scientific reports

OPEN

Optimizing brain stroke detection with a weighted voting ensemble machine learning model

Reeta Samuel & Thanapal Pandi✉

Brain stroke is a medical trauma that occurs when there is an impairment or decrease in blood circulation to a particular part of the brain, causing adjacent brain cells to die. Stroke diagnosis after an event is an ineffective method; other more labour-intensive and costly procedures exist for stroke diagnosis. This method involves directing a machine learning algorithm to a marked dataset to identify samples and irregularities indicative of stroke occurrence. This study focused on developing an ensemble machine learning model to predict brain stroke. The model combined the predictions of multiple individualistic classifiers, including random forest, eXtreme gradient boosting, and histogram-based gradient boosting, to improve accuracy. The proposed weighted voting-based ensemble (WVE) classifier model achieved an accuracy of 92.31% on a private stroke prediction dataset. The pre-assessment of stroke risk diagnosis, as suggested in this study, enables many people to take preventive actions well in advance, thereby lowering fatal effects. Our proposed method presents a feasible option for the early or initial diagnosis of stroke, as traditional methods, such as computed tomography (CT) scans and magnetic resonance imaging (MRIs), are time-consuming and costly. Future research could explore the use of intelligence-based optimization to enhance classification accuracy and address this limitation.

Keywords Brain stroke detection, EXtreme gradient boosting (XGBoost), Histogram-based gradient boosting (HGB), Machine learning (ML), Random forest (RF), Weighted voting

Stroke continues to be the second most common cause of mortality worldwide and is the primary contributor to long-term neurological impairment[1]. It is a primary contributor to global disability among major muscle-related disorders and holds a prominent position among the top three disorders. Cerebrovascular disease (CVD), characterized by stroke, is a cause of death and morbidity[2]. Moreover, 5 million people worldwide are chronically paralyzed due to stroke, which affects approximately 15 million people annually (Organization, 2015). Brain ischemic stroke, brain hemorrhage, and other severe brain traumas are caused by cerebral vascular disease, which arises from damaged brain blood vessels[3]. There are two categories of stroke: hemorrhagic and ischemic, both of which represent distinct types of occurrences[4]. Ischemic stroke occurs due to blood vessel blockage, whereas hemorrhagic stroke results from blood vessel rupture.

A momentary ischemic attack (TIA) is a form of ischemic stroke characterized by transient blood vessel obstruction[5]. This type of stroke does not cause long-term brain damage and lasts for no more than five hours, in contrast to ischemic stroke[6]. Ischemia or hemorrhage in the brain arteries causes stroke, which is also known as cerebrovascular injury. Stroke can result in many types of physical and cognitive impairments that risk functionality[7]. Brain stroke localization and identification are overly complex tasks that require an accurate affinity for the manner and location of the stroke to implement applicable behavioural interventions[8].

The prevalent cerebrovascular condition known as ischemic stroke (IS) or cerebral infarction is mostly caused by thrombi obstructing cerebral blood arteries, which results in ischemia and hypoxic necrosis of the brain tissue[9].

An ischemic stroke occurs when blood clots cause the brain's blood flow to halt too slowly[10]. After ischemic stroke, patients may also experience stroke bleeding, which can be a dangerous consequence of the disease[11]. After a stroke, the first month of recovery is quick and easy, but the next three to six months are slower[12].

However, bleeding occurs when a stroke occurs, blood leaks into the surrounding brain tissues, or blood vessels burst owing to their rigidity[13]. The most common causes of hemorrhagic brain stroke are bleeding diseases, aneurysms, arteriovenous disorders, hypertension, and injuries[1]. Hemorrhagic stroke is a severe condition with a high risk of morbidity and mortality (Banjan et al., 2023). AI in radiology, particularly in

School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore632 014, Tamil Nadu, India. ✉email: thanapal.p@vit.ac.in

computer vision and deep learning tasks, is gaining attention owing to advancements such as AlexNet (Liu, X et al.,2021).We also covered several issues in this study, along with potential fixes that need to be explored in further research. The primary outcomes of this study were as follows:

- Proposed a new approach for combining machine learning models using the Weighted Voting-based Ensemble
- (WVE) classifier model.
- Explored a neurological method for classifying brain stroke.
- Employed private data and techniques for diagnosis.
- Then compared our proposed model to various machine learning methods for stroke detection.
- Provides a comprehensive understanding of the patient's overall health and potential risk factors.
- Determined and examined the obstacles that still need to be solved before assessing the prospects for this field of study.

Motivation for proposed model.

