2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS) | 979-8-3315-2963-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICUIS64676.2024.10866013

A Novel Hybrid Approach Integrating Machine Learning and ANN for Stroke Risk Assessment

Dr.Nagagopiraju Vullam
Professor, Department of CSE
Chalapathi Institute of Engineering and Technology,Guntur,A.P,India
gopi.raju524@gmail.com

S.V.V.D.Jagadeesh
Sr.Assistant Professor, Department of AI&DS
LakiReddy Bali Reddy College of Engineering,Mylavaram
jagadeesh.53533@gmail.com

K.Siva Pavani
Assistant Professor,Department of AI&ML
Aditya University
Surampalem,India
kottapallisivapavani9@gmail.com

Maddala Janakidevi
Assistant professor, Department of CSE
SRKR Engineering College
Bhimavaram,A.P,India
janakidevimaddala9@gmail.com

Dr.A.Lakshmanarao
Assistant Professor, Department of IT
Aditya University
Surampalem,India
a.lakshmanarao@adityauniversity.in

Y S N Chandreswari
Associate Professor, Department of MCA
Bonam Venkata Chalamayya institute of technology and science, Batlapalem
tsnchchandreswari.bvts@bvcgroup.in

Abstract- **Brain stroke is a life-threatening condition that needs to be detected early and correct precision to improve patient wellness. This paper presents a method for predicting stroke using ML and ANN using a publicly available data from Kaggle. The dataset has class imbalance. This can be solved using SMOTE and ADASYN to guarantee a balanced class distribution. After pre-processing, six traditional ML models including Logistic Regression, DTC, Random Forest, KNN, SVM and Gradient Boosting were used to predict the risk of stroke. To further increase the accuracy of the forecast. Later, ensemble models also applied to further enhance performance. Finally, ANN was applied to capture complex patterns in the data. The results shown that the ensemble model significantly improves the prediction performance compared to the individual ML models. Later a hybrid ML and DL ANN is proposed. The proposed hybrid model combining ML best model and ANN technique improved the accuracy in predicting stroke and provide a more reliable system for early detection of stroke. It is observed that the hybrid method will outperform both individual ML models and ensemble models.**

.

Keywords— Brain Stroke, Machine Learning, Deep Learning, ANN, Ensemble.

I. Introduction

Brain stroke is one of the most prevalent and potentially fatal neurological conditions worldwide. Disruption in blood flow to the brain leads to brain stroke. As a result, brain tissue is starved of oxygen. This rupture may be caused by a blockage in the blood vessels that supply the brain . This causes blood to flow within or around the brain (ischemic stroke). Ischemic strokes account for nearly 85% of cases, and it is usually caused by the formation of a clot or narrowing of the blood vessel. High blood pressure or an aneurysm typically causes it. Hemorrhagic strokes, while less common, tend to be more severe and occur when blood vessels are weak, often due to high blood pressure or aneurysms. It ruptures and causes internal bleeding. Both types can cause permanent brain damage, permanent disability, or death. If not diagnosed and treated in a timely manner, it can be dangerous. It emphasizes the need for early detection and preventative care.

The challenge of early detection possible by identifying the risk factors and warning signs before a stroke symptom. Factors such as age, high BP, diabetes, heart disease and lifestyle habits can affect this situation. However, individual variation and the complex interaction of these factors often lead to challenging clinical prognoses. Therefore, predictive models using advanced computational techniques can help identify individuals at high risk for stroke before they begin to show serious symptoms. ML and DL algorithms offer promising solutions in disease detection by analysing large amounts of patient data, identify patterns, and make accurate predictions based on patterns.

In this work, an efficient prediction model for stroke detection was developed using ML and neural network-based approaches. A dataset from Kaggle with several variables like age, gender, high BP, and heart problems are used in the work. Due to class imbalance, some balancing techniques like SMOTE and ADASYN were applied. After balancing the data, several ML algorithms, including logistic regression, are used. DT, RF, SVM, and gradient boosting are applied. Later, an ensemble approach is used to combine the strengths of these algorithms. Next, a deep learning model ANN is also trained to further improve the prediction accuracy. Finally a hybrid ANN and best ML model applied to further increase accuracy. This paper aims to increase the reliability and accuracy of stroke risk prediction with ML and DL techniques. The algorithms used in this work are discussed below.

