Journal of Applied Information Science
12 (2) 2024, 13-21
http://www.publishingindia.com/jais

Smart Healthcare: Brain Stroke Prediction using Machine Learning

Amita Jain[1*], Sejal Mehunkar[2], Ruchi Jha[2], Saloni Patel[2], Sneha Tikar[2] and Tushar[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Prestige Institute of Engineering Management and Research, Indore, Madhya Pradesh, India.
Email: amita.jain01@gmail.com
[2]Department of Computer Science and Engineering, Prestige Institute of Engineering Management and Research, Indore, Madhya Pradesh, India.
[*]Corresponding Author

Abstract: The brain, which comprises of the cerebrum, cerebellum, and brainstem and is covered by the skull, is a very complex and intriguing organ in the human body. Stroke is the world's second-leading cause of mortality; as a result, it requires prompt treatment to avoid brain damage. Early detection of a brain stroke can help to prevent or lessen the severity of the stroke, which can lower death rates. Using machine learning algorithms to identify risk variables is a promising method. This paper proposed a model that included a methodology to achieve an accurate brain stroke forecast. Efficient data collection, data pre-processing, and data transformation methods have been applied to provide reliable information for our proposed model to be successful. A "brain stroke dataset" was employed to build up the model. The standardization technique is used to standardize data. In the training and testing procedure, Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbors (KNN) and Decision Tree (DT) classifiers are applied. The performance of each classifier has been estimated by adopting performance evaluation metrics such as accuracy, f1-score, precision, recall etc. The dataset was preprocessed by deleting unnecessary columns, dropping null values, removing outliers and by encoding the categorical values using One Hot Encoding. Oversampling technique is used to balance the dataset as it was highly unbalanced. After evaluation the findings reveal that the proposed stacking ensemble approach performs better than the single models, achieving an accuracy rate of nearly 99%.

Keywords: Brain stroke, Disease prediction, Machine learning, Random forest.

I. Introduction

The most severe and deadly disease in humans has long been thought to be brain stroke. The increased occurrence of brain stroke, which is associated with a high death rate, poses considerable risk and burden to healthcare systems worldwide. The brain is the most intricate element of the human body, as we all know. This three-pound organ is the brain's seat of intellect, as well as a sensation interpreter, movement creator, and behavior controller. It's a part of the brain that controls cognition, memory, emotion, touch, motor skills, vision, breathing, temperature, hunger, and other critical human activities. The brain, housed in a bone shell and kept clean by protective fluid, is the source of all the characteristics that define our humanity. Brain stroke occurs when blood-flow to a part of the brain is restricted or reduced, depriving brain tissue of oxygen and nutrients. Brain cells begin to die in a minute under this circumstance. The number of people suffering from stroke is increasing every day. Strokes in the brain are more common in males than women, especially in middle and older age. On the other hand, Stroke affects roughly 8% of children with sickle cell disease. A stroke affects 15 million individuals globally every year [1]. Five million of them die, and another five million are permanently crippled, putting a strain on families and communities.

To lower the morality of brain stroke patients, we built a prediction system that may help clinicians diagnose illnesses by feeding data from the patients into the processing system. The system analyzes the data, and the data is pre-processed by data pre-processing method. Finally, ML algorithms were used to identify brain stroke patients. This research aims to create a system that can predict brain stroke correctly. The stages that must be accomplished are as follows:

- Early diagnosis of diseases can aid in the reduction of risk.
- Brain stroke can destroy the global healthcare system.
- This method can be used to assist the physician.
- Early discovery of a brain stroke will also lead to advice on how to manage it in a more secure manner.
- Reduce the morality through brain stroke.
- The overall findings are used to assess the performance of the various models.

The accessibility of clinical data stored in electronic health records, known as EHRs, has facilitated the utilization of machine learning (ML) methods for identifying various medical conditions, including stroke, even in its early stages. By early detection of stroke, long term disability and mortality related with stroke can be prevented. Recently, machine learning techniques have been efficiently predicting risk of stroke. Ensemble techniques in machine learning involve merging the outcomes of multiple classification models and have demonstrated superior performance compared to individual models. Thus, the objective of this study is to employ an ensemble learning method for early-stage stroke detection using clinical data obtained from common health records.