1. Stroke data are complex and diverse, making it difficult for single models to capture all patterns.
2. Basic ensemble methods may not efficiently combine the strengths of these models.
3. Existing methods do not identify risks.
4. Working with private medical data while preserving privacy is a challenge.
5. Single-classifier approaches struggle to capture the complex nature of stroke risk factors.

The Weighted Voting-based Ensemble Classifier model offers several advantages, including achieving 92.31% accuracy, surpassing earlier single-classifier methods. It effectively captures the complex relationships among stroke risk factors, serves as a cost-effective alternative to expensive neuroimaging for diverse healthcare settings, and prioritizes clinical utility with interpretable risk assessments for preventive intervention. Additionally, the model employs an innovative weighted voting mechanism that dynamically adjusts the influence based on the performance.

Many classification algorithms have been developed in recent decades owing to the vital nature of robotized classification of pictures produced by Magnetic Resonance Imaging (MRI)[14]. It is essential to study and analyze the human mind[15]. MRI's rich data on delicate tissue living structures have greatly improved our understanding of cerebrum pathology and its remedies[16]. Figure 1 illustrates the risk factors for stroke.

Related work

Stroke is a significant global health issue, and early and precise detection is crucial for successful treatment. Machine learning has been used to organize brain foci using medical images that resemble CT and MRI scans. One approach involves extracting features from these images, such as texture and shape, and using them to guide machine learning paradigms, such as support vector machines, k-nearest neighbors, or decision trees. Several studies have shown that machine learning methods can effectively classify brain strokes Wang et al.[17]. For instance, a study that used CT scans and support vector machines was able to differentiate between ischemic and hemorrhagic strokes with an accuracy of 85.7%. Another study that used MRI images and convolutional neural networks to group infarctions and edema achieved an accuracy of 94.2%[18].

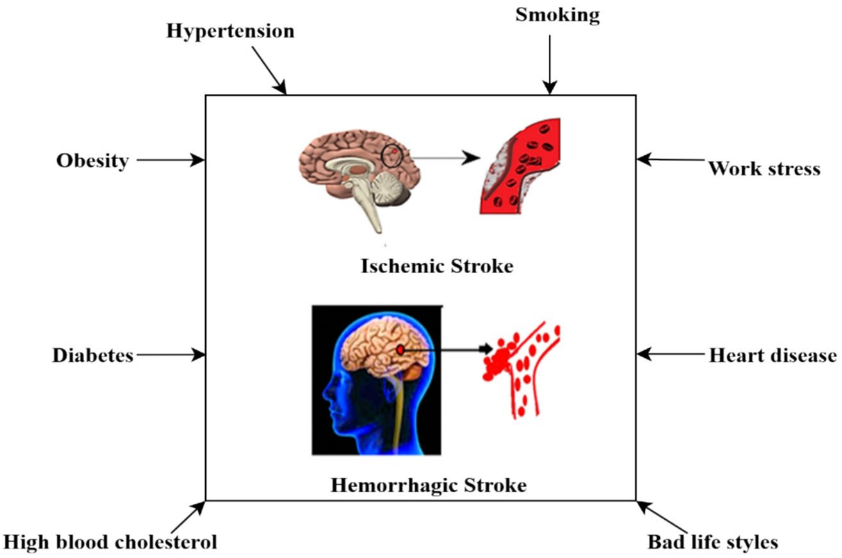


Fig. 1. Risk factors for stroke.

Jayachitra and Prasanth developed a new method for segmenting ischemic stroke lesions using fuzzy logic. After segmentation, they extracted features from the segmented regions and used them to train a weighted Gaussian Naive Bayes classifier. Their approach outperformed existing advanced methods in terms of accuracy[19].

Preprocessing, feature extraction, feature reduction, and brain MRI image classification are the four stages of the conventional classification model. Among all stages of the classification model, preprocessing is the easiest. A noise reduction algorithm was utilized during the preprocessing step to eliminate undesired elements from the images, such as the scalp and skull, and salt-and-pepper noise. The quality of the images was enhanced by removing noise[20].

Ultrasound (US) is one of the most used medical diagnostic methods and is an essential tool for medical imaging. It is crucial for both qualitative and quantitative disease classification and, medical assessments. The advantages of ultrasound imaging over other medical imaging technologies include its low cost, ease of use, lack of ionizing radioactivity, great understanding, and simultaneous imaging. Nonetheless, US imaging has certain issues compared with CT, MRI, and X-ray imaging. For example, increased noise and artifacts and reduced tissue contrast lead to boundary challenges. (Wu, L et al., 2017).

