2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December 2023, Cox's Bazar, Bangladesh

2023 26th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-5901-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCIT60459.2023.10440991

Development of an Intelligent System for Brain Stroke Prediction using Ensemble Feature Selection and Machine Learning Technique

Forhad Uddin Ahmed[1], Fahamida Hossain Mahi[2], Sayma Alam Suha[3], Muhammad Nazrul Islam[4],
[1,2,3]Department of CSE, Bangladesh Army International University of Science and Technology (BAIUST), Cumilla, Bangladesh
[4]Department of CSE, Military Institute of Science and Technology (MIST), Dhaka, Bangladesh
Email- forhad.uddin892@gmail.com, fahamidamahi@gmail.com, suha.mist@gmail.com, nazrul@cse.mist.ac.bd

***Abstract*—Brain stroke is a serious condition in human body, in which cerebral artery ruptures causing damage to the region and preventing oxygen and other substances from reaching the brain. The risk of brain stroke can be reduced dramatically by early detection through the analysis of various stroke warning indicators in the patient's body. In such cases, predictive analytic of brain stroke through an ensemble feature selection and machine learning using patients' various clinical symptoms can be an excellent solution. The goal of this research is to develop a predictive intelligent system that would estimate the possibility of occurring brain stroke using patient symptom information applying machine learning techniques. To attain the objective, the relevant patient dataset has been collected, analyzed and visualized. The dataset has undergone rigorous preprocessing using multiple data exploration techniques. Five different feature selection method are followed by an ensemble methodology with the maximum voting approach to extract the most important feature and generate a dataset with reduced attributes. Following that, ten types of machine learning classifiers are trained, examined and put to the test to predict the likelihood of brain stroke. Each of the classification models are evaluated using multiple performance parameters. The algorithm that handled this challenge the best is Logistic regression model, which has provided an accuracy of almost 82.7% employing the dataset with reduced clinical features for stroke prediction.**

Keywords*— *Brain Stroke, Machine Learning, classification, neural network.

I. Introduction

A fatal and serious illness is brain stroke which can manifest itself in both hemorrhagic and thrombotic forms, where following a stroke, inflammation causes severe neurological damage in patient [1]. Brain stroke is linked to physical abnormalities such as inattention, impaired coordination, hearing, and communication difficulties, as well as difficulty in arranging objects. Brain strokes, caused by blood clots or hemorrhages, can be detected early through CT and MRI, reducing fatalities and aiding in patient management through early detection of associated symptoms.

Machine learning can significantly predict brain stroke by identifying patterns in large medical data, which humans struggle to discern. By examining patient histories, genetic information, digital health records, and clinical imaging, machine learning algorithms can identify risk factors, symptoms, and potential outcomes associated with various human anomalies, including brain stroke [2]. These algorithms may use this information to create prediction models that help doctors spot stroke suspects before patients have symptoms. Additionally, clinicians may immediately notice people at a high chance of having a stroke and take urgent action to prevent or cure it by using a machine learning to evaluate patient data in real-time. Overall, machine learning has the potential to be a potent weapon in the battle against stroke, allowing medical professionals to quickly identify individuals take precautions to reduce their risk and improve patient outcomes for those who are at high risk.

Therefore, examining the effectiveness of machine learning classifier models for brain stroke estimation is the aim of this research. To accomplish this objective, the study uses a clinical based data-set. To make the dataset more approachable for the machine literacy model, it is initially sanctified and data is pre-processed. Prior to utilizing for prediction, the dataset is checked for null values. In order to convert categorical values into numerical representations, encoding is also utilized. Applies five different feature selection strategies after data preprocessing, each of which leverages the datasets to extract the significant attributes. The results from different feature selection techniques are aggregated through ensemble majority voting to explore the reduced set of features. Data preparation is followed by dividing with and without the dataset with reduced features into train and test data. Following that, to make the predictive analysis a number of machine learning models are used which includes classic, ensemble and neural network based machine learning models. The accuracy, precision, recall and F1-score of each of these algorithms is determined. The study pinpoints the model that performs the best in predicting brain strokes. The study's core contributions are: assembling the dataset after gathering entries, cleaning and preprocessing the data; train, test and Feature selection and evaluating multiple machine learning classification models to find out the model providing the best performance in prediction.

