International Journal of Bioinformatics Research and Applications

ISSN online: 1744-5493 - ISSN print: 1744-5485
https://www.inderscience.com/ijbra

Performance evaluation of machine learning classifiers for brain stroke prediction

Drishti Arora, Rakesh Garg, Farhan Asif, Ritvik Garg, Neetu Singla

DOI: 10.1504/IJBRA.2024.10060748

Article History:

Received:	06 June 2023
Last revised:	11 July 2023
Accepted:	24 July 2023
Published online:	14 March 2024

Performance evaluation of machine learning classifiers for brain stroke prediction

Drishti Arora, Rakesh Garg* and Farhan Asif

Department of Computer Science and Engineering,
Amity University Uttar Pradesh,
Noida, India
Email: drishtiarora1102@gmail.com
Email: rkgarg06@gmail.com
Email: fasif168@gmail.com
*Corresponding author

Ritvik Garg

Department of Computer and Communication Engineering,
Manipal University,
Jaipur, Rajasthan, India
Email: ritvikgarg0510@gmail.com

Neetu Singla

Department of Computer Science and Engineering,
The NorthCap University,
Gurugram, India
Email: neetusingla@ncuindia.edu

Abstract: A cerebral vascular accident, commonly known as a stroke, is a pathological condition that impacts the brain due to the rupture of capillaries. It occurs when there is a disturbance in the typical blood circulation and essential physiological processes of the brain. As per the WHO, stroke is the foremost aetiology of mortality, a significant public health concern. While there has been considerable research on the prognosis of heart attacks, investigating the risk factors associated with strokes has been relatively limited. Considering this, a plethora of advanced machine learning models has been leveraged to prognosticate the probability of an impending stroke event. The prime focus of this study is the performance evaluation of eight distinct machine learning classification models as support vector classifier, K-nearest neighbour, logistic regressor, decision tree classifier, random forest classifier, Naïve Bayes classifier, AdaBoost classifier, and XGBoost classifier used for brain stroke prediction. The performance statistics obtained through experimental setup shows that the XGBoost algorithm demonstrated remarkable accuracy, yielding prediction results of approximately 92.75%, making it the preeminent model for precise and reliable stroke prediction.

Keywords: brain stroke prediction; machine learning classifiers; accuracy; AUC score.

Reference to this paper should be made as follows: Arora, D., Garg, R., Asif, F., Garg, R. and Singla, N. (2024) 'Performance evaluation of machine learning classifiers for brain stroke prediction', *Int. J. Bioinformatics Research and Applications*, Vol. 20, No. 1, pp.61–77.

Biographical notes: Drishti Arora is pursuing her Bachelor of Technology in the Department of Computer Science and Engineering at Amity University, Uttar Pradesh, Noida, India. Her research interests include artificial intelligence, machine learning and optimisation.

Rakesh Garg is a Professor in the Department of Computer Science and Engineering at Amity University, Uttar Pradesh, Noida, India. He received his BE, MTech and PhD in Computer Science and Engineering in 2003, 2006, and 2016 respectively. His research interests include component-based software engineering, software reliability and multi-criteria decision-making optimisation. He has published a number of research papers in reputed international journals and guided a number of master's and doctoral degree dissertations.

Farhan Asif is pursuing his Bachelor of Technology in the Department of Computer Science and Engineering at Amity University, Uttar Pradesh, Noida, India. His research interests include artificial intelligence, machine learning and optimisation.

Ritvik Garg is pursuing his Bachelor of Technology in the Department of Computer and Communication Engineering at Manipal University, Jaipur, Rajasthan, India. His research interests include artificial intelligence, machine learning and optimisation, multi-criteria decision making, etc.

Neetu Singla is currently working as an Assistant Professor in The NorthCap University. She is pursuing her PhD from the Department of Computer Science and Engineering, Netaji Subhas University of Technology, New Delhi. She has received her MTech in Computer Engineering from the Maharishi Markandeshwar University and BTech in Information Technology from the Kurukshetra University. She has 12 years of experience in teaching. Her research interests include computer vision, video analytics, and video forensics.

1 Introduction

A brain stroke ensues when the blood perfusion to specific brain regions is impeded, leading to an acute deprivation of vital oxygen to the brain cells. It is a perilous medical condition that requires prompt intervention. There are two primary classifications of strokes: ischemic strokes, resulting from an obstruction of a blood vessel supplying the brain, and hemorrhagic strokes, which arise from a ruptured or leaking blood vessel in the brain. The incidence of brain stroke is on a constant rise worldwide, with middle-income countries being the most affected. The complications arising from brain stroke can cause dysfunctions in various parts of the body, leading to impaired motor functions, cognitive deficits, and other neurological issues (Sandhu et al., 2022). WHO estimates that approximately 15 million people experience strokes annually (Wolfe, 2000). The statistics surrounding brain stroke are staggering. As per the World Health Organization (WHO), strokes are the second most prevalent cause of death and the third most common

cause of disability across the globe. In the USA, a person suffers from a stroke every 40 seconds, leading to approximately 140,000 deaths annually.

The alarming statistics on the prevalence and impact of stroke highlight the urgent need for timely diagnosis and treatment. Early detection of stroke symptoms and accurate diagnosis can help prevent long-term disabilities and reduce the risk of death. Therefore, there is a pressing need for ongoing research into the development of effective stroke prediction models and diagnostic tools that can accurately identify high-risk individuals and facilitate early intervention. The prediction of brain stroke entails the evaluation of diverse risk factors, including but not limited to hypertension, tobacco use, hyperlipidaemia, diabetes, adiposity, sedentary lifestyle, familial history of stroke, advanced age, gender, and other comorbidities that may heighten the propensity for a cerebrovascular event (Dev et al., 2022). Machine learning (ML) algorithms have proven invaluable, providing accurate analysis and reliable predictions. ML algorithms trained on historical data to learn the intricate relationships between risk factors and stroke outcomes. Subsequently, these models can be employed to analyse new individual data, providing personalised stroke risk assessments based on their unique risk factors (Saber et al., 2019). The accuracy of these models can be continuously enhanced through ongoing refinement and validation using real-world data. These models can be utilised in conjunction with comprehensive medical assessments by qualified healthcare professionals to ensure precise risk assessment and appropriate management strategies (Govindarajan et al., 2020).