In this work, a publicly available dataset was utilized for detecting strokes. The dataset was at first preprocessed by deleting unnecessary columns, dropping null values, detecting and removing outliers and converting the categorical values to numerical values using Label Encoding. The dataset was highly unbalanced, so we used oversampling to balance the data. Then various models were selected and evaluated on the processed dataset. The primary focus of this study is a stacking method, where five different models were used as base models and one as meta model to enhance the overall performance.

Different performance metrics such as accuracy, precision, recall, F-1 score, 7 auc score of the classification models are determined and compared with each other to decide which one predicts more accurately on the dataset. The experimental results of the proposed methods were also compared with the existing works henceforth to show the novelty of the work.

II. Literature Survey

In terms of early illness prediction, machine learning is a promising technology. Many studies have been conducted to predict diseases such as cancer, skin disease, stroke illness, and so on. Some researchers have employed different techniques for finding patterns for Mycobacterium Tuberculosis Complex (MTBC) [2-5]. There is virtually little study on the prediction of brain stroke illness. Multiple research studies have utilized machine learning models, resulting in significant findings regarding the detection of stroke.

G. A. P. Singh and P. K. Gupta, Performance analysis of various machine learning-based approaches for detection and classification of stroke. Detect and prediction of the stroke using different Machine Learning algorithms [8].

In [6] the authors developed and evaluated various models including Logistic regression, stochastic gradient descent (SGD), K-NN, decision trees, random forests, multi-layer perception, and Naive

Bayes. They further explored ensemble techniques, specifically majority voting and stacking method. The stacking model was most efficient with an accuracy of 98%.

Moreover, in [7], the authors explored diverse factors and employed several ML algorithms for accurate stroke prediction. They also created a user-friendly HTML page with a Flask application, enabling users to input parameters and obtain results. Notably, the NBC (Naïve Bayes Classification) algorithm performed the best with an 82% accuracy rate. Furthermore, the study suggests future extractions involving Neural Networks for model training and the collection of image datasets.

Similarly, in [8] several physiological measurements and machine learning techniques, including the Voting Classifier, Logistic Regression (LR), Decision Tree (DT) Classification, and Random Forest (RF) Classification, were utilized to create four separate predictive models. The Random Forest algorithm achieved an accuracy of around 96 percent for brain stroke forecasting. A future scope was suggested where broadening the dataset and exploring advanced machine learning techniques like SVM, AdaBoost and Bagging will enhance the framework model.

Furthermore, in [9] the author analyzed eleven models including AdaBoost, SVM, Gradient Boosting, Random Forest, Nearest Centroid, Multi-Layer Perception, Voting Classifier, Decision Tree, KNN, and Naïve Bayes to achieve more accurate stroke diagnosis with imbalanced data. This research resulted in the creation of a mobile app and web page for real-time risk assessment. Among the eleven models, the SVM and RF models demonstrated the highest accuracy rates, achieving 99.99% and 99.87%, respectively.

And finally, in [10], the authors compared their stroke prediction approach with alternative methods. They collected and preprocessed data, performed feature selection using the C4.5 algorithm with Decision Trees, applied Principal Component Analysis (PCA) for dimensionality reduction, and employed Artificial Neural Networks (ANN) for classification. Their proposed method exhibited superior accuracy when compared to two other methods. Hence, this paper proposes a Stroke Detection framework that incorporates various data preprocessing methods and a stacked ensemble learning approach for detecting stroke.

G. Vijayadeep *et al*. [11] proposed a hybrid feature extraction-based system to predict brain stroke applying a random forest classifier. K-cross-validation technique used for extracting the feature and the obtained accuracy was 98.23%.

A Real-time gait monitoring system for stroke prediction service referred by S. J. Park *et al*. [12]. Their work was conducted with Gait parameters of 63 stroke patients and 208 healthy patients. They used RT, CART-, C5.0, SVM, LR and LSVM classifier in their model and generate highest accuracy AUC: 0.995, Gini: 0.993 with C5.0 classifier. They resulted in the lowest accuracy AUC: 0.908, Gini: 0.816 with LSVM classifier.

T. Badriyah *et al*. [13] build a classification model for predicting two sub-types of stroke disease, Ischemic stroke, and stroke hemorrhage depend on CT scan data those obtained 102 patients. They used eight machines learning algorithm to generate the accuracy. Their model resulted in 95.97% accuracy with random forest algorithm and lowest 71.18% accuracy with Naive Bayes algorithm.