The remaining components of this research endeavor are

979-8-3503-5901-5/23/$31.00 ©2023 IEEE

organized as follows: The literature review is covered in Section II; Section III details the methodology; Results Analysis is examined in Section IV; Section V includes the discussion and conclusion.

II. Literature Review

A few researchers have recently concentrated on predicting brain strokes in patients using machine learning. For example, V.Bandi et al. [3] worked on several prediction research done by machine learning algorithm to predict the stroke,like evaluating danger factors was relevant to different kinds of strokes. Identifying the potential causes of a variety of strokes is difficult given current information. P.Govindarajan et al. [4] proposes a stammer using text mining and machine learning techniques to categorize strokes using common and distinctive characteristics from incident files. Biswas et al. [5] made comparisons between various ML classifiers for stroke prediction.Eleven classifiers that are used in this investigation. T.I.Shoily et al. [6] focused on the diagnosis of the stroke disease. To identify the potential kind of stroke may happen or has already happened,they looked at a person's medical records and physical condition using machine learning algorithms. M.S.Azam et al. [7] evaluates machine learning algorithms' performance in predicting stroke risk, examining dataset components and assessing their effectiveness in this context. Dritsas et al. [8] focused on utilizing a machine learning techniques to forecast the chance of having a stroke. Y.Wu et al. [9] proposed machine learning models for stroke prediction in an elderly Chinese inhabitant using imbalanced record. Between 2012 and 2014, information was collected from a prospective cohort of 1131 people and ML techniques are used for stroke prediction with imbalanced data. J.Yang et al. [10] proposed to evaluate the connection between the climate and stroke in this study with ML techniques. K.A. Mahesh et al. [11] work uses project risk variables to estimate stroke risk in older people, provide personalized precautions and lifestyle messages via web application, and use a prediction model for stroke prediction. R.Jeena et al. [12] worked with the various physiological markers used to calculate the risk of stroke are examined in this study. A.Sudha et al. [13] proposed the classification of strokes using the Decision Tree, Neural Network and Bayesian Classifier. They have 1000 records in their dataset. The PCA technique was used to reduce the number of dimensions. C.S.nwosu et al. [14] predicted brain stroke using multi-layer perception, random forest and decision tree models.

Moreover, B. Akter et al. [15] developed a model for accurate brain stroke forecasts using effective data collection, pre-processing, and transformation methods, utilizing a brain stroke dataset for its success. A.N. Tusher et al. [16] proposed method created a technique that allows us to detect brain strokes early and accurately. This system makes use of several categorization techniques. AF Abdel-Gawad et al. [17] research aims to develop a reliable stroke prediction model using machine learning algorithms and a diverse dataset, which could be used in healthcare settings to identify high-risk individuals. AR Abd Mizwar et al. [18] proposed method use of one of the ensemble learning techniques for predicts strokes, particularly the Xtreme Gradient Boosting algorithm, this research seeks to boost accuracy. S Mushtaq et al. [19] select the best machine learning algorithms for predicting brain strokes by examining existing research publications. The majority of research focuses on death rates and functional outcomes, indicating that future work should focus on predicting these strokes more effectively.