The field of stroke prognostication has seen significant advances in recent years with the application of ML techniques to predict stroke outcomes based on historical data and individual characteristics. However, there is still a need for a comparison of the existing work in this field to establish the most effective ML classifiers for stroke prognosis. This research aims to fill this gap by conducting a comparative study of various ML classifiers built by people based on historical stroke data and individual patient characteristics. The study will evaluate the performance of different classifiers, including LR, SVM, D-trees, KNN, Naïve Bayes, and ensemble learners like RF, AdaBoost and XGboost, using various evaluation metrics such as accuracy, precision, recall, F-measure and area under the curve (AUC).

The subsequent sections of the paper are organised as follows: in Section 2, a concise overview of previous studies concerning the comparison and evaluation of classification techniques is presented. Section 3 outlines the methodology employed in this paper. Section 4 provides a detailed description of the dataset, exploratory data analysis, data pre-processing, feature selection techniques, and an overview of the classification techniques investigated in this study, along with the experimental results. The outcomes of different classifiers, evaluated using various performance metrics, are discussed in Section 5. Finally, Section 6 presents the conclusions derived from this research.

2 Literature review

In recent times, ML algorithms have demonstrated promising efficacy in predicting stroke risk through analysing extensive datasets and identifying intricate patterns among diverse risk factors. A thorough evaluation was undertaken to critically appraise the efficacy of ML-based predictive models for stroke risk assessment. A systematic search

of esteemed databases was undertaken, utilising strategically chosen keywords, including 'stroke', 'ML', 'risk prediction', and 'predictive analytics'. Emon et al. (2020) proposed an advanced weighted voting method that seamlessly incorporated a range of diverse base classifiers, including LR, SGD, D-tree, AdaBoost, Gaussian, QDA, multi-layer perceptron, KNN, gradient boosting, and XGBoost. The approach combined the predictions of these classifiers to achieve an impressive accuracy of 97%, indicating that their approach effectively harnessed the unique capabilities of these classifiers, resulting in highly accurate classification performance. Hung et al. (2017) proposed a novel algorithm that combined DNN and GBDT techniques, resulting in an impressive performance in high UAR and AUC scores. However, their approach could not assess the feasibility of implementing DNN in a clinical setting, suggesting that while their algorithm showed promising results, further investigation and validation in real-world clinical scenarios would be necessary to ascertain its practical applicability in a clinical context. Khosla et al. (2010) utilised a margin-based censored regression approach and a conservative mean heuristic for feature selection in predicting stroke. However, it is acknowledged that this algorithm's effectiveness may be reduced in datasets with highly correlated features. In such scenarios, an alternative approach, such as an L1 regularised feature selection algorithm, could be employed before applying the conservative mean heuristic for more accurate fine-tuning of feature selection, highlighting the potential need for tailored feature selection methods.

Sirsat et al. (2020) reviewed 50 studies published between 2010 and 2019, focusing on ML techniques for brain stroke prediction. Their analysis classified these techniques into four categories and determined that the SVM classifier and RF classifier emerged as the optimal models for this purpose. This finding underscores the potential of SVM and RF classifiers as promising approaches for stroke prediction in ML research. Prentzas et al. (2019) introduced a novel approach that utilised argumentation techniques in conjunction with ML to construct explainable artificial intelligence (XAI) models. Their methodology demonstrated superior performance compared to SVM and RF classifiers, surpassing their accuracy. However, compared to probabilistic neural networks (PNN), the latter yielded lower performance with an accuracy of 64%, suggesting the potential effectiveness of incorporating argumentation techniques in ML-based XAI models for improved performance and interpretability. Bandi et al. (2020) introduced a refined iteration of the RF algorithm to analyse and assess stroke risk levels. Their modified algorithm demonstrated improved performance, achieving an impressive accuracy of 96.97%, suggesting the potential of utilising their approach for accurately predicting and assessing the risk levels associated with strokes, which could have significant implications in clinical practice and patient care. Dritsas and Trigka (2022) proposed a novel stacking classifier technique for predicting the risk of stroke, which yielded impressive results with an AUC score of 98.9% and an overall accuracy of 98%. However, it should be noted that their approach may not be suitable for image datasets, as it was not designed or explicitly tested for such data types. Therefore, it is imperative to conduct further research to establish and validate stroke risk prediction models that are specifically optimised for image datasets. Rajora et al. (2021) conducted a comparative analysis of Naïve Bayes, LR, D-tree, RF, and gradient boosting algorithms for predicting stroke risk. Nevertheless, the results obtained from their study revealed that the accuracy of the models could be further enhanced through meticulous fine-tuning of algorithm parameters or hyperparameters. This implies that additional optimisation endeavours may yield superior performance in predicting stroke risk with heightened accuracy.