T. Liu *et al*. [14] proposed a hybrid machine learning approach to predict cerebral strokes. For their proposed method, they used HealthData.gov, which was utilized from the benchmark dataset from Kaggle. They counted the accuracy of KNN, LR, RF DECT depending on smoking status and SVM, RFR, BayRid, and GBM with BMI status variable and obtained 49.6% accuracy from the KNN classifier and 87.6% accuracy from the RFR classifier.

G. Fang *et al*. [15] proposed a machine learning approach for predicting stroke prognosis. They used an IST dataset that contained 19435 patient data of 467 hospitals. They used SVC, MLP, RF, and AdaBoost classifiers to predict RVISINF of acute stroke. They obtained the highest accuracy on the RVISINF model.

P. Govindarajan *et al*. [16] proposed a Machine learning model to detect stroke patients. Their work used 507 patient data collected from Sugam Multispecialty hospital, Kumbakonam. They used ANN, SVM, DT, LR, Boosting, and bagging techniques. ANN resulted in the highest accuracy for their approach that was 95.3%.

M. Emon *et al*. [17] proposed a machine learning approach for stroke prediction. In their research, they used 5110 people's data from the medical clinic of Bangladesh. They used LR, SGD, DT, AdaBoost, Gaussian, Quadratic Discriminant Analysis, MLP, KNN, and XGBoost Classifier for predicting the stroke. They obtained 97% accuracy from the Weighted Voting classifier and 65% accuracy from SGD.

A novel prediction model was introduced in the paper of T. Shoily *et al*. [18] with several general classifications and different combinations of features techniques. They used NB, J48, KNN, and RF Classifiers to build their model. Obtained accuracy of 99.8% with the J48, KNN, RF classifier, and 85.6% with NB Classifier.

The essential risk factors were identified in the study by M. Amin *et al*. [19], and machine learning models (k-NN, DT, NB, LR, SVM, Neural Network, and a hybrid of voting with NB and LR) were used to undertake a comparative analysis. According to their findings, it was founded that when combined with the selected attributes, the hybrid model attained an accuracy of 87.41.

For prediction, M. Ashraf *et al*. [20] employed individual learning algorithms and ensemble techniques such as Bayes Net, J48, KNN, MLP, NB, RT, and RF. J48 was the most accurate, with a score of 70.77 percent. They then used cutting-edge approaches, with KERAS achieving an accuracy rate of 80. From the above studies, it is observed that there is very little research on brain stroke prediction, and they adopted tiny datasets to implement their works.

III. Proposed Methodology

The primary objective of this study is to develop a system capable of identifying strokes. In pursuit of this goal, we devised two distinct approaches. The first approach involves using the results from each separate model, while the second approach involves combining these models into an ensemble. To achieve accurate stroke detection, we employed different machine learning models such as SVM, Decision Tree, Random Forest, KNN and Logistic Regression.

A. Introduction to the Dataset

A dataset is a collection of data. In the case of tabular data, a dataset corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the dataset in question. The data set lists values for each of the variables, such as gender, age and bmi of the person. Datasets can also consist of a collection of documents or files. The stroke prediction dataset was used to perform the study. There were 5110 occurrences and 15 attributes in this dataset. Fourteen attributes are utilized as inputs, with one attribute functioning as an output. The output column's stroke has a value of either 1 or 0. A stroke risk was found when the value 1 was detected, but a stroke risk was not found when the value 0 was displayed. In this dataset, there is a greater chance of a 0 in the output column (stroke) than a 1. The stroke column alone has 249 rows with a value of 1, whereas 4861 rows have a value of 0. To improve accuracy, data preprocessing is used to balance the data.

B. Data Preprocessing

i) Removing Unnecessary Columns

The columns that are not related to the target feature need to be deleted. In this work the "id" column has no relevance with stroke. So, this column has been deleted.

ii) Dropping Null Values

There were a total of 201 null values identified in the BMI feature which need to be handled. After counting these null values, they were dropped for proper functioning of the machine learning algorithm. Dropping null values may lead to valuable information loss.

iii) Detecting and Removing Outliners

Outliers are data points that are significantly different from the majority of the data from dataset. They can result from measurement or execution errors. In this work a total number of 649 outliers were detected by using the Interquartile Range (IQR) method and then they were removed from the dataset. Before removing the outliers there were a total of 4909 samples and after removing outliers there were total 4260 samples.