Prior research has used machine learning techniques to forecast motor and working healing during the critical and subacute cycles of stroke[21]. Machine learning models have been used to predict stroke patients' recovery of motor or intellectual abilities[22]. Researchers have developed various screening techniques to improve the effectiveness of stroke screening, which can be broadly categorized into four categories. By implementing revisions to the detection evaluation form for the stroke population, the screening efficiency was enhanced[23].

Hung et al. extended a deep convolutional network representation using Taiwan's stroke diagnosis record (EMC) and contrasted its outcomes with those of conventional machine learning techniques. The AUC of 0.92 represented the highest accuracy between the two divisions (stroke or no stroke). Although the trial range in this study was rather large and could identify strokes with some precision, the overall result was not favourable[24].

S. J. Hegland et al. developed a predictive model for acute strokes using CNN and brain MRI data. The best result for the deep convolutional network for ischemic stroke prediction was an AUC of 0.88 ± 0.12[25]. Immediately after stroke, brain impairment results from blood flow disruption[26].

This study aimed to enhance the effectiveness of brain–computer interface (BCI) technology for stroke recovery. The researchers created a new method for recognizing a patient's intention to move, utilizing a technique called"time series shapelets."This innovative approach demonstrated superior performance compared to existing methods in both offline analysis and simulated real-time scenarios[27].

Z. Gong et al., developed a new approach to microwave medical sensing (MMS) is presented, enabling rapid and accurate classification and localization of strokes. By dividing the examination area into sections and utilizing decision tree learning, this method efficiently identifies the stroke characteristics. Compared with traditional methods, it achieves a 14.1% and 19.2% improvement in classification rates, reduces localization time by 21.1%, and attains a localization accuracy of over 0.90. This innovative space-division-based technique is particularly suitable for wearable devices and offers a promising solution for localizing brain strokes without the need for imaging[28].

This research explored the application of Huygens' principle (HP) imaging for stroke detection in the brain. Although the intricate structure of the brain poses challenges, recent advancements in artificial intelligence and deep learning have enhanced the accuracy of stroke detection and classification. Through simulations using the Finite Difference Time Domain (FDTD) method, we revealed that combining the magnitude and phase information from HP imaging improves stroke detection and classification accuracy. The proposed approach was validated using real-world data from two patients[29]. Table 1 presents a comparison of the existing models.

Proposed model

We propose a Weighted Voting-based ensemble classifier and compare it with seven common machine learning algorithms: Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), gradient tree boosting (GB), K-Nearest Neighbor (KNN), and Naive Bayes (NB) and employed a voting classifier in this study.

A Voting classifier combines multiple models to make predictions by choosing the class with the maximum possibility. It is often used to forecast outcomes, such as voting results. Because the weighted voting ensemble considers each base classifier's optimism in its prediction, compared to hard voting, which sums the number of times each model has been recognized by crucial classifiers, it is usually considered more accurate and dependable. It is an easy-to-use technique that can be used to improve the performance of a machine-learning model in both classification and regression scenarios. Figure 2 illustrates the proposed model.

Three basic classifiers are utilized in the proposed WVE classifier: Random Forest, eXtreme gradient boosting, and histogram-based gradient boosting. The nomenclature is presented in Table 2. The following is an overview of each base classifiers.

Random forest

Random Forest (RF), a supervised learning technique, is used for regression and classification tasks. It consists of decision trees (also called "forest"), bagging, feature randomness, and voting. It can handle high-dimensional data, prevent overfitting, manage missing values, and is interpretable.

The hyperparameters control the ensemble size, depth, features, and sample split. Common applications include image classification and regression. At the Random Forest level, the average feature importance across all trees was the final measure of significance. The importance value of each characteristic of each tree was added together and divided by the total number of trees. For regression, a random forest prediction was used.