However, each of the above mentioned articles contained some limitations. Such as, V.Bandi et al. [3] had the inability to anticipate risk levels and many stroke kinds. BISWAS et al. [5] study cannot accuracy achieved utilizing machine learning. T.I.Shoily et al. [6] Work is constrained by the fact that the data collection is not symmetrical. M.S.Azam et al. [7] cannot anticipate the risk of stroke and analyze performance using big data. E.Dritsas et al. [8] worked was founded on an openly accessible dataset. Y.Wu et al. [9] model proposed only data accessibility; the population we studied was too small. X.Xia et al. [20] researchers conducted their initial interviews with subjects in neighborhood hospitals. Because patients with severe impairments were excluded from the study, the prevalence of stroke may have been overestimated. Second, because the specific traits of the various stroke sub types were not assessed. Third, socioeconomic research is subject to regional bias because it compares the provinces. J.Heo et al. [21] several limitations on this study exist. Being a single-center study, this investigation must be supported by data from multiple sources.The variables that are typically accessible in most cases served as inputs to the machine learning algorithms.

III. Methodology

The suggested system's technique is explained in this section.Multiple Kaggle datasets were considered in an effort to advance the implementation.From all of these datasets,pick one that will work for analysis.To come: data visualization and analysis. Pre-processing of the data from the datasets includes the filling in of missing or null values, the removal of extraneous columns, and the use of data encoding creating numerical values from category variables.Using smote technique,after then,the data is balanced.The test and train data are then segregated from the pre-processed data. The suggested technique is referred to as a hybrid technique because it has two main stages. The first stage is feature engineering, which has been carried out in a number of stages to identify and choose the ideal set of features required for forecasting stroke. The second stage is classification, where a machine learning classification model will be used to determine whether the patient has a stroke anomaly or not.A variety of Utilising machine learning methods load the training data and forecast how a stroke will turn out.Analysis of performance is last. Fig. 1 shows the technique for the suggested system in a flow chart format.

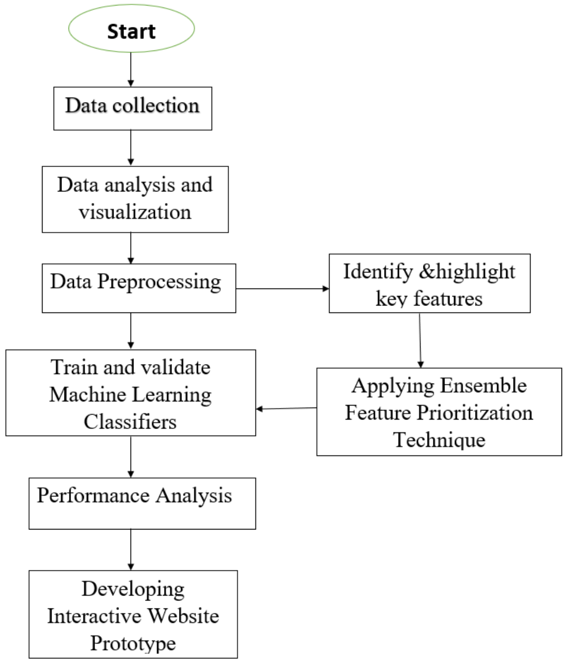

Start
Data collection
Data analysis and visualization
Data Preprocessing
Identify &highlight key features
Train and validate Machine Learning Classifiers
Applying Ensemble Feature Prioritization Technique
Performance Analysis
Developing Interactive Website Prototype

Fig. 1. Flowchart of Methodology

A. Data Collection and Analysis

data collection from Kaggle titled ”Stroke Prediction Dataset” [22]. This specific dataset has 12 characteristics and 5110 unique patient records (brief description has been shown in Table I).One of two values can be found in the output column ”stroke”: ”1” or ”0.” A risk of ”0” indicates no stroke risk, but a risk of ”1” indicates a possible danger. This dataset is particularly unbalanced since there is a higher probability of a value of ”0” than a value of ”1” in the output column (”stroke”).There are only 249 entries in the stroke column with the value ”1” and 4861 rows with the value ”0”.Additionally,the dataset included categorical type information such as gender,ever-married status, work type,residence type and smoking status.