A. Logistic Regression

It is a general statistical model applying for binary classification, for a particular type of yes or no predictions. However, it can be also be applied for multi-class classification. In the simplest use It is one of the two most frequently used ML algorithms for classifying binary data translated to 0 or 1 as input..

B. Deision Tree

This is one of ML supervised learning technique with tree notation. Each node represents a feature and each option is represented by a connection that connects nodes. Each child node represents the results.

C. KNN

The general concept used to describe the family of classification and regression algorithms known as KNN. It is a instance-based learning. This is also sometimes called lazy learning. It mainly depends on closest neighbours to group data into categories.

D. KNN

The general concept used to describe the family of classification and regression algorithms known as KNN. It is a instance-based learning. This is also sometimes called lazy learning. It mainly depends on closest neighbours to group data into categories.

E. SVC

It is a model that is popular for its efficiency and versatility. And it is used in many fields such as image recognition, text classification. and medical diagnosis.

F. Random Forest

It uses a collection of decision trees to classify data into categories. It can be used with both regression, classification tasks.

G. Gradient Boosting

It combines multiple weak models to create a unique and more accurate prediction model. It uses the power of weak learners and makes it stronger.

H. Rtificial Neural Networks

In this work, ANN used to predict brain stroke. It takes advantage of its ability to capture complex forms in the data comprising the input pool of features. Multiple hidden layers using the ReLU update function and output layers uses the sigmoid function for binary classification. The network is trained via backpropagation using an optimization algorithm, reducing prediction errors in iterations. ANN model increases prediction accuracy and complements traditional ML models in a hybrid approach.

II. LITERATURE SURVEY

J. Parvathi et al. [1] focused on early detection and prevention of stroke. They divided stroke into ischemic and haemorrhagic types and emphasized the importance of immediate treatment, such as giving clot-busting drugs or coagulating drugs. Their work highlights the importance of detecting early warning signs of stroke to reduce stroke severity. The study aimed to predict the probability of brain damage at an early stage with the help of a machine learning algorithm. They applied several classification models, including KNN, Logistic Regression, RF, EGB. Ahmad Hassan et al. [2] discussed challenges in stroke prediction by dealing with unbalanced and missing data. They used three imputation techniques to deal with the imbalance. They validated several advanced models using k-fold cross-validation on unbalanced and balanced datasets. The main predictors identified were age, BMI, glycose levels, heart disease, and high blood pressure and marital status. Dense Stacking Ensemble model used as a meta-classifier. The authors explored the integration of modern technology and health services in the treatment and prevention of stroke, which is an important public health problem. They conducted an extensive literature review on the use of ML, DL models in stroke prediction. By synthesizing existing research, identifying trends, best practices and gaps in research, they gain valuable insights for further research. Their work has led to the development of robust prediction models.

T. Kavitha et al. [4] studied cerebrovascular disease, which is characterized by insufficient blood flow to the brain which can lead to cell death. They identified that 85 percent of strokes can be detected early with proper techniques. They mainly used ML techniques such as LGR, DTC, KNN, RF and SVM. Their findings showed that SVM was accurate to 94.6% in stroke prediction. It further enhanced by RF and XGB algorithms, which were enhanced with feature selection techniques such as Chi-Square and Information Gain. The authors in [5] examined the increased incidence of stroke and emphasized the critical need for early forecasts to minimize impacts. They studied on the effectiveness of ML algorithms in predicting disease by suing a well-known dataset that included various factors that affects the risk of brain injury. Harshit Kumar Singhai et al.,[6] explored the complexity of the brain and emphasize the urgency of timely treatment for stroke, which is the leading cause of death worldwide. They noted that early identification of risk factors for brain injury would be beneficial and can greatly reduce the death rate. They highlighted the role of ML algorithms in this process. They proposed a model that uses the "Stroke Prediction Dataset" from Kaggle. With ML approaches, they achieved good results. Ritesh Kumari et al. [7] explored the serious problem of brain stroke with the percentage of the population affected by this in every year. They emphasized the importance of early intervention to save people's lives and discuss the role of ML as a tool for prevention based on symptoms, lifestyle, and medical history. Various ML classifiers including NN, SVM RF etc. are used to classify patients at risk for stroke. The performance of these models was compared to find the best performance. The authors also used LIME and SHAP to provide insights into the policymaking to identify best-performing method.

