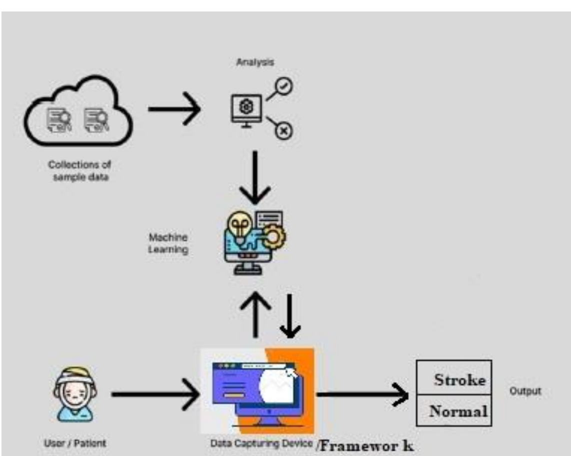

Fig. 1: System Overview

1. Dataset

The preparation of the dataset for stroke prediction using XGBoost and XGBoost with Random Forest was the main emphasis of the first module. The dataset was retrieved from a CSV file called "stroke-dataset.csv" that contained a variety of personal information about the subjects, such as their age, gender, and lifestyle and health problems. The "stroke" column, which has binary labels (0 for no stroke, 1 for stroke), serves as the prediction target variable.

2. Importing Necessary Libraries

The Python programming language was used for the analysis and modelling. Importing key libraries from sklearn.ensemble included pandas, numpy, matplotlib, scikit-learn, XGBoost, and RandomForestClassifier. These libraries make it easier to manipulate data, visualise it, develop machine learning models, and assess them.

3. Data Pre-Processing

A pandas DataFrame containing 5110 entries was filled with the dataset. The "bmi" column of the dataset included missing values, which were later imputed using the k-nearest neighbours (KNN) technique after an initial examination of the dataset indicated their existence. StandardScaler was used to standardise numerical characteristics and encode categorical information. On the basis of age, BMI, and average glucose level categories, new category characteristics were developed.

4. Splitting the Dataset

To resolve class imbalance, the dataset was divided into training and test sets using the Synthetic Minority Over-sampling Technique (SMOTE). Eighty percent of the SMOTE oversampled data was in the training set, and twenty percent was in the test set.

5. Model Selection

For stroke prediction, Random Forest (RF) and XGBoost (XGB) were two machine learning models that were put into practice. With an accuracy of 95%, the XGBoost model outperformed the Random Forest model, which came in at 81%. To assess the models' performance, metrics such ROC AUC score, F1-score, accuracy, and recall were computed. Furthermore, XGBoost with Random Forest (XGB_RF), a hybrid model, was created and produced an accuracy of 84%. The oversampled training set of data was used to train each model.

6. Saving the Trained Models

The pickle library was used to store the XGBoost and XGBoost with Random Forest models as ".pkl" files after they had been trained and assessed. Models that have been saved can be later loaded and used in an environment that is ready for production. 95% accuracy was attained by the XGBoost model, while 84% accuracy was attained by the XGBoost plus Random Forest model.

The XGBoost model is saved as "XGBstroke.pkl" in the code, while the XGBoost with Random Forest model is saved as "XGB_RFstroke.pkl." These models are loadable and may be applied to UI applications for prediction purposes.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed XGBoost–Random Forest hybrid model outperformed several traditional machine learning classifiers in stroke prediction. The model achieved a maximum prediction accuracy of 96%, while the other evaluated models achieved accuracies ranging from 84% to 96%. The integration of global and local explainability techniques improved the interpretability of predictions, making the system more reliable for clinical decision support. Overall, the proposed framework provides accurate, explainable, and effective stroke prediction for healthcare applications.