B. Data Visualization

Data visualization is the process of representing data and information graphically to facilitate understanding, analysis, and decision-making.The type of data and the research topics you are addressing determine the visualization approaches you choose. In this work, typical data visualizations are employed. Fig. 2 displays data visualization using the pair-plot method.In machine learning and data analysis,The link between several variables in a dataset is examined using a pair-plot visualisation approach. The scatter-plots of the variables that are paired together are presented in the off-diagonal plots, and the histogram or kernel density estimate is displayed in the diagonal plots. This creates a grid of scatter-plots and histograms.

TABLE I
BRIEF DESCRIPTION OF THE ATTRIBUTED IN THE DATASET

SL.No	Attribute	Type(Values)	Description
1	id	Integer	a distinct patient-specific integer value
2	gender	String(Male, Female,Other)	the gender of the patient
3	age	Integer	Age of the patient
4	hypertension	Integer(0,1)	determines whether a patient has high blood pressure or not
5	heart_disease	Integer(0,1)	identifies the presence or absence of cardiac disease in the patient
6	ever_married	String(Yes,No)	It indicates whether or not the patient is married
7	work_type	String(children, govt_job, Never_worked, Private, Self_employed)	It provides several work categories
8	residence_type	String(Urban, Rural)	The type of patient's residence is saved
9	avg_glucose _level	Floating number	provides the value of the blood's average glucose level
10	bmi	Floating number	revels the patient's body mass index value
11	smoking_status	String(formerly smoked, never smoked, smokes, unknown)	It provides the patient's smoking history
12	stroke	Integer(0,1)	output column that displays the status of the stroke.

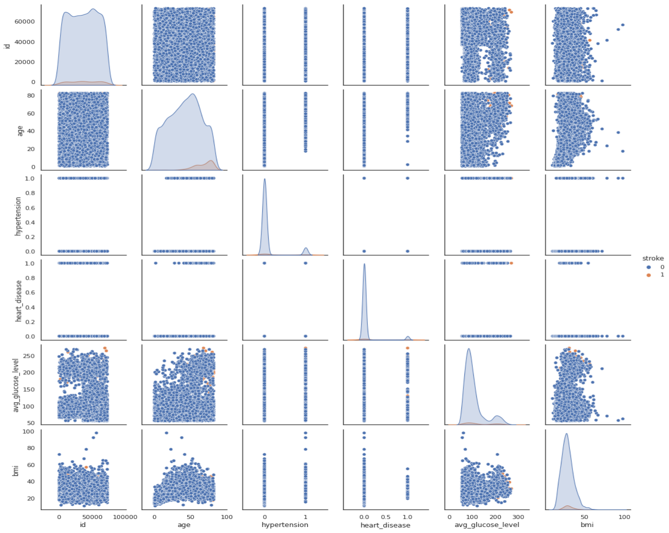


Fig. 2. dataset visualization

C. Data Pre-processing

To avoid having the model train incorrectly, undesirable noise and outliers from the dataset must be removed before creating the model. This phase entails removing any obstacles to the model's improved performance.After choosing the appropriate dataset has been gathered, For model creation, the data has to be purified.We have applied some preprocessing technique in our work.There are:

1.By using the SMOTE approach, balance imbalance data: The dataset utilized to perform the stroke prediction challenge is somewhat uneven. 5110 rows make up the entire dataset,of which 249 indicate a stroke's likelihood to occur and 4861 indicate a stroke's lack thereof. Fig.3 presents a graphic illustration of the imbalance. The results are inaccurate and the prediction is ineffective if such imbalanced data is not addressed.The unbalanced data must first be handled in order to create a useful model.The following approach is utilized for this.By matching the minority class to the under-sampled majority class, the Smote approach equalizes the data.In this case, compared to the class with a value of ”1”, the class with a value of ”0” is under-sampled.The dataset will ultimately include 249 rows with the value ”0” and 249 rows with the value ”1”. Fig. 4. provides a output column from the final dataset is shown graphically.