Zhang et al. (2021) underscored the promising potential of DL techniques in advancing the field of brain stroke diagnosis and prediction. Their findings shed light on the significant contributions of DL in enhancing the accuracy and efficiency of stroke diagnosis and risk prediction, suggesting the potential for further advancements in this area of research. Heo et al. (2020) employed natural language processing (NLP)-based ML and DL techniques to predict stroke based on brain MRI text reports. However, the model is limited to text in English and has not been validated, potentially indicating the need for further research and validation before widespread application in clinical settings. Shashank et al. (2020) utilised D-tree, Naïve Bayes, and NN algorithms to identify patients at risk of stroke. This project has the potential to be extended for further exploration of stroke probabilities, potentially opening up new avenues for research in this area. Amann (2022) conducted a comprehensive survey of current practices to reduce stroke incidence and mortalities. The study shed light on challenges related to data gathering, app development, and deployment in practice, providing valuable insights for further research and development in this field.

Tazin et al. (2021) employed various ML algorithms, including LR, D-tree classification, RF classification, and voting classifier, to predict stroke risk. Among them, random forest classifier demonstrated superior performance with 96% accuracy. The study further suggested that the accuracy could be enhanced by incorporating AdaBoost, SVM, and bagging techniques into the model. Krittanawong et al. (2020) conducted a comparative analysis of multiple ML and DL algorithms for stroke prognostic. The findings revealed that support vector machine (SVM) outperformed other algorithms, achieving an AUC of 0.92. Al-Zebari and Sengur (2019) utilised a variety of ML algorithms, including D-trees, LR, discriminant analysis (DA), SVM, KNN, and ensemble learners for stroke prediction. The results indicated that the LR method attained the highest accuracy score of 77.9%, outperforming other algorithms in the study. Alaka et al. (2020) employed a variety of ML algorithms, including RF, classification and regression tree (CART), SVM, ABM, LASSO regression, and LR models for stroke prediction. The achieved performance metrics included an AUC range of [0.66–0.71] and an MCC range of [0.34–0.42]. Wu and Fang (2020) evaluated various ML algorithms for predicting stroke risk. The findings revealed that the RF algorithm exhibited predictive accuracy, with an AUC of 0.768. Asadi et al. (2014) proposed a model based on a D-tree algorithm for stroke prediction. The results demonstrated that the model exhibited good predictive performance and could potentially assist clinicians in making improved decisions about patient care. Lin et al. (2020) implemented several ML algorithms, including SVM, RF, and ANN, and also evaluated a hybrid artificial neural network using a ten-fold CV approach. The results suggested that these models achieved a high AUC of 0.94 predicting stroke, indicating their potential for accurate stroke risk prediction. Arslan et al. (2016) introduced a medical data mining approach for predicting brain stroke. The study employed three data mining algorithms, namely C4.5 D-tree, KNN, and SVM, to forecast the likelihood of stroke. The findings revealed that the SVM algorithm exhibited superior accuracy, true positive rate (TPR), and specificity performance compared to the other two algorithms, highlighting its potential for accurately predicting ischemic stroke. The existing major studies in the concerned area are given in Table 1.

In conclusion, ML algorithms have exhibited promising potential in prognosticating stroke risk by analysing extensive datasets and discerning patterns among risk factors.

Various ML-based approaches have been investigated for stroke risk prediction, including SVM classifier, RF classifier, Naïve Bayes, LR, D-tree, and gradient boosting algorithms. Furthermore, cutting-edge DL techniques such as NLP-based ML and DL have also shown promise in augmenting accuracy and efficiency in stroke diagnosis and risk prediction. Nevertheless, challenges related to data acquisition, application development, and real-world deployment must be addressed to reduce stroke incidence and associated mortality rates. The main objective of this research paper is to perform an extensive comparative analysis of advanced classification algorithms employed for predicting stroke occurrences. Through meticulous evaluation of their strengths, limitations, and performance metrics, our study aims to provide valuable insights into the relative effectiveness of these algorithms in accurately predicting stroke risk. This comparative assessment will involve rigorous benchmarking against established criteria, such as accuracy, TPR, TNR, and AUC score, to discern their respective merits and drawbacks. Such a systematic comparison will facilitate the identification of the most suitable algorithm(s) for precise and reliable stroke prediction, thereby contributing to advancing stroke research and clinical practice.

Table 1 Existing major studies

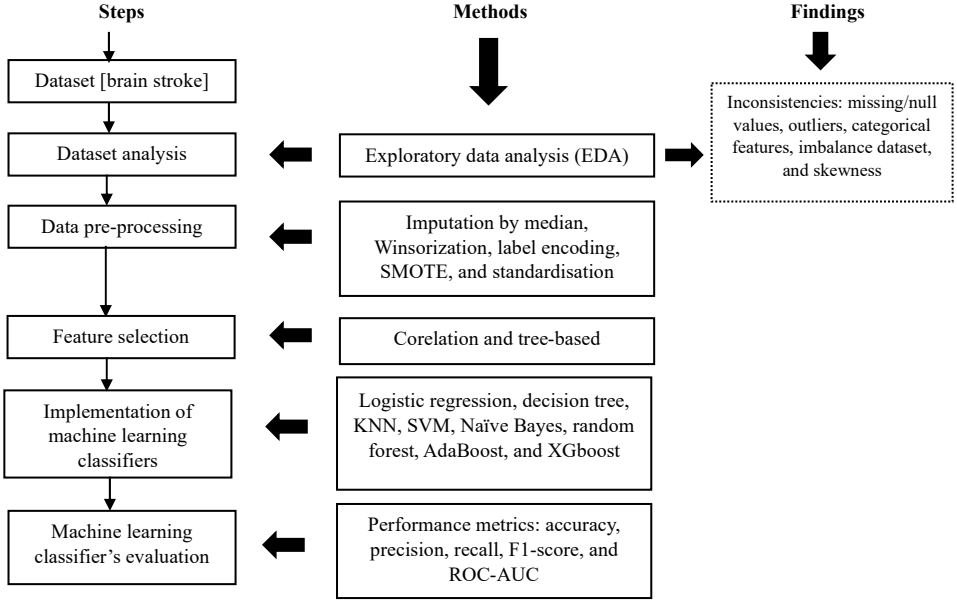
Authors	*Classification techniques*
Emon et al. (2020)	Voting, SGD, D-tree, AdaBoost, Gaussian classifier, QDA, MLP, and KNN
Hung et al. (2017)	DNN and GBDT
Sirsat et al. (2020)	Support vector machine and random forest
Dritsas and Trigka (2022)	Stacking classifier
Rajora et al. (2021)	Naïve Bayes, LR, D-tree, RF and gradient boosting
Zhang et al. (2021)	CNN, RNN, LSTM, and GRU
Heo et al. (2020)	Multi-CNN and CNN
Shashank et al. (2020)	D-tree, Naïve Bayes, and NN
Tazin et al. (2021)	LR, D-tree, RF, and voting classifier
Krittanawong et al. (2020)	SVM and boosting algorithms
Al-Zebari and Sengur (2019)	D-trees, LR, DA, SVM, KNN, and ensemble learners
Alaka et al. (2020)	RF, CART, C5.0 D-tree, SVM, ABM, LASSO, and LR
Wu and Fang (2020)	LR, D-tree, RF, Naïve Bayes, and SVM
Asadi et al. (2014)	LR, RF SVM, KNN, and ANN
Lin et al. (2020)	SVM and SGB
Arslan et al. (2016)	D-tree, Naïve Bayes, KNN, LR, RF, and SVM