Key technique	Model	Research performance	Limitation	References
Machine learning and deep learning	CNN, LSTM, KNN, XGB, and majority voting ensemble	Proposed model obtained the highest classification performance based on all evaluation metrics on all datasets	Potential limitations in generalizability to other populations or datasets, need for further validation, May require significant computational resources	30
Deep learning	CNN-GRU, SMOTE Method	Higher classification accuracy compared to other existing models	Potential limitations in generalizability to other datasets or environments	31
Ensemble learning, data mining techniques	Weighted ensemble model using genetic algorithm	Improved performance compared to individual classifiers	May require significant computational resources	32
Remote monitoring	Web application for remote monitoring and management, Real-time monitoring and alerts	Effective in monitoring and managing high-risk pregnancies	Limited to healthcare professionals, not designed for patient use	32
Ensemble-based deep learning model	CNN, LSTM, XGBoost, KNN	Outperformed existing models, demonstrating superiority in cardiovascular disease prediction	Lack of interpretability of the model's predictions due to the complexity of the ensemble architecture	33
Semantic relatedness and similarity measures	natural language, machine learning algorithms	Using students' answers as feedback considerably improved the accuracy and performance of these measures	The dataset used is relatively small	34
Machine learning	Neural networks, SVM, KNN	remarkable accuracy and minimal loss	Limited to a single dataset, potential variation with other datasets	35
Machine learning	Nomogram prediction model	Successfully identified several parameters associated with stroke risk, demonstrated superior predictive accuracy	Potential limitations in generalizability to other populations, need for further validation	36
Machine learning (ML)	Random forest (RF), KNN, DT, AdaBoost, XGBoost, SVM, ANN	RF achieved highest performance	Potential limitations in generalizability to other populations or datasets, need for further validation, May require significant computational resources	37
Ensemble Machine Learning	Soft Voting Classifier (Random Forest, Extremely Randomized Trees, Histogram-Based Gradient Boosting)	Achieved an accuracy of 96.88%, improved accuracy and robustness compared to single classifiers	Potential limitations in handling complex interactions between features, need for further optimization	18
Face Detection using Yolo v8	Stroke monitoring strategy	Achieved high accuracy of 98.43%	Limited availability of stroke patient data	38
Modified Vision Transformer (ViT) integrated approach	End to end ViT Architecture, CNN	87.51% classification accuracy for brain CT scan slices	Improvement needed for stroke diagnosis	43
A deep-learning-based Microwave-induced thermo acoustic tomography MITAT (DL-MITAT) Technique	A residual attention U-Net (ResAttU-Net)	effectively eliminated image artifacts and accurately restored hemorrhage spots as small as 3 mm	No performance metrics for increased accuracy; training sets are constructed only using the simulation approach	44
AutoML	A combination of AutoML, Vision Transformers (ViT), and CNN	The model achieved 87% accuracy for single-slice level predictions and 92% accuracy for patient-wise predictions	Small sample size, complexity of the integrated architecture	45

Table 1. Comparison of existing model in literature.

$$\widehat{y_i} = \frac{1}{N}\sum_{n=1}^{N} T_n(x_i) \tag{1}$$

$$\widehat{y_i} = mode\left(T_1\left(x_i\right), T_2(x_i), \ldots., T_N(x_i)\right) \tag{2}$$

where N is the sum of trees and $T_n(x_i)$ is the calculation from the n-th tree for input (x_i). For classification, the final prediction is the mode (majority vote) of the class prediction from all trees: where $T_N(x_i)$ is the prediction from the n-th tree for input (x_i). Random Forest builds various individual decision trees and associates them through averaging (regression) or voting (classification). However, Random Forest is not always the best choice, therefore, it is essential to experiment and compare its performance with that of other algorithms. Regarded as one of the most potent and resilient algorithms available, Random Forest is simple to operate and capable of processing a multitude of characteristics and categorical variables. In addition, it is less likely to overfit than a single decision tree. Figure 3 shows the Random Forest graphic depiction.

XGBOOST, or extreme gradient boosting

A very effective and adaptable gradient boosting framework is called Extreme Gradient Boosting (XGBoost). It is intended to outperform conventional gradient boosting techniques in terms of scalability, regularization, and speed. Owing to its exceptional performance, XGBoost has become popular in both real-world applications and machine learning competitions.

Key characteristics of regularization with XGBoost: XGBoost uses L1 and L2 regularization to reduce overfitting and improve generalization. System Optimization: It suited for huge datasets because it is optimized for parallel and distributed computing. Flexibility: XGBoost can process both arithmetic and categorical attributes across a wide range of data formats. Efficiency: It is appropriate for large-scale challenges because of