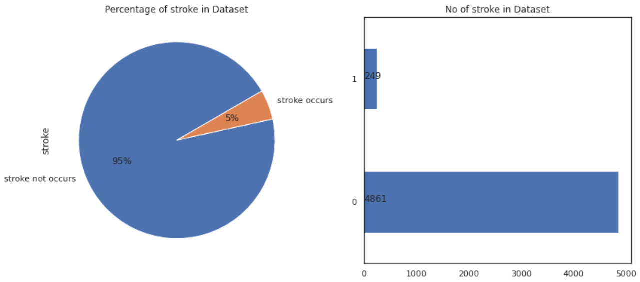


Fig. 3. Before applying smote technique (imbalanced data)

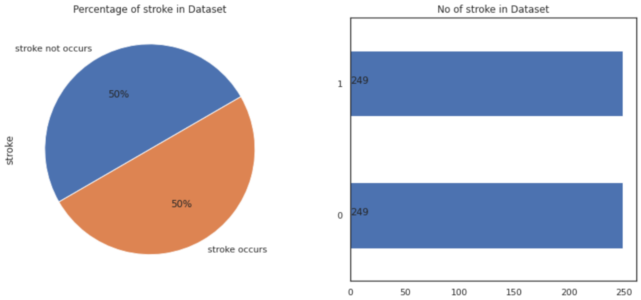


Fig. 4. After applying smote technique(balanced Data)

2.Null value remove: The datasets identified null values are subsequently filled in. Here, the ”bmi” column’s null values are the null values, which stand in for the column’s mean.Null values are removed, cleaning up the dataset.

3.Drop unnecessary column: The absence of the column ”id” does not significantly affect model construction.

4.Data encoding: The act of transforming a category value to a numerical value that can be quickly included into a machine learning method is known as data encoding.String values are contained in five columns with the names ”gender,” ”ever married,” ”work type,” ”residence type,” and ”smoking status.” We use the ”One Hot Encoding” technique for ”smoking status” and ”work type.” These string values are all transformed into a collection of numbers. The characteristics in the data frame now all have numerical values that might be utilised to create predictions. The dataset at this point has 249 records with 16 numerical features and one target attribute that may be used in a prediction model after preprocessing.

D. Feature selection and Prioritization

In order to limit the feature space as effectively as possible in accordance with a given criterion, a subset of the original characteristics are selected using the feature selection process. Five different feature selection approaches are utilised in this study, and the findings are combined using a majority voting mechanism in order to examine the most prevalent characteristics among the 17 numerical qualities. Each technique has chosen the top 12 characteristics according to their respective procedures, and all the qualities are then ranked based on the votes. The 12 features with the highest number of votes according to feature selection techniques are chosen to be applied in ML models, and the remaining 5 features are subsequently deleted from the data frame.The techniques for feature selection include the Pearson’s correlation coefficient approach, Principal component analysis, Chi-square method, recursive feature elimination, and the Light Gradient Boosting Machine. To demonstrate, however, how this step of feature selection dominates the effectiveness of predictive analytic, use the ensemble approach.

E. Applying Machine Learning Models

The procedure of building the model follows dealing with the unbalanced dataset and finishing data prepossessing.To increase the precision and productivity of this activity, the sampled data are split into train and test sets. After being split,To train the model,a number of classification techniques are utilized. In this study, ten different types of wisely utilized machine learning algorithms are used including Decision Tree, Naive Bayes,Neural Network Model, Random Forest, Stochastic Gradient Descent,Support Vector Classification and Nearest Neighbors, Adaboost classifier, XGboost(extreme Gradient Boosting), Logistic Regression. With the use of numerous performance indicators and training and testing each of these classifiers on the dataset both with and without the decreased feature sets, the top performing method has been identified.