The authors of [8] discussed the severe consequences of strokes, which occur when a part of the brain fails to receive adequate blood flow, often due to arterial blockages or bleeding. They explained that prolonged oxygen deprivation leads to brain cell death, resulting in irreversible damage and loss of functionality. Their study highlighted the critical importance of timely restoration of blood flow to minimize damage, emphasizing that stroke treatment is highly time-sensitive. The authors also noted that symptoms vary depending on the specific brain regions affected, as different regions control distinct functions. Their research focused on predicting strokes using machine learning (ML) models. The authors of [9] emphasized the essential role of blood vessels in delivering oxygen and nutrients to the brain and highlighted cerebrovascular disease as one of the most severe global health threats. They applied ML models, including K-Nearest Neighbors (KNN) and Random Forest (RF), achieving promising results in stroke prediction. Sachin Sharma et al. [10] focused on developing a model to identify individuals at risk of stroke based on demographic, clinical, and lifestyle factors. Their study utilized a comprehensive dataset comprising a large cohort of stroke and non-stroke patients, including detailed information on various risk factors. ML algorithms such as Logistic Regression, XGBoost, and Naïve Bayes (NB) were applied and evaluated on this dataset. Additionally, feature selection techniques were employed to identify key risk factors. The results indicated that the XGBoost model demonstrated the highest accuracy in predicting stroke risk. In [11], stroke prediction was studied using various ML models, including Naïve Bayes (NB), Logistic Regression, Decision Tree Classifier (DTC), KNN, AdaBoost, and XGBoost. A comparative analysis of these algorithms was performed to evaluate their efficiency in stroke prediction. The findings revealed that AdaBoost and XGBoost achieved the highest probability scores, demonstrating superior performance for stroke prediction.

III. Proposed Method

The proposed method was depicted in figure 1. The proposed method for brain stroke prediction consists of several steps. Initially, a dataset is collected from Kaggle with features like age, gender, high BP etc. Later, pre-processing steps performed on dataset. The dataset is imbalanced. Handling class imbalance is important step before applying ML models. For that, two methods SMOTE and ADASYN are employed to generate synthetic instances of the minority class (stroke cases), ensuring a balanced representation of both classes. After balancing, several ML models like Logistic Regression, DTC, Random Forest, KNN, SVM and Gradient Boosting applied as foundation for model creation. The results with these ML methods are tabulated for future comparisons. After identifying best ML models, an ensemble with best techniques is implemented to enhance performance of brain stroke prediction. Later, a deep learning ANN also applied to further enhance performance of model. After training the ANN model, it is combined with the best ML models to create hybrid approach. After applying all these techniques, the best model for brain stroke prediction was identified and used for predictions.

IV. Experimental Results

A. Data Collection

The data collection process collected health-related information from kaggle. It includes information about age, gender, marital status and type of residence et. Along with these, this data also includes medical history of high blood pressure, heart disease, and stroke. It was extracted from patient records and public health banks. Lifestyle factors such as smoking, type of occupation, and average blood sugar levels It will be collected through in-person surveys or wearable devices. while clinical data such as body mass index (BMI) and blood sugar levels will also covered. This data collection guarantees a rich and robust data set for analyzing brain stroke risk factors. It consists of 5,110 samples.