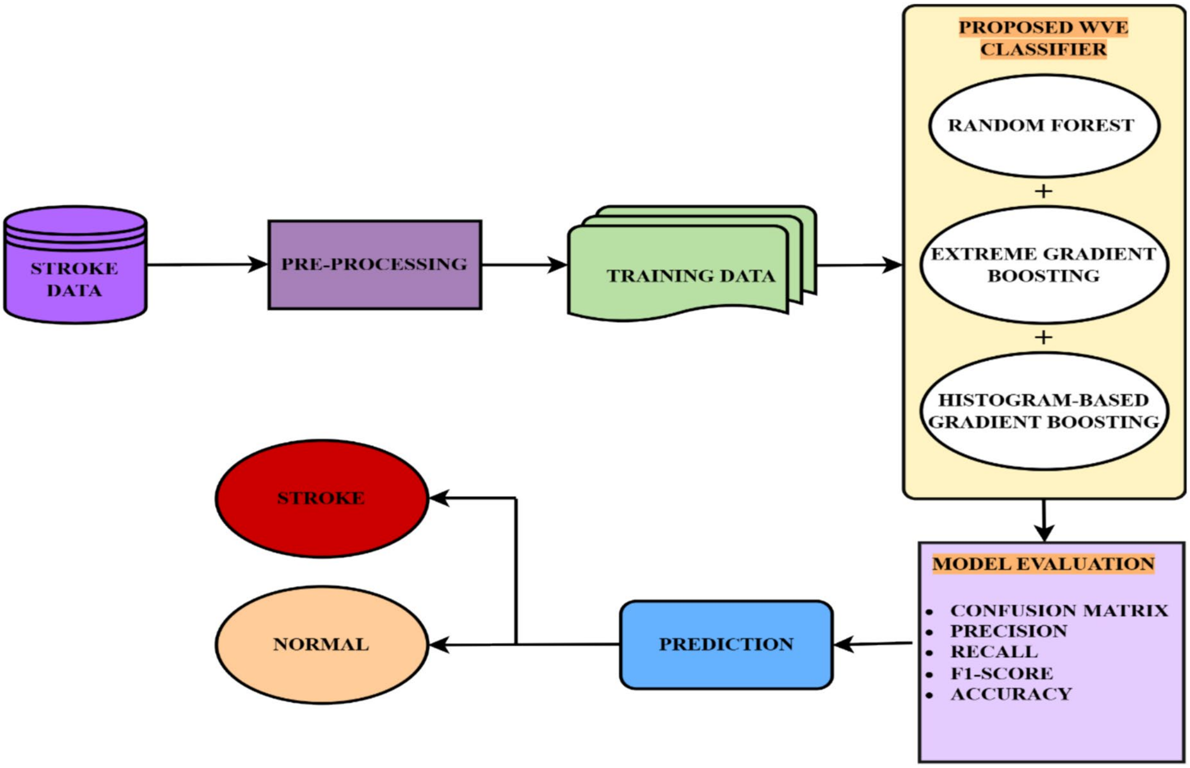


Fig. 2. Proposed weighted voting-based ensemble classifier model.

its computational efficiency. XGBoost contains built-in techniques for handling missing data. Customizable Loss Functions: Custom loss functions can be built to adapt the algorithm to challenges. The regularized objective function with two components is reduced by XGBoost as follows:

$$L\left(\theta\right) = \sum_{i=1}^{n} l(\mathrm{u}_i, \widehat{u_i}) + \sum_{k=1}^{k} \Omega(f_k) \tag{3}$$

where, $l(\mathrm{u}_i, \widehat{u_i})$ is the loss function that measures how well the model fits the data, represents the total number of data points, u_i is the actual observed value for the i$^{\mathrm{th}}$ data point, and $\widehat{u_i}$ is the predicted value for the i$^{\mathrm{th}}$ data point. $\Omega(f_k)$ is the regularization term for the complexity of the k$^{\mathrm{th}}$ tree, which is usually defined as

$$\Omega(f_k) = \gamma^T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

where:

- γ is a regularization parameter that controls the number of leaves in the tree.
- T represents the number of terminal nodes in the tree.
- w_j is the weight of each leaf.
- λ is a regularization parameter for leaf weights. The predicted value for an input x is calculated by adding the outputs of all the trees:

where $f_k\ (x_i)$ is the prediction output of the k$^{\mathrm{th}}$ tree, and K is the total number of trees.

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i) \tag{5}$$

Histogram based gradient boosting

Histogram-Based Gradient Boosting, or HBGB for simple terms, remains the machine learning equal to the Gradient Boosting algorithm. A sophisticated ensemble technique called gradient boosting constructs a model by integrating the predictions from numerous ineffective models, each of which is instructed to fix the mistakes of the previous models. Rather than employing a single decision tree, as in the past, HBGB uses histograms to