F. Performance Metrics

Metrics for evaluation are connected to machine learning activities.The effectiveness of a classification model is evaluated using classification metrics. Other frequent measures include accuracy the percentage of appropriately predicted events, precision as the proportion of correctly identified events over all correctly identified events, recall as measured by the percentage of accurately identified incidents over all correctly identified events, and F1 score as the harmonic mean of precision and recall. We should be able to enhance our model’s overall predictive power utilising a range of performance assessment indicators before applying it to production data. without properly assessing the machine learning model utilising several assessment measures. The accuracy, precision, recall and F1 score formulas look like this:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

IV. Result Analysis

A. *Feature Selection*

In this work, five different feature selection techniques were used: the Pearson correlation coefficient methodology, Recursive Feature Elimination(RFE), Embedded Random Forest, Tree-based Feature Selection and Light GBM. Following data preprocessing, each of them utilized a different approach to select the 12 features out of a total of 17 features that they felt were most crucial, leaving out the other 5 non-prioritized qualities. Votes for each attribute are then calculated using a variety of methods and the results are shown in Table II. As a result,the new dataset only comprises the 12 traits that the assessment decided to be the most important, leaving out the five qualities mentioned earlier.

TABLE II
Attribute voting based on feature selection techniques

SL	Features	Pearson	Chi-2	RFE	Randm Forest	Lightagbm	Total
1	avg_glucose _level	TRUE	TRUE	TRUE	TRUE	TRUE	5
2	age _	TRUE	TRUE	TRUE	TRUE	TRUE	5
3	work_type _Govt_job	TRUE	TRUE	TRUE	FALSE	FALSE	3
4	smoking_status_ formerly smoked	TRUE	TRUE	TRUE	FALSE	FALSE	3
5	smoking_status _Unknown	TRUE	TRUE	TRUE	FALSE	FALSE	3
6	hypertension _	TRUE	TRUE	TRUE	FALSE	FALSE	3
7	heart _disease	TRUE	TRUE	TRUE	FALSE	FALSE	3
8	ever _married	TRUE	TRUE	TRUE	FALSE	FALSE	3
9	bmi _	TRUE	FALSE	FALSE	TRUE	TRUE	3
10	work_type _children	TRUE	TRUE	FALSE	FALSE	FALSE	2
11	work_type_ Self-employed	TRUE	TRUE	FALSE	FALSE	FALSE	2
12	work_type_ Never_worked	TRUE	TRUE	FALSE	FALSE	FALSE	2

B. *ML Classification Model*

Ten different types of ML classification-based models have been used in this study to estimate the stroke rate. To examine the effects of the features reduction process, each of these models has been trained and tested using the dataset with complete features as well as a reduced features set. Table III displays the results of several ML model types using the four performance measures accuracy , precision, recall and F1 score. The results also demonstrate that for all classifiers, accuracy increases when the feature set is decreased. The best accuracy in this case is 78% when all of the features from the dataset are used, while the best accuracy when utilizing the smaller feature set is 82.7% .Fig 5. graphically shows the comparative R- Squared analysis of ML models.This means that, when using the stroke dataset for prediction, ensemble forms of machine learning techniques often produce higher results.Additionally, a smaller dataset with the 12 most important features and a " Logistic regression" model can get the best prediction accuracy with 82.7% accuracy , 82.6% precision, 82.6% recall and 82.6% F1 score.

TABLE III
Performance Analysis from machine learning techniques

ML Classifier	with all attributes				with reduced attributes			
	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Decision Tree	0.647	0.64	0.639	0.64	0.72	0.72	0.721	0.72
Naive Bayes	0.72	0.737	0.735	0.72	0.627	0.764	0.601	0.544
Neural Network	0.767	0.779	0.78	0.767	0.807	0.808	0.809	0.807
Random Forest	0.773	0.784	0.786	0.773	0.813	0.815	0.81	0.811
Stochastic Gradient Descent	0.607	0.672	0.552	0.485	0.633	0.768	0.608	0.555
support vector classification	0.78	0.792	0.793	0.78	0.82	0.821	0.817	0.818
Nearest neighbors	0.693	0.702	0.668	0.667	0.687	0.72	0.698	0.682
Adaboost classifier	0.747	0.757	0.758	0.747	0.787	0.786	0.786	0.786
XGBoost	0.747	0.745	0.749	0.745	0.78	0.779	0.779	0.779
Logistic Regression	0.767	0.768	0.772	0.766	0.827	0.826	0.826	0.826