B. Data Preprocessing

In the next step, data is checked for errors and inconsistency issues. During this phase, the only modification required is to handle missing or incorrect values. For the feature "smoking", there are some filed values with "unknown" values. These are replaced with most frequent value in that column "never smoked". This step is done to keep the data set consistent and complete. There are no other inconsistencies found in the dataset. Of course there are categorical variables are these, but those can be handled by encoding easily. Now, this data is ready for the modelling phase.

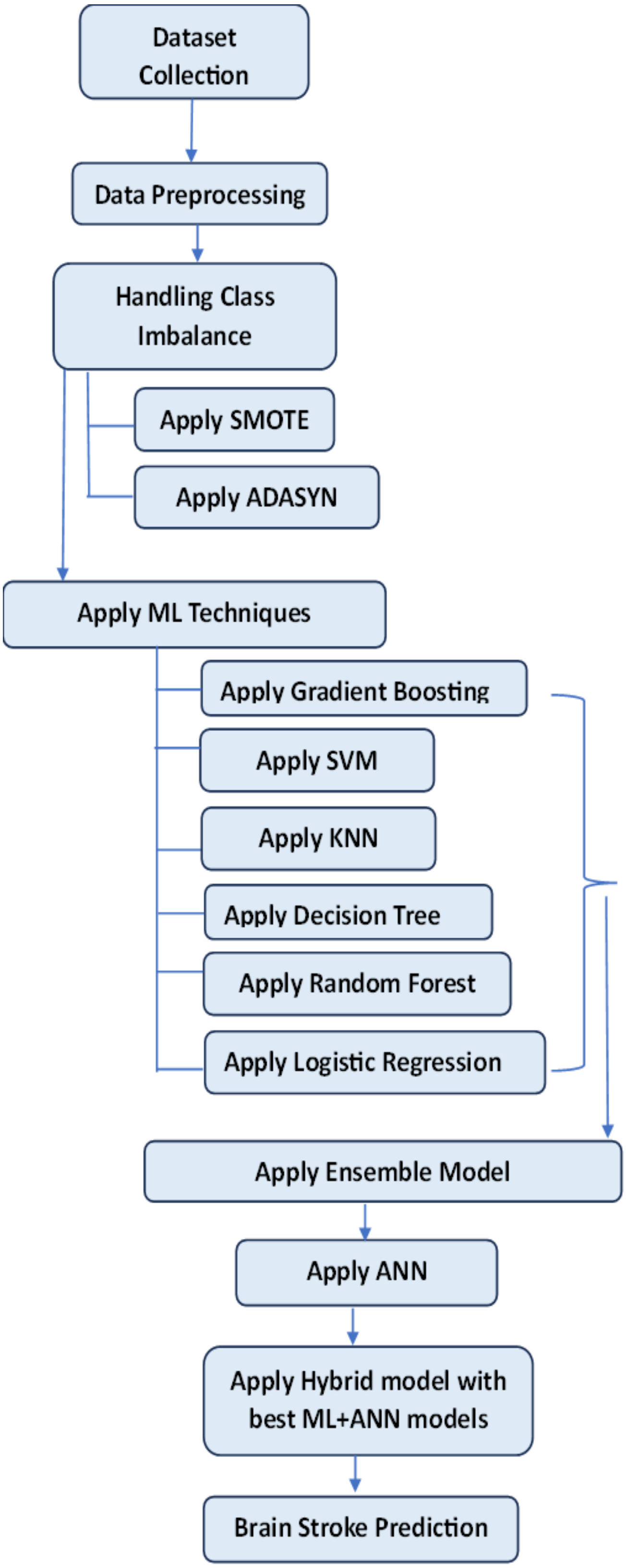

Fig. 1. Proposed Methodology

979-8-3315-2963-5/24/$31.00 ©2024 IEEE

C. Handling Class Imbalance

This dataset shows a large class inequality. There were 4,861 samples labelled as "No stroke" and only 249 samples were labelled "no stroke”. So, it is compulsory to handle class imbalance, otherwise ML models may not perform well. Therefore, two techniques namely SMOTE and ADASYN applied. These methods create a synthetic sample of subclasses (stroke cases) and guarantee a balanced database. This step is critical to avoid biasing the model towards the majority and to increase the accuracy of prediction.