Notation	Description
$\widehat{y_i}$	Final predicted output for the ith input sample
x_i	The i^{th} input data sample
$T_n(x_i)$	Predicted class label by the n^{th} model or client for input x_i
$\sum_{n=1}^{N}$	Summation overall N models or clients
$\frac{1}{N}$	Averaging factor to compute the mean prediction from all contributors
N	Total number of models or clients
$mode()$	Statistical mode function that returns the most frequent class label
$L(\theta)$	Total loss function with parameters (θ)
n	Total number of data samples
$l(u_i, \widehat{u_i})$	Loss between ground truth u_i and predicted output $\widehat{u_i}$
$\sum_{i=1}^{n}$	Summation overall n training samples
$\sum_{k=1}^{k}$	Summation over all K model components
$\Omega(f_k)$	Regularization term for the kth model component
(f_k)	Model parameters of the k^{th} component
(θ)	Overall set of model parameters
γ^T	Bias or constant term related to iteration T
λ	Regularization coefficient
T	Total number of training iterations or time steps
w_j^2	Model weight parameter at step j
$\sum_{j=1}^{T} w_j^2$	Sum of squared weights
$\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$	L2 regularization term
$f_k(x_i)$	Output of the k^{th} model when applied to input x_i
K	Total number of models contributing to the aggregation
$\sum_{k=1}^{K}$	Summation over all K models
$\widehat{Y}_k$	An estimated value at index k
$\sum_{j=0}^{n}$	Summation from j = 0 to j = n, so summing up n + 1 terms
$Y_k^{(j)}$	Classifier
w_j	Weight assigned to j^{th} classifier

Table 2. Summary of notations.

approximate the underlying data distribution. It creates a histogram for each characteristic and divides the data into distinct bins based on the histogram of each feature. Subsequently, each bin was assigned a decision tree model. The HBGB aims to reduce an objective function that consists of a loss term and regularization term. The most used loss function for regression is the mean squared error (MSE). For classification, the log loss is declared as Eq. 7.

$$l(u_i, \widehat{u_i}) = (u_i, \widehat{u_i})^2 \tag{6}$$

$$l(u_i, \widehat{u_i}) = -u_i \log(\widehat{u_i}) - (1 - u_i)\log(1 - \widehat{u_i}) \tag{7}$$

where is the loss function that quantifies how well the model fits the data, n is the total number of data points, u_i is the true value for the i^{th} data point, and $\widehat{u_i}$ is the predicted value for the i^{th} data point. This can improve performance by enabling the procedure to more closely approach the data's essential spreading Fig. 4 shows the structure of histogram based gradient boosting. Because HBGB can conduct these patterns better than the usual gradient boosting, it is especially well-suited for datasets with many features, severely skewed data, or data with outliers. It can also be parallelized to expedite training, and is reasonably quick and simple to use.

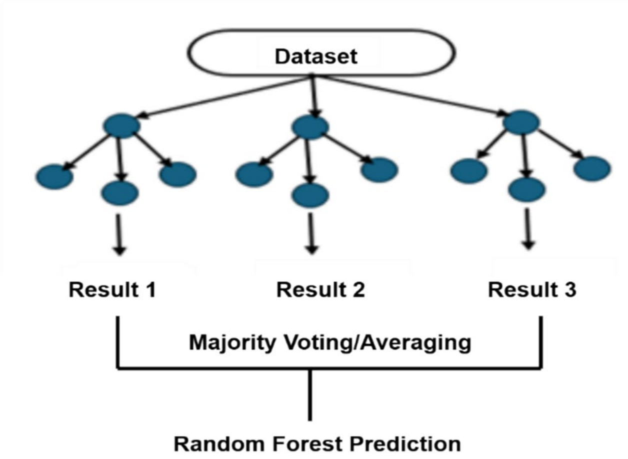


Fig. 3. Random forest classifier.