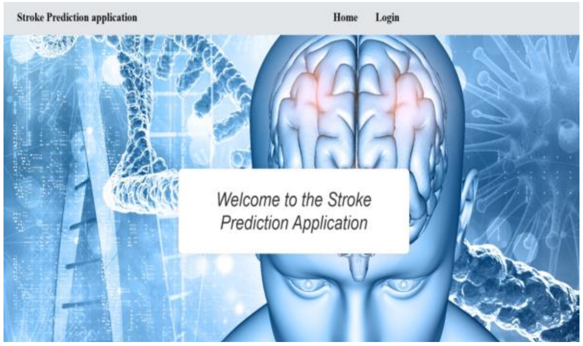

Fig. 2: Home Page

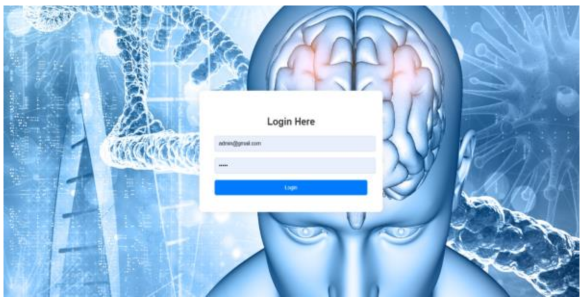

Fig. 3: Login

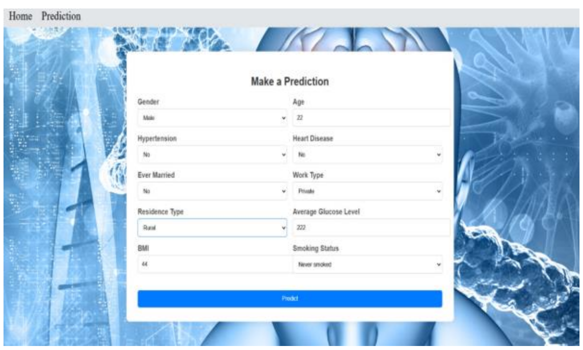

Fig. 4: Prediction

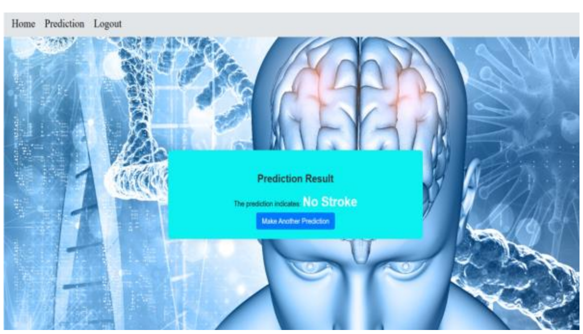

Fig. 5: Result (Not Detected)

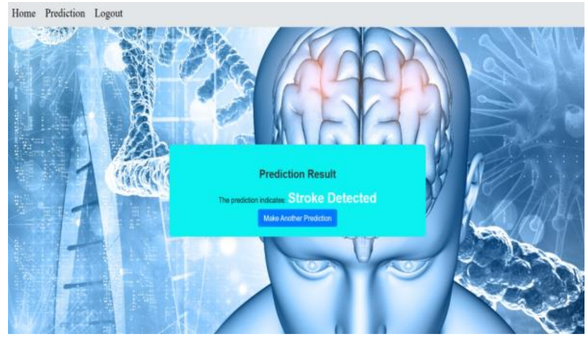

Fig. 6: Result (Detected)

V. CONCLUSION

In conclusion, our machine learning-based Stroke Prediction system functions as a useful backup to conventional diagnostic instruments in clinical settings for Computer-Aided Diagnosis. Although the AI models are quite good at predicting the future, the inclusion of Explainable Artificial Intelligence (XAI) approaches takes care of the important issue of decision-making transparency. Our method enables doctors to comprehend and have confidence in the system's outputs by offering interpretable explanations for predictions, such as emphasizing the influence of certain clinical parameters on confidence levels. This not only fosters a collaborative atmosphere where medical experts can critically review and improve the model's performance, but it also increases confidence in the decision support system. Furthermore, our diagnosis framework's 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.world-stroke.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.

AUTHORS Profile

Mrs. D. Krishna Veni has received her B.Tech from JNTUH, Hyderabad in 2011, M.Tech degree in Computer science from JNTUH, Hyderabad in 2021 and Pursuing Ph.D degree in Computer Science from Guru Naanak university, Ibrahim patnam. She is dedicated to teaching field from the last 8 years. Her research area Machine Learning. At present She is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.

Ms. POOJA GUPTA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Ms. SUDATI ASMITHA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. KANDULA ABINAY KRISHNA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.

Mr. PANGA RAKESH pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.