iv) Data Balancing

Imbalanced data is a prevalent issue in classification tasks where one target class label is significantly greater than the other class label. In this work the data was highly imbalanced with only 136 samples indicating a stroke, while a vast majority of 4,124 samples exhibited no signs of stroke. In this study, to match the number of samples between majority class (No Stroke) and minority class (Stroke) oversampling method was used.

v) Data Encoding

Data encoding is a way of transforming the categorical values to numerical values as many machine learning algorithms require numerical inputs to perform calculations and make predictions. In this work one Label Encoding has been used for encoding creating binary columns for each category.

vi) Data Scaling

Data scaling is a process which is used to transform the features or variables to a similar scale. It ensures equal contribution of all features to the learning algorithm. In this study Standard Scaler has been used for scaling.

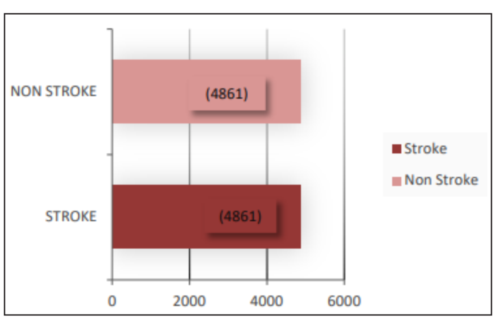

Fig. 1: Total Count of Stroke and Non-Stroke Data After Pre-Processing

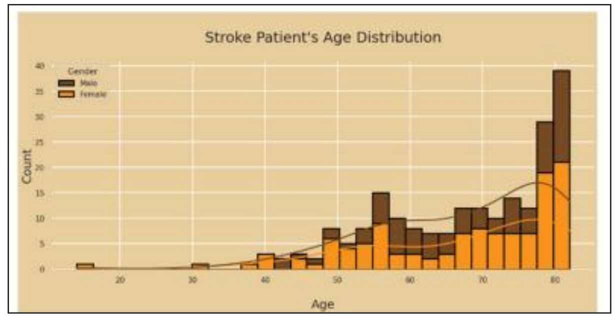

Fig. 2: Age Distribution of Stroke Patients

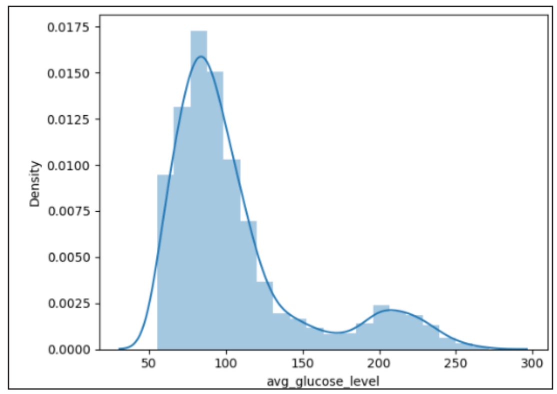

Fig. 3: Average Glucose Level Distribution

C. Model Description

i) Logistic Regression

Logistic Regression is a simple and interpretable supervised machine learning model which is generally employed for binary classification problems, though it can be extended to solve multiclass classification problems. It is used when the data is linearly separable, and the outcome is binary or dichotomous in nature [21]. The core of this model relies on the sigmoid function, creating its characteristic "S"-shaped curve. The output of this model may be yes or no, 0 or 1, true or false, etc. In this work the output belongs either to class "Stroke" or class "No Stroke".

ii) Decision Tree

Decision Tree (DT) is a widely used Machine Learning model applicable to both regression and classification tasks. It is known for its interpretability which makes it easier to understand decision-making. It forms a hierarchical structure where data is divided into branches, leading to predictions at the leaf nodes. In this model, internal nodes represent features, while leaf nodes correspond to classes. There are two main components of a decision tree: decision nodes, where data is split, and leaf nodes, where predictions are made. If the tree is deep it may lead to overfitting.

iii) Random Forest

Random Forest (RF) is an algorithm used for supervised machine learning and widely employed for both regression and classification tasks. It operates as an ensemble method, consisting of numerous separate decision trees, each trained on a randomly selected subset of the data. For final predictions, a voting mechanism is applied. In this study RF achieved the best accuracy. One of the notable strengths of the Random Forest algorithm is its resistance to overfitting, a common concern in machine learning. If there are enough trees present in the forest, the classifier will avoid overfitting the model.