D. Applying ML Techniques

After dealing with class imbalance, ML models applied on the dataset. Six traditional ML models applied. Each technique is trained on train data and tested with testing data. Key evaluation indicators such as precision, precision, recall, F1 score were calculated to evaluate the effectiveness of each model. Table I and figure 2 depicts the results with SMOTE and ADASYN datasets.

For SMOTE dataset, the Logistic Regression model achieved an accuracy of 78% with a precision of 76%, a recall of 78%, and an F1 score of 77%. DTC performed well, with an accuracy of 82%., 81% precision, 83% recall, and 82% F1 score. Random Forest achieved the highest accuracy of 89% with 90% precision, 88% recall, and 88% F1 score, indicating superior performance. KNN provided balanced results with 79% accuracy, while SVM had lower accuracy of 74% accuracy. GBC given 77% accuracy. Random Forest emerged as the best performing model in this comparison.

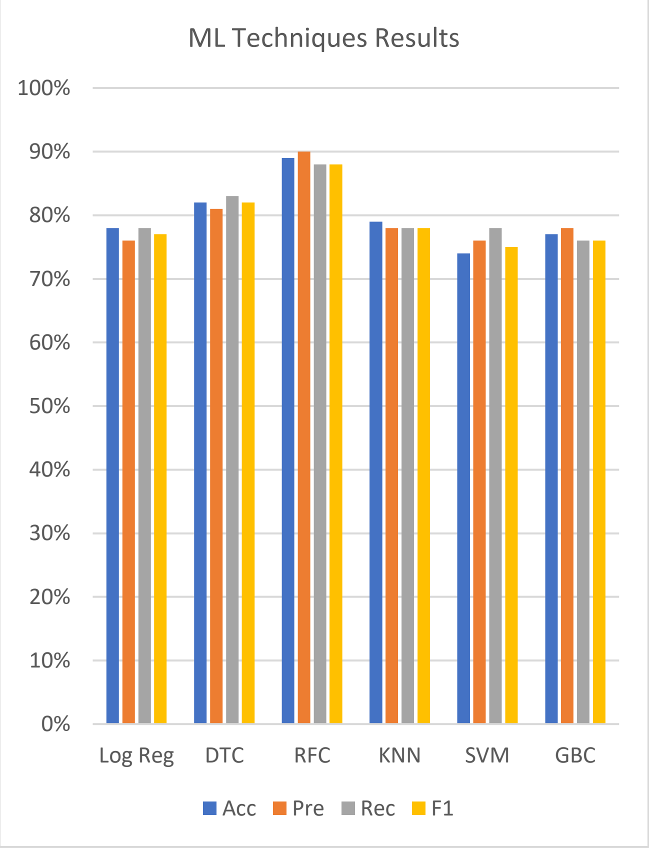

Fig. 2. ML Techniques Results with SMOTE dataset

TABLE I. RESULTS WITH ML TECHNIQUES AND SMOTE

Model	Acc	Pre	Rec	F1
Log Reg	78%	76%	78%	77%
DTC	82%	81%	83%	82%
RFC	89%	90%	88%	88%
KNN	79%	78%	78%	78%
SVM	74%	76%	78%	75%
GBC	77%	78%	76%	76%

Table II and figure 3 shows results of ML models with ADASYN dataset. RFC achieved the highest accuracy of 90%, with impressive precision (92%), recall (87%), and an F1-score of 90%. Similarly, DTC and KNN exhibited an accuracy of 80%, with DTC achieving a balanced precision and recall of 80%. The KNN given a precision of 83% and logistic regression given 79% accuracy. Overall, the ADASYN technique provided a balanced class distribution.