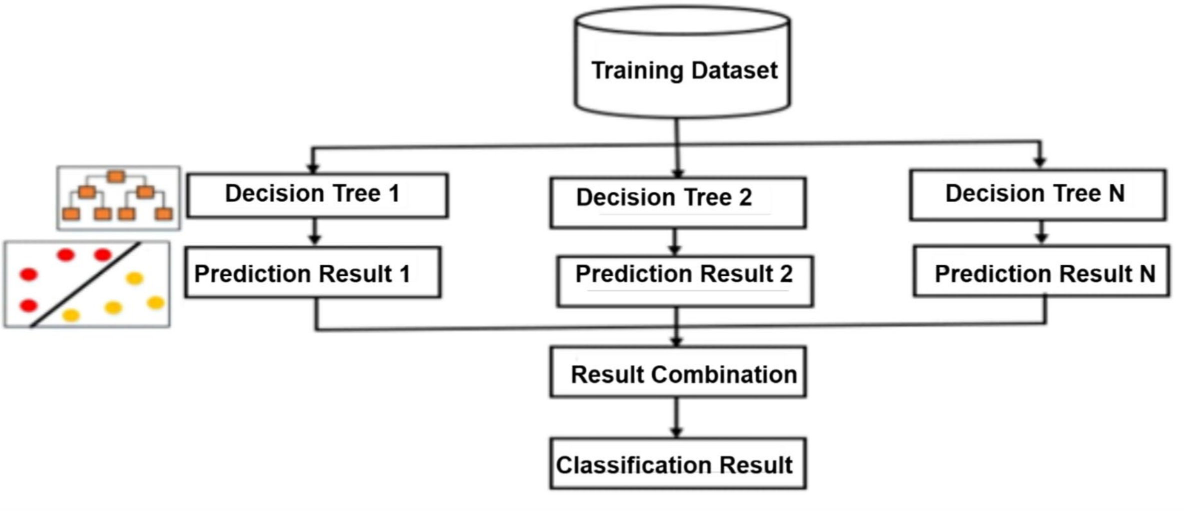


Fig. 4. Histogram-based gradient boosting structure.

Findings and discussion

This section discusses the results of the proposed WVE classifier and comparative analysis. A private dataset comprising stroke patient records was collected from text files (Excel file) at KC Multi-Specialty Hospital in Chennai, India. We confirm that all methods and experiments conducted were purely computational and did not involve any human subjects directly. The patient's details were highly confidential.

- We confirm that all methods were carried out in accordance with relevant guidelines and regulations.
- We confirm that all experimental protocols were approved by the KC Multispecialty Hospital, Chennai, India.
- We confirm that informed consent was obtained from all subjects and/or their legal guardian(s).

This dataset was used to compare various machine learning algorithms with the proposed model. This study used a private dataset comprising 280 records. After implementing quality assurance measures, 261 high-quality records were selected for analysis. Among these, 87 records were labeled as stroke cases (assigned a value of 1), and the remaining 174 were classified as normal (assigned a value of 0). To evaluate the model performance, the dataset was divided into a train set (80%) and a test set (20%). Extensive data preprocessing was performed to ensure consistency and accuracy of the analysis. This dataset provides valuable and distinct insights. To avoid the model deviating from the intended training data, data preprocessing is necessary before model construction to eliminate superfluous noise and outliers from the dataset. The dataset contains 11 characteristics. Table 3 shows the data sample format. The performance of these algorithms was evaluated using standard metrics, including accuracy, precision, recall, and F1-score.

Gender	Age	Hypertension	Heart_ disease	Ever_ married	Job type	Residence type	Avg_ Glucose level	BMI	Smoking status	Stroke
Male	57	0	1	No	Govt	Urban	217.08	33.80841	Unknown	1
Male	58	0	0	Yes	Private	Rural	189.84	31.37853	Unknown	1
Female	58	0	0	Yes	Private	Urban	71.2	30.00388	Unknown	1
Male	58	0	0	Yes	Private	Urban	82.3	30.19957	smokes	1
Female	59	0	0	Yes	Private	Rural	211.78	33.48457	formerly smoked	1
Male	79	0	1	Yes	Private	Urban	57.08	22	formerly smoked	0
Female	37	0	0	Yes	Private	Rural	162.96	39.4	never smoked	0
Female	37	0	0	Yes	Private	Rural	73.5	26.1	formerly smoked	0

Table 3. Sample data format.

Feature	Description
Gender	Male, female, others
Age	Age of the patient
Hypertension	0 = no hypertension, 1 = has hypertension
Heart disease	0 = no heart disease, 1 = has heart disease
Ever married	Patient’s marital status
Job type	Patient’s work type
Residence_type	Patient’s residence type
Avg_Glucose level	The average glucose level in the blood
BMI	body mass index
Smoking status	Smoking status: formerly smoking/never smoked/smoked
Stroke (Target)	0 (zero) = no stroke, 1(one) = has stroke

Table 4. Feature description for the dataset.

The dataset contains 11 features for each sample and a target variable. The target variable was binary, with 1 representing stroke and 0 representing no stroke. A brief overview of these features is provided in Table 4. Figure 5 presents an overview of the implementation process.

The dataset was pre-processed to handle missing values and normalize the numerical features. We maintained the original class distribution without applying any resampling techniques to preserve the real-world imbalance in stroke occurrence. Our dataset consisted of clinical and demographic data, including variables such as age, gender, hypertension, and heart disease. The data are tabular and are used to predict stroke risk or diagnose stroke.