iv) KNN

K-Nearest Neighbors is a distance-based classifier used in supervised learning techniques. It is considered as a nonparametric algorithm because it doesn't rely on any assumptions about the underlying data distribution. KNN is a "lazy" algorithm, which means it doesn't immediately train on the dataset. Instead, it stores the dataset and performs actions on it during classification. The choice of K is very critical, as small value of K leads to overfitting and large value of K leads to underfitting. KNN is useful when the dataset is not very large. Its flexibility in handling complex data patterns makes it a valuable tool in various machine learning applications.

v) Support Vector Machine

SVM is a type of supervised Machine Learning algorithm utilized tasks in both classification and regression, though especially well-suited for classification tasks. This classification identifies the optimal hyperplane that separates the dataset into two classes [22] (See Fig. 6).

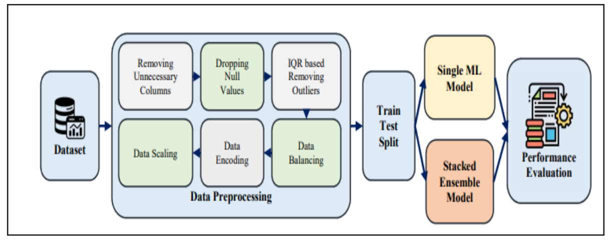

Fig. 4: Flow Diagram of Proposed Stroke Detection Technique

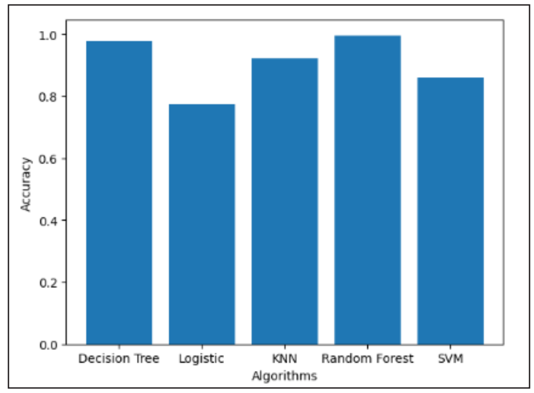

Fig. 5: Accuracy Comparison of Single Models

D. Ensemble Learning

In this research, we utilized stacking, an ensemble learning technique that combines diverse classifiers within a voting classifier. Our approach involves training base models (Decision Tree, Logistic Regression, SVM, Random Forest and KNN) on the training data and then using their predictions to train a voting-classifier.

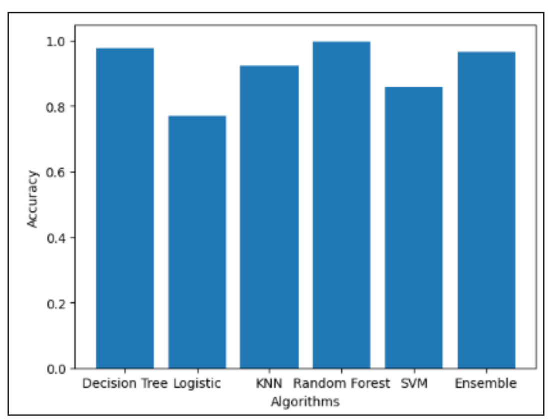

Fig. 6: Accuracy Comparison of Single and Ensemble

IV. Results

We used a "Brain stroke prediction dataset" to build our model. This dataset contained a total of 5110 patient observations and 15 attributes. The attributes were gradually gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose, cholesterol level, BMI, smoking status, alcohol consumption, diabetes and stroke. Each attribute holds different symptoms, which helps to make a decision. We evaluated the correlation with each attribute using heat-map (See Fig. 7).

In this study, we initially assessed the performance of different individual machine learning models by employing a confusion matrix. Fig. 2 and Table I show the performance of Decision Tree, SVM, Logistic Regression, Random Forest and KNN models and proposed stacked ensemble model.