TABLE II. RESULTS WITH ML TECHNIQUES AND ADASYN

Model	Acc	Pre	Rec	F1
Log Reg	79%	79%	78%	78%
DTC	80%	80%	80%	80%
RFC	90%	92%	87%	90%
KNN	80%	83%	78%	81%
SVM	75%	74%	76%	74%
GBC	75%	75%	75%	75%

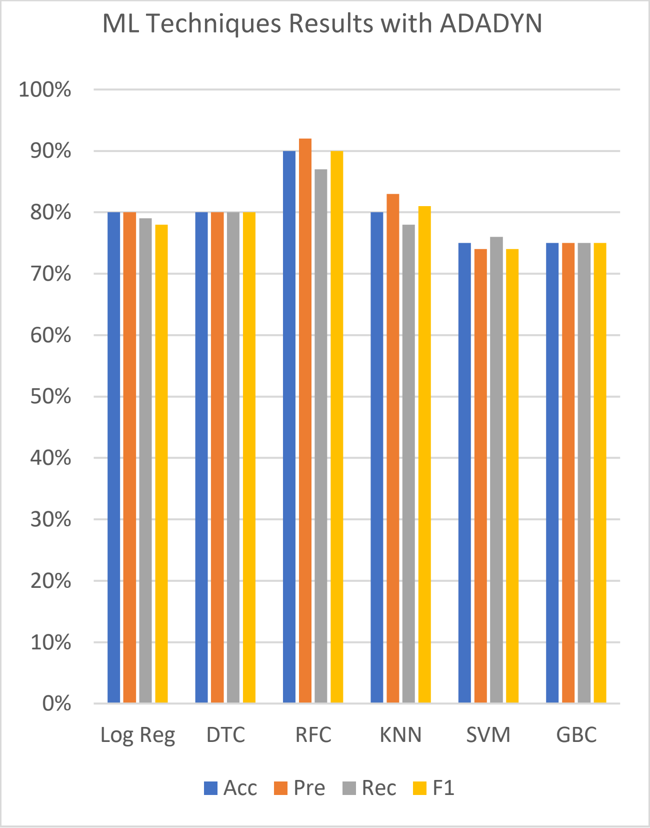

Fig. 3. ML Techniques Results with ADASYN dataset

E. *Applying Ensemble Model*

In this step, an ensemble cascading approach using Random Forest (RF) as a metaclassifier and best three ML models in previous step to further enhance the prediction performance. DTC, RF, and KNN serve as basic learners. As ADASYN dataset given high accuracy, it is used in ensemble approach. By combining the predictions of these basic models with the use of RF models as meta-learners for final decision-making. The joint approach can take the advantage of the strengths of each model. This stacking unit achieved an incredible 93% accuracy and 92% recall surpassing the performance of individual models.

F. *Applying ANN*

The ANN was integrated into the stroke prediction model to capture complex, non-linear relationships within the dataset that traditional ML models might miss. While the ML models effectively handle structured patterns and provide robust baseline predictions, ANNs excel in identifying subtle interactions among features due to their layered architecture and activation functions.

In this step, Artificial neural networks (ANN) have been used to further improve stroke prediction. The ANN model consists of three layers, with 40, 20, and 10 neurons in the hidden layer, respectively. It has a sigmoid function in the output layer and ReLU in hidden layers. ANN achieved an accuracy of 94 %, demonstrating its ability to better capture complex patterns and subtle relationships in the data. Figure 4 shows epoch wise accuracy for this model.