3 Methodology

This subsection showcases a visual representation, in the form of a flowchart, illustrating the methodology employed in this paper. The flowchart in Figure 1 serves as a graphical depiction of the step-by-step approach undertaken by the researchers to conduct their study. It elucidates the systematic procedures, techniques, and processes implemented

throughout the research process, offering a clear and concise overview of the methodology utilised in this paper.

Figure 1 Stepwise procedure adopted for this research



4 Experimental setup

This subsection outlines the adherence to the seven-step data science process employed to effectively tackle the given problem statement. By adhering to this well-defined process, the chances of achieving success were significantly maximised, while also ensuring that the work conducted remained in alignment with the goals and expectations of the project.

4.1 Dataset description

The initial step involved clearly defining the study's objective. Subsequently, a dataset was gathered from Kaggle, a comprehensive data repository, and the corresponding code. This dataset assessed a patient's stroke risk based on input characteristics such as gender, age, and various illnesses, including heart disease and hypertension. Each row of the dataset contained relevant patient data. It contains 5,110 records and 12 fields (Sailasya and Kumari, 2021). The description of the dataset taken is provided in Table 2.

4.2 Exploratory data analysis

The exploratory data analysis phase was then conducted to explore and analyse the dataset. This phase aimed to identify patterns, distributions, and statistical measures through visualisations, leading to valuable inferences. During the EDA phase, a

meticulous examination of the data was carried out, encompassing various statistical techniques and visualisations. This process enabled the identification of data distribution, outliers, missing values, and correlations between variables. By leveraging advanced statistical methods and employing sophisticated visualisation techniques, hidden relationships and key patterns were uncovered that may influence the subsequent steps of the data science process.

Table 2 Stroke prediction dataset

S. no.	*Predictors*	*Reference values*
1	gender	Female/male
2	age	0–82
3	hypertension	0/1
4	heart_disease	0/1
5	ever_married	Yes/no
6	work_type	Private/self-employed/children/never_worked
7	Residence_type	Urban/rural
8	avg_glucose_level	55.12–271.74
9	bmi	10.3–97.6
10	smoking_status	Formerly smoked/never smoked/smokes/unknown
11	stroke	0/1
12	id	-

4.3 Data pre-processing

Next, the data pre-processing phase was undertaken to clean and process the raw data before feeding it into the model. The data may be incomplete, inconsistent, or invalid, which could lead to erroneous findings. Thus, data processing was necessary to enhance model accuracy. This involved manipulating and transforming the data, handling missing values and outliers, and numerically encoding categorical features, as many ML models only accept numerical data. Additionally, data scaling techniques were applied to normalise the range of distinct features in the dataset. Synthetic minority over-sampling technique (SMOTE) was used to handle the imbalance of the dataset.

4.4 Feature selection

The next step in the process was feature selection, which entailed selecting critical features while discarding unnecessary ones that did not significantly contribute to the output variable. This was done by assessing the contribution of independent features to the target feature and finding the correlation between the features.

4.5 Implementation of existing classification algorithms for brain stroke prediction

In the subsequent section, we shall delve into an array of sophisticated classification algorithms that can be harnessed for stroke prediction. Through their advanced

computational techniques and statistical algorithms, these powerful tools offer promising avenues for accurate and reliable stroke risk prediction, thus paving the way for timely and targeted preventive interventions. The brief description of the eight ML classifiers taken up for their performance evaluation in this study is given here.

4.5.1 Logistic regression

It is a supervised algorithm used to solve binary classification problems. The output variable must be discrete or categorical. The input features should be linearly separable. This algorithm assumes that the independent features should not be multi-collinear, i.e., the independent features should be correlated. The target is to fit an S-shaped Sigmoid or Logistic curve to the given data points by updating the coefficient until the error is reduced. The approach utilises the sigmoid activation function to map any real value ranging from 0 to 1. Further, it uses a threshold value that classifies predicted values above the threshold as one and below the threshold as zero (LaValley, 2008). This algorithm is easy to implement and is efficient for linearly separable data. However, in real-world problems, data is rarely linearly separable. Also, it may lead to overfitting when the dataset is large (Kleinbaum et al., 2002).