A. Simple Model Evaluation

i) Logistic Regression

Table I illustrates the performance of Logistic Regression. The accuracy achieved with this algorithm is approximately 77%. Additional metrics evaluating this model's efficiency include precision (74%), recall (81%), and F-1 score (77%) for the "stroke" case. In this study Logistic Regression performed the worst out of all the models.

ii) Decision Tree

Table I depicts the performance of Decision Tree which achieved an accuracy of approximately 98%. The precision, recall and F-1 score for the "stroke" case are 96%, 100% and 98% respectively.

iii) Random Forest

Table I depicts the performance matrix of Random Forest (RF). In this work this model achieved the best accuracy out of all the models which is 99.74%. The recall, F1-score and precision are also very high which are 100%, 99% and 99% respectively for stroke class.

iv) KNN

The matrix in Table I shows a high performance of KNN model. This model achieved a high accuracy of around 92%. For the stroke class precision and F1-score reached 86% and 92%, respectively, while recall was 100%.

v) Support Vector Machine.

The matrix in Table I reveals that the accuracy achieved by SVM model is around 85%. The precision, recall, and F1-score for the "stroke" class are 80%, 93%, and 86%, respectively.

B. Ensemble Model Evaluation

The matrix depicted in Table II reveals that the stacking ensemble model achieved an accuracy rate of 95%. For the stroke class, the precision is 90%, while the recall and F1-score are both 100% and 95% respectively.

It presents a comparison between the proposed model and the individual models. Our findings indicate that the proposed model has performed well like all other models.

Table I: Performance Evaluation Metrics for Individual Models

Model	Accuracy (%)	F1 Score	Precision Score	Recall Score	AUC Score
Decision Tree	97.686	0.977	0.956	1.0	0.976
Logistic Regression	77.018	0.779	0.749	0.811	0.770
k-Nearest Neighbors	92.185	0.927	0.865	1.0	0.922
Random Forest	99.743	0.997	0.995	1.0	0.997
Support Vector Machine	85.861	0.869	0.809	0.938	0.859

Table II: Performance Evaluation Metrics for Ensemble Model with Individual Models

Model	Accuracy (%)	F1 Score	Precision Score	Recall Score	AUC Score
Decision Tree	97.686	0.977	0.956	1.0	0.976
Logistic Regression	77.018	0.779	0.749	0.811	0.770
k-Nearest Neighbors	92.185	0.927	0.865	1.0	0.922
Random Forest	99.743	0.997	0.995	1.0	0.997
Support Vector Machine	85.861	0.869	0.809	0.938	0.859
Ensemble Model	95.013	0.952	0.909	1.0	0.950

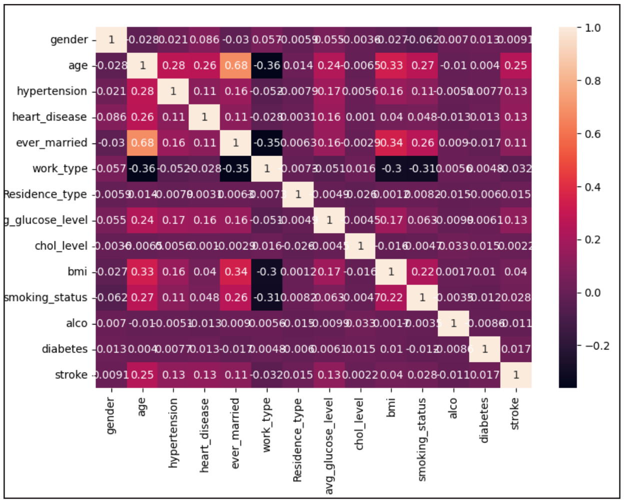

Fig. 7: Correlation Among Attributes by Heat-Map

V. Conclusion

In this study we proposed a stacking ensemble approach and compared its performance with traditional single classifiers, namely SVM, Decision Tree, KNN, Logistic Regression and Random Forest. We began by collecting a dataset, preprocessing it, selecting various classifiers, evaluating them and finally comparing their performance with our proposed stacking approach. The preprocessing process includes deleting unnecessary columns, dropping null values, detecting and removing outliers, data balancing using oversampling method, data encoding for converting categorical values to numerical values and data scaling by using Standard scaler.

Stroke is a potentially fatal medical condition that needs to be treated right away to prevent future consequences. The creation of a machine learning ML model could help with stroke early diagnosis and subsequent reduction of its severe consequences. This study examines how well different machine learning ML algorithms predict stroke based on various biological factors. With a classification accuracy of 99%, random forest classification exceeds the other investigated techniques. According to the study, the random forest method performs better than other methods when forecasting brain strokes using cross-validation measures. By using this system, we can predict brain stroke earlier and take the required measures to decrease the effect of the stroke. So that it saves the lives of the patients without going to death.