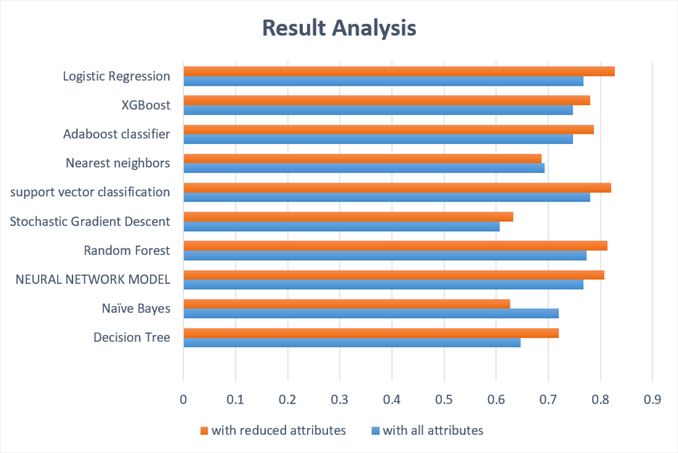
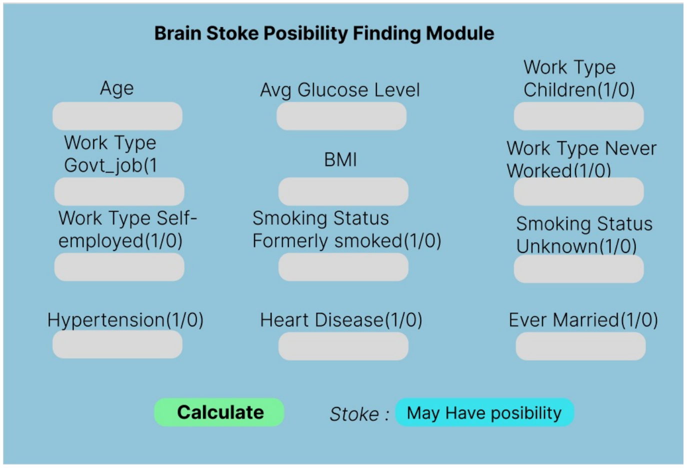


Fig. 5. Evaluation of Each Method's Performance

C. *Developing Interactive Website*

The existence of stroke disease can be predicted using an interactive online prototype. The ensemble technique was used to create a prediction system once the best-performing model had been determined. By using the dataset we utilised for our research, this online prototype has been evaluated. Screenshots of the website are shown in Fig. 6, giving visitors a look at its user interface. The dashboard allows users to enter their data, which is processed by the suggested algorithm to identify the existence of stroke illness. In order to quickly get the forecast result, users can click the calculate button. To assist users in taking preventive steps for their health, this online prototype leverages cutting-edge machine learning.


Brain Stoke Posibility Finding Module
Age
Avg Glucose Level
Work Type Children(1/0)
Work Type Govt_job(1
BMI
Work Type Never Worked(1/0)
Work Type Self-employed(1/0)
Smoking Status Formerly smoked(1/0)
Smoking Status Unknown(1/0)
Hypertension(1/0)
Heart Disease(1/0)
Ever Married(1/0)
Calculate

Stoke : May Have posibility

Fig. 6. User Interface of Dashboard from website prototype

V. Discussion & Conclusion

To prevent worsening the serious medical condition brain stroke, it must be treated immediately at an early stage. The biggest global reason for demise and condition according to the WHO, is stroke. Strokes, which impact the central nervous system, have recently risen to the top of the list of fatalities. To handle this situation machine learning based predictive analytic can be a pioneer. In this work, machine learning methods are used to diagnose,classify and forecast stroke using clinical data. In order to help with the early detection of strokes and lessen the severity of their effects,one can use a machine learning technique. This study investigates stroke prediction using several physiological variables from patients and then apply Machine Learning techniques to forecast. In this work, it has been attempted to prepare the dataset using several preprocessing techniques. Following that, the ten machine learning algorithms are applied to forecast a patient's risk of suffering a stroke. Logistics regression surpasses all other methods with an accuracy of 82.7%.Additionally, an interactive online application based on this effective paradigm has been created for simpler access.This initiative is regarded as the cornerstone of the healthcare system for stroke sufferers. Future studies in this domain may include examining performance and predicting the risk of stroke using dataset from a vast region as well as applying a range of modern machine learning technologies including deep learning, explainable AI, federated learning techniques etc.