4.5.2 K-nearest neighbours

KNN is an elementary nonlinear classification ML algorithm based on the proximity between the data points. The algorithm is characterised as non-parametric, as it abstains from making assumptions about the underlying dataset's distribution or structure. Instead of learning during the training phase, the algorithm stores the data points for future use during the testing phase, earning the moniker lazy learner. To categorise a new data point using KNN, the algorithm calculates the distance between the new data point and the existing data points using various measures such as cosine, Manhattan, Euclidean, and others. This distance-based approach allows the algorithm to determine the appropriate class label without updating the model during training. Then, these distances are arranged in ascending order, and the first k distances are selected. Majority voting is applied to these points to classify the new data in a category. It is tricky to choose the value of k (Peterson, 2009). Therefore, the selection of k requires domain knowledge. This algorithm is easy to implement and effectively deals with smaller datasets. It also has some constraints. It is not robust to outliers, and the computation cost is high as the algorithm aims to calculate the distance between data points (Prasad et al., 2019).

4.5.3 Support vector classifier

SVM is a powerful statistical approach that aims to find the optimal hyperplane for a training dataset by maximising the margin between the two categories, known as support vectors. The margin, which represents the proximity between the support vectors and the hyperplane, is sought to be maximised to improve classification accuracy. While linear SVM assumes that the categories can be separated by a straight line, the reality is that many real-world datasets are not linearly separable. To overcome this limitation, SVM supports various kernels, such as polynomial, RBF, and Sigmoid, which enable SVM to map datasets to higher-dimensional spaces and find more complex decision boundaries. Selecting a kernel is part of hyperparametric tuning (Noble, 2006). This algorithm is

efficient for linearly separable datasets or datasets with high dimensionality. SVM does not work well if the target classes overlap. Also, its efficacy reduces with large datasets (Meyer and Wien, 2001).

4.5.4 Decision tree classifier

D-tree classifier is a nonlinear supervised algorithm with a tree-like structure. The internal nodes, also known as decision nodes, represent the features or attributes of a dataset. The branches emanating from these decision nodes depict the possible outcomes or choices based on the values of these features. Finally, the leaf nodes of the decision tree represent the class labels or predicted outcomes for the instances that reach that particular node after traversing the decision path from the root to the leaf. The tree begins at the root node, which represents a feature from the dataset and is split into several branches according to the categories in the attribute. The splitting technique followed by the decision nodes is the same as the root node. The split at the decision nodes should be pure for a good model. There are two metrics to check the purity of a split, namely, entropy and the Gini Index. Entropy refers to the randomness in a dataset. The entropy for a pure split is 0. The second metric is the Gini Index which also measures the purity of a split. The lower the Gini Index, the purer the split. Entropy requires logarithmic calculations, while the analysis of the Gini Index involves basic mathematical operations. Thus, the time taken to compute entropy is more than the Gini Index. Hence, for large datasets, Gini Index is used to assess splits' purity, while entropy is used for smaller datasets (Tangirala, 2020). The feature to be split is selected by calculating information gain. It measures the information provided by a feature. Information gain is maximised in a decision tree. The attribute whose splitting gives the highest information gain is chosen. The advantage of this model is minimal data pre-processing. The major limitation of this algorithm is that it leads to an overfitting condition. It can be reduced by tuning the hyperparameters like min_sample_split, min_samples_leaf, etc. Another way to avoid overfitting is pruning. It implies removing the unwanted branches of the decision tree (Jijo and Abdulazeez, 2021). The depth of the tree can be specified before the model building, known as pre-pruning, or unnecessary components can be removed after a model building, known as post-pruning.

4.5.5 Naïve Bayes classifier

The algorithm is referred to as a probabilistic supervised learning method commonly employed for classification tasks. It is based on Bayes' theorem, a probability theory principle. This theorem utilises conditional probability to estimate the likelihood of an event happening, given the occurrence of another event. In this algorithm, it is assumed that each input feature is independent and has an equal contribution to the prediction of the target variable. This assumption allows the algorithm to make predictions based on the joint probability of the input features and target variables. This probabilistic approach makes it a popular choice for classification problems where the relationships between features and the target variable must be considered probabilistic. In Gaussian Naïve Bayes classifier, the input features should be normally distributed. Its probability density function plot should be a symmetrical bell-shaped curve about the mean value. This approach has a short training phase. Hence it is fast (Vembandasamy et al., 2015). In contrast, the assumption of the algorithm of independence of features is rarely met in

real-world datasets. It also suffers from the zero-frequency problem, i.e., if a value from a categorical variable is not seen in the training data but appears in the test dataset, it is assigned to 0 for the category (Murphy, 2006).

4.5.6 Random forest classifier

RF algorithm employs an ensemble approach by combining several decision trees trained on subsets of the training dataset, and the final output is evaluated by simple majority rule. The hyperparameters chosen for every decision tree cannot be analysed. Hence, it is a black-box model. Bootstrap aggregation, also called Bagging, is an ensemble technique that selects random samples with the replacement for the training dataset. This technique of row sampling with replacement is called bootstrap. The algorithm gives different samples to multiple models (decision trees) to train them independently to predict the target class. The final prediction is obtained through aggregation, where outcomes from different models are combined using majority voting (Liu et al., 2012). This approach overcomes the issue of overfitting, often associated with decision tree algorithms, by leveraging the collective decision-making of multiple trees. The random forest algorithm gives high accuracy for massive datasets with high dimensionality. It prevents overfitting issues by aggregation of the results. However, the learning phase requires a long time as several decision trees must be trained (Azar et al., 2014).

4.5.7 AdaBoost classifier

AdaBoost, or adaptive boosting, is a boosting ensemble supervised learning algorithm for classification problems. The boosting techniques combine several weak learner algorithms with building a robust learner model. The method is constructing multiple weak models that progressively rectify incorrectly classified outcomes (Emanet et al., 2014). Models are built successively until the error is minimised. The algorithm combines several decision trees with one level known as decision stumps. A weak classifier is trained using the training data based on the weighted samples. Equal weights are assigned at this step to all the data points. The decision stump with the lowest Gini Index is the initial stump.