Acknowledgement

The authors would like to thank the Department of Computer Science and Engineering, Prestige Institute of Engineering Management and Research, Indore for providing support.

References

[1] Worldwide Stroke Affect, WHO. Accessed: Oct. 06, 2021. [Online]. Available: http://www.emro.who.int/health-topics/stroke-cerebrovascularaccident/index.html

[2] A. Jain, and K. R. Pardasani, "Mining fuzzy amino acid associations in peptide sequences of mycobacterium tuberculosis complex (MTBC)," *Netw Model Anal Health Inform Bioinforma*, vol. 4, no. 3, 2015, doi: https://doi.org/10.1007/s13721-015-0075-4.

[3] A. Jain, and K. R. Pardasani, "Soft fuzzy model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex," *Current Science*, vol. 110, no. 4, pp. 603-618, 2016.

[4] A. Jain, and K. R. Pardasani, "Fuzzy soft set model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex (MTBC)," *Journal of Intelligent & Fuzzy Systems*, vol. 31, no. 1, pp. 259-273, 2016.

[5] A. Jain, and K. R. Pardasani, "Fuzzy-soft-fuzzy set model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex (MTBC)," *International Journal of Data Mining and Bioinformatics*, vol. 17, no. 1, pp. 1-24, 2017.

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

[7] G. Sailasya, and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ML classification algorithms," *International*

Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.

[8] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke disease detection and prediction using robust learning approaches,” *Journal of Healthcare Engineering*, 2021.

[9] N. Biswas, K. M. M. Uddin, S. T. Rikta, and S. K. Dey, “A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach,” *Healthcare Analytics*, vol. 2, p. 100116, 2022.

[10] M. S. Singh, and P. Choudhary, “Stroke prediction using artificial intelligence,” in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, IEEE, Aug. 2017, pp. 158-161.

[11] G. Vijayadeep, and N. N. M. Rao, “A hybrid feature extraction based optimized random forest learning model for brain stroke prediction,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* vol. 11, no. 3, pp. 1152-1165, 2020.

[12] S. J. Park, I. Hussain, S. Hong, D. Kim, H. Park, and H. C. Benjamin, “Realtime GAIT monitoring system for consumer stroke prediction service,” in *IEEE International Conference on Consumer Electronics (ICCE),* 2020, pp. 1-4.

[13] T. Badriyah, N. Sakinah, I. Syarif, and D. R. Syarif, “Machine learning algorithm for stroke disease classification,” in *International Conference on Electrical, Communication, and Computer Engineering (ICECCE),* 2020, pp. 1- 5,

[14] T. Liu, W. Fan, and C. Wu, “A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,” *Artificial Intelligence in Medicine*, vol. 101, p. 101723, 2019.

[15] G. Fang, W. Liu, and L. Wang, “A machine learning approach to select features important to stroke prognosis,” *Computational Biology and Chemistry*, vol. 88, p. 107316, 2020.

[16] 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.

[17] M. Emon, M. Keya, T. Meghla, M. Rahman, M. A. Mamun, and M. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA),* IEEE, 2020, pp. 1464-1469.

[18] T. Shoily, T. Islam, and S. Jannat, “Detection of stroke disease using machine learning algorithms,” in *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT),* IEEE, 2019, pp. 1-6.

[19] M. Amin, Y. Chiam, and K. Varathan, “Identification of significant features and data mining techniques in predicting heart disease,” *Telematics and Informatics*, vol. 36, pp. 82-93, 2019.

[20] M. Ashraf, S. Ahmad, N. Ganai, R. Shah, M. Zamanand, and S. Khan, “Prediction of cardiovascular disease through cutting-edge deep learning technologies - An empirical study based on TENSORFLOW, PYTORCH and KERAS,” in *International Conference on Innovative Computing and Communications*, Springer, Singapore, 2021, pp. 239-255.

[21] H. Bonthu, “An introduction to logistic regression,” *Analytics Vidhya*, 2021.

[22] S. R. Amendolia, G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala, and G. M. Mura, “A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening,” *Chemometrics and Intelligent Laboratory Systems*, vol. 69, no. 1-2, pp. 13-20, 2003.