References

[1] C. Beuker, D. Schafflick, J.-K. Strecker, M. Heming, X. Li, J. Wolbert, A. Schmidt-Pogoda, C. Thomas, T. Kuhlmann, I. Aranda-Pardos *et al.*, "Stroke induces disease-specific myeloid cells in the brain parenchyma and pia," *Nature Communications*, vol. 13, no. 1, p. 945, 2022.

[2] S. A. Suha and M. N. Islam, "Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique," *Heliyon*, vol. 9, no. 3, 2023.

[3] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning." *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020.

[4] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," *Neural Computing and Applications*, vol. 32, no. 3, pp. 817–828, 2020.

[5] "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Healthcare Analytics*, vol. 2, p. 100116, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772442522000569

[6] T. I. Shoily, T. Islam, S. Jannat, S. A. Tanna, T. M. Alif, and R. R. Ema, "Detection of stroke disease using machine learning algorithms," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019, pp. 1–6.

[7] M. S. Azam, M. Habibullah, and H. K. Rana, "Performance analysis of various machine learning approaches in stroke prediction," *International Journal of Computer Applications*, vol. 175, no. 21, pp. 11–15, 2020.

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

[9] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older chinese," *International journal of environmental research and public health*, vol. 17, no. 6, p. 1828, 2020.

[10] J. Yang, L. Ji, Q. Wang, and X. Lu, "The prediction model of stroke on climate factors by multiple regression," in *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*. IEEE, 2016, pp. 587–591.

[11] K. A. MAHESH, H. Shashank, S. Srikanth, and A. Thejas, "Prediction of stroke using machine learning," 2020.

[12] R. Jeena and S. Kumar, "Stroke prediction using svm," in *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*. IEEE, 2016, pp. 600–602.

[13] A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," *International Journal of Computer Applications*, vol. 43, no. 14, pp. 26–31, 2012.

[14] C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli, and D. John, "Predicting stroke from electronic health records," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 5704–5707.

[15] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas, and U. Sara, "A machine learning approach to detect the brain stroke disease," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2022, pp. 897–901.

[16] A. N. Tusher, M. S. Sadik, and M. T. Islam, "Early brain stroke prediction using machine learning," in *2022 11th International Conference on System Modeling Advancement in Research Trends (SMART)*, 2022, pp. 1280–1284.

[17] A. F. Abdel-Gawad, S. El-Sayed, M. M. Ismail *et al.*, "From data to diagnosis: Applied machine learning for stroke prediction in computational healthcare," *Journal of Artificial Intelligence and Metaheuristics*, vol. 3, no. 1, pp. 51–1, 2023.

[18] A. R. Abd Mizwar, A. Sunyoto, and M. R. Arief, "Stroke prediction using machine learning method with extreme gradient boosting algorithm," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 595–606, 2022.

[19] S. Mushtaq and K. S. Saini, "A review on predicting brain stroke using machine learning," in *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2023, pp. 667–673.

[20] 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," *Journal of neurology*, vol. 266, no. 6, pp. 1449–1458, 2019.

[21] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning–based model for prediction of outcomes in acute stroke," *Stroke*, vol. 50, no. 5, pp. 1263–1265, 2019.

[22] R. Choudhary. Stroke prediction dataset. [Online]. Available: https://www.kaggle.com/code/rishabh057/healthcare-dataset-stroke-data/input. Nov. 2020.