Further, the performance of the stump is assessed by calculating the total error, based on which the weights are updated. The outcomes that are incorrectly predicted are given higher weights. Next, the data points are divided into buckets based on the new sample weights. The algorithm then chooses random numbers between 0 and 1. The probability of the numbers being selected to fall in the buckets with misclassified data points is high as their bucket size is larger. Hence, most misclassified data points are passed to the next model as training data (An and Kim, 2010).

4.5.8 XGBoost classifier

XGBoost stands for extreme gradient boosting, a boosting ensemble supervised learning technique. It is a sequential modelling algorithm that uses decision stumps. This approach is a combination of boosting techniques and gradient descent. XGBoost uses decision trees and progressively uses if conditions to make decisions at every step. Various parameters can be tweaked to enhance the performance of the model. The quality score or similarity score is a measure that is calculated for the residuals at each split as an attribute

selection measure (Carmona et al., 2022). The various features of XGBoost are regularisation, missing data handling, parallel processing, tree pruning, etc. The algorithm has inbuilt stopping criteria for splitting the decision trees. The two main advantages of the algorithm are its high execution speed and excellent model performance. It saves time by auto tree pruning, cache optimisation, and parallel processing. It can handle missing values and outliers. With their diverse strengths and capabilities, these advanced classification algorithms can significantly reduce the incidence of strokes and improve patient outcomes (Liew et al., 2021).

4.6 Performance evaluation of classifiers

The performance of various classifiers was assessed on various performance metrics as given in Table 2.

Table 2 Performance metrics used in the present research

Metric	*Description/formula*
Accuracy	$\text{Accuracy score} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$
Precision	$\text{Precision score} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
Recall	$\text{Recall score} = \frac{\text{FP}}{\text{FP} + \text{FN}}$
F1-score	$\text{F1-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
ROC-AUC score	The area under the receiver operating characteristic (ROC) curve is the ability of a classification model to distinguish between the categories

Notes: Here, TP is true positive, FP is false positive (type 1 error), FN is false negative (type 2 error), and TN is true negative.

The training and testing accuracy comparative statistics obtained for all eight-ML classification taken up in this study are given in Figure 2 where as performance based on rest of the metrics are presented in Table 3.

5 Results and discussion

The major findings of the present research are discussed in this section.

- Table 3 presents the evaluation metrics, namely precision, recall, and F1-score and AUC score, for various classification algorithms. It is evident that LR, KNN, D-trees, SVC, Naïve Bayes, RF, AdaBoost, and XGBoost classifier perform commendably in accurately identifying instances associated with the negative class (class 0). These algorithms demonstrate high precision, recall, and F1-score for class 0. On the other hand, when it comes to the positive class (class 1), the algorithms generally exhibit lower precision, recall, and F1-score compared to class 0. Consequently, when selecting an algorithm for a specific classification task, it is crucial to consider the performance on both classes.

Figure 2 Training and testing accuracy comparative statistics for eight classifiers (see online version for colours)

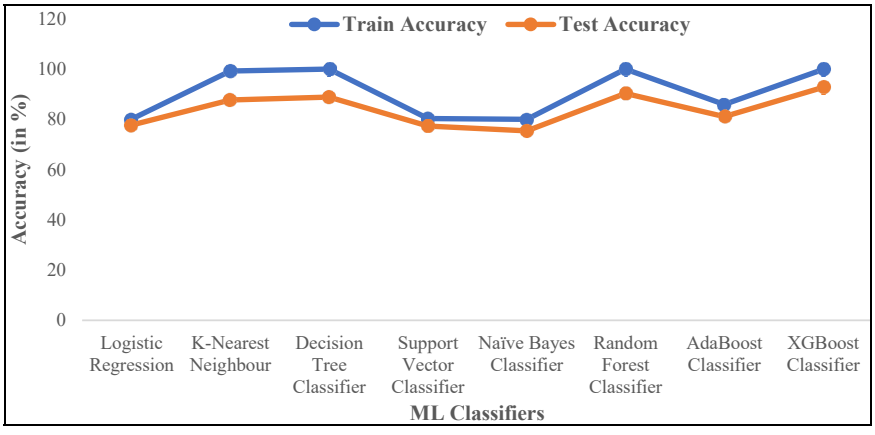


Table 3 Calculated values of statistical parameters for classification techniques

Classification algorithm	*AUC score*	*Precision*	*Recall*	*F1-score*	*Class*
Logistic regression	73.2151	0.97	0.79	0.87	0
		0.12	0.50	0.19	1
K-nearest neighbour	65.1190	0.96	0.91	0.93	0
		0.16	0.30	0.21	1
Decision tree classifier	57.3921	0.96	0.91	0.93	0
		0.16	0.30	0.20	1
Support vector classifier	75.7652	0.97	0.79	0.87	0
		0.13	0.57	0.22	1
Naïve Bayes classifier	75.9374	0.97	0.76	0.85	0
		0.12	0.59	0.20	1
Random forest classifier	71.3652	0.95	0.95	0.95	0
		0.13	0.13	0.13	1
AdaBoost classifier	74.9522	0.96	0.82	0.88	0
		0.12	0..46	0.20	1
XGBoost classifier	*75.8207*	*0.95*	*0.97*	*0.96*	*0*
		0.22	*0.15*	*0.18*	*1*