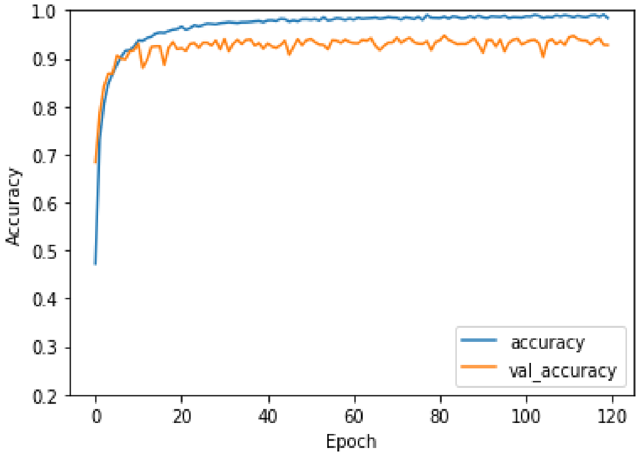

Fig. 4. Epochwise accuracy with ANN for train and test samples

G. *Applying Hybrid model of ML and ANN model*

Finally, a hybrid model approach is used to combine both ML and ANN models. This fusion model combines the power of best ML classifiers RF and ANN to take advantage of both traditional ML and DL capabilities to handle complex data. In this process, the predictions from the Random Forest model and the ANN were combined, allowing the hybrid model has advantage from the strengths of both approaches. Hybrid model shown the best performance with accuracy of 95.4%, which greatly improved the accuracy of stroke prediction compared to ML alone and ANN alone models.

V. CONCLUSION

This research study successfully demonstrates the effectiveness of various ML techniques including traditional ML methods and DL to predict the risk of stroke The results indicate that random forest achieved the highest accuracy of 90% with ADASYN, while the ensemble model combining best three ML classifiers achieved an impressive accuracy of 93%. Moreover, the application of ANN has significantly improved its predictive ability with an accuracy of 94%. Combining the strengths of random forest and ANN, the hybrid model outperformed each approach with an accuracy of 95.4%. This combined method not only improves the accuracy of stroke prediction, but it also highlights its potential tool for healthcare professionals in early diagnosis and intervention. Future work could explore the integration of real-time patient monitoring and personalized risk factors to further enhance the model's predictive capability and adaptability in clinical settings.

REFERENCES

[1] Jarapala Parvathi, “Machine Learning based Brain Stroke Prediction using Light Gradient Boosting Machine Algorithm”, Int J Intell Syst Appl Eng, vol. 12, no. 4, Aug. 2024.

[2] A. Hassan et al., “Predictive modelling and identification of key risk factors for stroke using machine learning,” Scientific Reports, vol. 14, no. 1. Springer Science and Business Media LLC, May 20, 2024.

[3] R. M. Mandhare et al., "Analysis of AI Driven Brain Stroke Prediction Using Machine Learning and Deep Learning," 2024 International Conference on Communication, Computer Sciences and Engineering , Gautam Buddha Nagar, India, 2024.

[4] T. Kavitha et al.,"Enhanced Machine Learning Algorithm to Predict Brain Stroke," International Conference on Inventive Computation Technologies, Lalitpur, Nepal, 2024.

[5] V. Jain et al., "Performance Enhancement of Machine Learning Algorithms for Predicting Stroke," 2024 7th International Conference on Circuit Power and Computing Technologies, Kollam, India, 2024.

[6] H. K. Singhai et al., "An Experimental Analysis of Brain Stroke Prediction Using Machine Learning Algorithms," International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation, Gwalior, India, 2024.

[7] R. Kumari et al., "Interpretation and Analysis of Machine Learning Models for Brain Stroke Prediction," International Conference on Information Systems and Computer Networks, Mathura, India, 2023.

[8] N. Devarakonda et al., "Brain Stroke Prediction Using Machine Learning Techniques," 2023 Fifth International Conference on Electrical, Computer and Communication Technologies , Erode, India, 2023.

[9] T. N. Deepthi et al., "Prediction of Brain Stroke in Human Beings using Machine Learning," International Conference on Electronics and Renewable Systems , Tuticorin, India, 2023.

[10] S. Sharma, "Stroke Prediction Using XGB Classifier, Logistic Regression, GaussianNB and BernaulliNB Classifier," International Conference on Circuit Power and Computing Technologies, Kollam, India, 2023.

[11] S. Gupta et al., "Stroke Prediction using Machine Learning Methods," 2022 12th International Conference on Cloud Computing, Data Science & Engineering, Noida, India, 2022.

[12] https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset/data.