- Figure 2 in Subsection 4.6 depicts that logistic regression, SVC, and Naïve Bayes classifier have relatively similar train accuracy and test accuracy, indicating consistent performance on training and test data. The decision tree classifier has a train accuracy of 100.00%, suggesting it may have overfitting issues as it achieves perfect accuracy on the training data but lower accuracy on the test data. K-nearest neighbour has a very high train accuracy of 99.2037%. AUC score is relatively lower compared to some other algorithms, suggesting a potential issue of overfitting or lack of generalisation ability on unseen data. Random forest classifier has a high train accuracy of 99.9486%, indicating potential overfitting, but its test accuracy of 90.3131% is still relatively high, suggesting good generalisation ability. AdaBoost

classifier shows a moderate improvement in test accuracy compared to the train accuracy, suggesting it may have some ability to generalise well to unseen data. XGBoost classifier also offers promising performance with a test accuracy of 92.7593%, suggesting its potential for accurate stroke risk prediction. Choosing an algorithm for a specific classification task, the performance on both classes should be considered. Hence, XGBoost classifier or RF classifier may be more suitable choices.

6 Conclusions and future scope

The present study aims on the performance comparison of eight different ML classifiers used for brain stroke prediction. The performance statistics depict that XGBoost has emerged as the most promising contender, exhibiting a remarkable test accuracy of 92.7592% and an impressive ROC-AUC score of 75.8207. These superior performance metrics can be attributed to XGBoost's incorporation of advanced features such as regularisation, parallel processing, tree-pruning, integrated cross-validation, and robust handling of missing values. Collectively, these state-of-the-art capabilities endow XGBoost with a distinct competitive advantage, positioning it as the optimal choice for accurate and reliable stroke prediction in formal research and clinical settings.

Furthermore, this research can be extended by integrating custom ensemble techniques to attain even higher accuracy in stroke prediction. Custom ensemble approaches combine or stack multiple weak learner models to create a robust, high-performing learner with boosted accuracy. By leveraging the strengths of various algorithms, the limitations inherent in individual models can be mitigated. Mechanisms such as voting classifiers, weighted averages, and other aggregation techniques can integrate predictions from different models and generate more accurate and reliable stroke risk predictions. This approach holds promise in overcoming the limitations of individual algorithms and enhancing the predictive accuracy of stroke prediction models.

References

Alaka, S.A., Menon, B.K., Brobbey, A., Williamson, T., Goyal, M., Demchuk, A.M. and Sajobi, T.T. (2020) 'Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and regression models', *Frontiers in Neurology*, Vol. 11, p.889.

Al-Zebari, A. and Sengur, A. (2019) 'Performance comparison of machine learning techniques on diabetes disease detection', in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, IEEE, November, pp.1–4.

Amann, J. (2022) 'Machine learning (ML) in stroke medicine: opportunities and challenges for risk prediction and prevention', *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*, pp.57–71.

An, T.K. and Kim, M.H. (2010) 'A new diverse AdaBoost classifier', in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, IEEE, October, Vol. 1, pp.359–363.

Arslan, A.K., Colak, C. and Sarihan, M.E. (2016) 'Different medical data mining approaches based prediction of ischemic stroke', *Computer Methods and Programs in Biomedicine*, Vol. 130, pp.87–92.

Asadi, H., Dowling, R., Yan, B. and Mitchell, P. (2014) 'Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy', *PloS One*, Vol. 9, No. 2, p.e88225.

Azar, A.T., Elshazly, H.I., Hassanien, A.E. and Elkorany, A.M. (2014) 'A random forest classifier for lymph diseases', *Computer Methods and Programs in Biomedicine*, Vol. 113, No. 2, pp.465–473.

Bandi, V., Bhattacharyya, D. and Midhunchakkravarthy, D. (2020) 'Prediction of brain stroke severity using machine learning', *Rev. d'Intelligence Artif.*, Vol. 34, No. 6, pp.753–761.

Carmona, P., Dwekat, A. and Mardawi, Z. (2022) 'No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure', *Research in International Business and Finance*, Vol. 61, p.101649.

Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B. and John, D. (2022) 'A predictive analytics approach for stroke prediction using machine learning and neural networks', *Healthcare Analytics*, Vol. 2, p.100032.

Dritsas, E. and Trigka, M. (2022) 'Stroke risk prediction with machine learning techniques', *Sensors*, Vol. 22, No. 13, p.4670.

Emanet, N., Öz, H.R., Bayram, N. and Delen, D. (2014) 'A comparative analysis of machine learning methods for classification type decision problems in healthcare', *Decision Analytics*, Vol. 1, pp.1–20.

Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, M.M., Al Mamun, M.S. and Kaiser, M.S. (2020) 'Performance analysis of machine learning approaches in stroke prediction', in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, November, pp.1464–1469.

Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P. and Manikandan, R. (2020) 'Classification of stroke disease using machine learning algorithms', *Neural Computing and Applications*, Vol. 32, pp.817–828.

Heo, T.S., Kim, Y.S., Choi, J.M., Jeong, Y.S., Seo, S.Y., Lee, J.H. and Kim, C. (2020) 'Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI', *Journal of Personalized Medicine*, Vol. 10, No. 4, p.286.

Hung, C.Y., Chen, W.C., Lai, P.T., Lin, C.H. and Lee, C.C. (2017) 'Comparing deep neural networks and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database', in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, July, pp.3110–3113.

Jijo, B.T. and Abdulazeez, A.M. (2021) 'Classification based on decision tree algorithm for machine learning', *Journal of Applied Science and Technology Trends*, Vol. 2. pp.20–28.

Khosla, A., Cao, Y., Lin, C.C.Y., Chiu, H.K., Hu, J. and Lee, H. (2010) 'An integrated machine learning approach to stroke prediction', in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July, pp.183–192.

Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M. (2002) *Logistic Regression*, p.536, Springer-Verlag, New York.

Krittanawong, C., Virk, H.U.H., Bangalore, S., Wang, Z., Johnson, K.W., Pinotti, R. and Tang, W.W. (2020) 'Machine learning prediction in cardiovascular diseases: a meta-analysis', *Scientific Reports*, Vol. 10, No. 1, p.16057.

LaValley, M.P. (2008) 'Logistic regression', *Circulation*, Vol. 117, No. 18, pp.2395–2399.

Liew, X.Y., Hameed, N. and Clos, J. (2021) 'An investigation of XGBoost-based algorithm for breast cancer classification', *Machine Learning with Applications*, Vol. 6, p.100154.

Lin, C.H., Hsu, K.C., Johnson, K.R., Fann, Y.C., Tsai, C.H., Sun, Y.U. and Taiwan Stroke Registry Investigators (2020) 'Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry', *Computer Methods and Programs in Biomedicine*, Vol. 190, p.105381.

Liu, Y., Wang, Y. and Zhang, J. (2012) 'New machine learning algorithm: random forest', in *Information Computing and Applications: Third International Conference, ICICA 2012, Proceedings*, Springer, Berlin, Heidelberg, Chengde, China, 14–16 September, No. 3, pp.246–252.

Meyer, D. and Wien, F.T. (2001) 'Support vector machines', *R News*, Vol. 1, No. 3, pp.23–26.

Murphy, K.P. (2006) 'Naïve Bayes classifiers', *University of British Columbia*, Vol. 18, No. 60, pp.1–8.

Noble, W.S. (2006) 'What is a support vector machine?', *Nature Biotechnology*, Vol. 24, No. 12, pp.1565–1567.

Peterson, L.E. (2009) 'K-nearest neighbour', *Scholarpedia*, Vol. 4, No. 2, p.1883.

Prasad, D., Goyal, S.K., Sharma, A., Bindal, A. and Kushwah, V.S. (2019) 'System model for predictive analytics using k-nearest neighbours algorithm', *Journal of Computational and Theoretical Nanoscience*, Vol. 16, No. 10, pp.4425–4430.

Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A. and Pattichis, C. (2019) 'Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction', in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, IEEE, October, pp.817–821.

Rajora, M., Rathod, M. and Naik, N.S. (2021) 'Stroke prediction using machine learning in a distributed environment', in *International Conference on Distributed Computing and Internet Technology*, Springer, Cham, January, pp.238–252.

Saber, H., Somai, M., Rajah, G.B., Scalzo, F. and Liebeskind, D.S. (2019) 'Predictive analytics and machine learning in stroke and neurovascular medicine', *Neurological Research*, Vol. 41, No. 8, pp.681–690.

Sailasya, G. and Kumari, G.L.A. (2021) 'Analyzing the performance of stroke prediction using ML classification algorithms', *International Journal of Advanced Computer Science and Applications*, Vol. 12, No. 6.

Sandhu, J.K., Lilhore, U.K., Poongodi, M., Kaur, N., Band, S.S., Hamdi, M. and Mosavi, A.H. (2022) *Predicting the Risk of Heart Failure Based on Clinical Data*, Vol. 57, DOI: 10.22967/HCIS.2022.12.057.

Shashank, H.N., Srikanth, S. and Thejas, A.M. (2020) *Prediction of Stroke using Machine Learning*, Doctoral dissertation, CMR Institute of Technology. Bangalore.

Sirsat, M.S., Fermé, E. and Câmara, J. (2020) 'Machine learning for brain stroke: a review', *Journal of Stroke and Cerebrovascular Diseases*, Vol. 29, No. 10, p.105162.

Tangirala, S. (2020) 'Evaluating the impact of Gini Index and information gain on classification using decision tree classifier algorithm', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 2, pp.612–619.

Tazin, T., Alam, M.N., Dola, N.N., Bari, M.S., Bourouis, S. and Khan, M.M. (2021) 'Stroke disease detection and prediction using robust learning approaches', *Journal of Healthcare Engineering*, Vol. 2021, Article ID 7633381, 12pp.

Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) 'Heart disease detection using Naïve Bayes algorithm', *International Journal of Innovative Science, Engineering & Technology*, Vol. 2, No. 9, pp.441–444.

Wolfe, C.D. (2000) 'The impact of stroke', *British Medical Bulletin*, Vol. 56, No. 2, pp.275–286.

Wu, Y. and Fang, Y. (2020) 'Stroke prediction with machine learning methods among older Chinese', *International Journal of Environmental Research and Public Health*, Vol. 17, No. 6, p.1828.

Zhang, S., Zhang, M., Ma, S., Wang, Q., Qu, Y., Sun, Z. and Yang, T. (2021) 'Research progress of deep learning in the diagnosis and prevention of stroke', *BioMed Research International*, ID 5213550.

Abbreviations

AdaBoost	Adaptive boosting
ANN	Artificial neural network
AUC	Area under the curve
CART	Classification and regression tree
CV	Cross validation
DA	Discriminant analysis
DL	Deep learning
DNN	Deep neural network
D-tree	Decision tree
GBDT	Gradient boosting decision tree
KNN	K-nearest neighbours
MRI	Magnetic resonance imaging
LR	Logistic regression
MCC	Matthews correlation coefficient
ML	Machine learning
LASSO	Least absolute shrinkage and selection operator
SGD	Stochastic gradient descent
SVM	Support vector machine
TNR	True negative rate
TPR	True positive rate
UAR	Unweighted average recall
WHO	World Health Organization
XAI	Explainable artificial intelligence
XGBoost	Extreme gradient boosting
NN	Neural network
PNN	Probabilistic neural network
QDA	Quadratic discriminant analysis
RBF	Radial basis function
RF	Random forest
NLP	Natural language processing
ROC-AUC score	Receiver operating characteristic-area under the curve score
SMOTE	Synthetic minority over sampling technique