- Total samples: 261 selected from 280 records
- Feature dimensions: 11 features after one-hot encoding
- Class distribution: 87 positive (stroke) and 174 negative stroke cases
- Data source: Clinical records were collected from hospital data and stored as CSV file.

We used the Stroke Prediction Dataset, which has the following features:

- Demographic information (age, gender)
- Medical history (hypertension, heart disease)
- Lifestyle factors (smoking status, BMI)
- Socioeconomic indicators (work type, residence type, average glucose level)
- Laboratory results (average glucose level, etc.)

Stroke affects older individuals, with most patients aged between 60 and 80 years. While men experience strokes earlier, typically in their mid-50 s to 80 s, women are commonly affected between their late 40 s and 80 s of age. The data revealed that a substantial proportion of patients, particularly men, were overweight or obese. Some patients have extremely high BMIs. Interestingly, although heart disease is not prevalent among patients with stroke, high blood pressure is not a common risk factor. Additionally, a larger number of patients maintained normal blood sugar levels. Table 5 presents the performance of different machine learning methods with the proposed model in predicting brain strokes. Implemented a machine learning (ML) technique was implemented using a WVE classifier in the proposed system. The proposed approach is tested using several machine learning techniques, with Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Tree Boosting, KNN, and Naive Bayes. Based on their accuracy scores, the best individual classifiers were used in the ensemble voting classifier. Figure 6 shows a comparative evaluation of the suggested framework with other models. The general F1 score obtained in this study was 92%. This model was fine-tuned to the highest possible degree after several

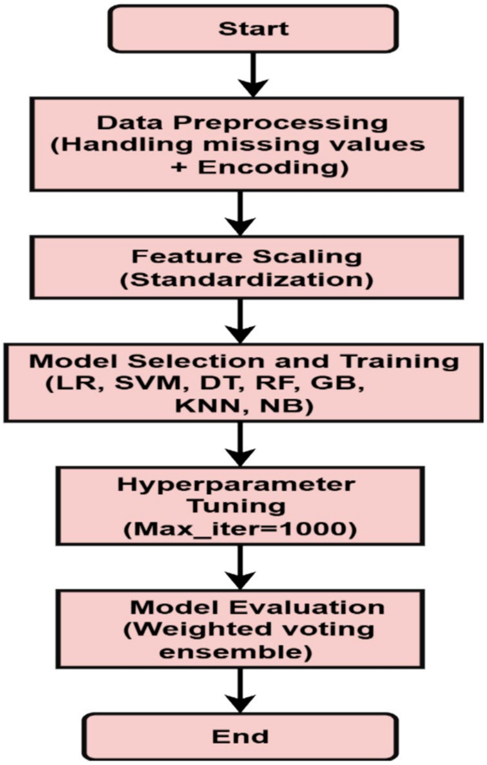

Start
Data Preprocessing (Handling missing values + Encoding)
Feature Scaling (Standardization)
Model Selection and Training (LR, SVM, DT, RF, GB, KNN, NB)
Hyperparameter Tuning (Max_iter=1000)
Model Evaluation (Weighted voting ensemble)
End

Fig. 5. Overview of the implementation process.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Logistic regression	76	74	75	77
Support vector machines	86	70	71	77
Decision tree	83	60	57	69
Random forest	90	80	82	85
Gradient tree boosting	86	70	71	77
K-nearest neighbor	76	74	75	76
Naive bayes	76	74	75	76
Proposed model	**94**	**90**	**92**	**92**

Table 5. Evaluation of performance metrics of the proposed model.

iterations. The model attained an accuracy of 92.31%. The proposed model achieved a maximum accuracy of 92%, recall of 90%, F1-score of 92%, and precision of 94%. Figure 7 shows the precision, recall, and F1-score.

A confusion matrix is a pictorial tool commonly used in machine learning to evaluate the execution of classification models. It presents a tabular representation of the expected and real class labels, providing a comprehensive evaluation of the model's accuracy. The rows of the matrix represent the true class labels, and the columns represent the predicted class labels of the model. The slanted elements indicate correct classifications, whereas the off-diagonal elements denote the misclassifications.

Many methods are currently employed to identify stroke disease, but the most underutilized method is preliminary stroke risk assessment based on critical factors, including age, blood glucose level, hypertension, and body mass index. A comprehensive analysis of the predicted data for the new patients was conducted. BMI was classified as normal (18.5–24.9), overweight (25–29.9), obese (30–34.9), or extremely obese (>34.9). Additionally, glucose levels were classified as normal (170–200), elevated (190–230), or high (220–300). These findings, combined with stroke risk levels, provide a comprehensive understanding of a patient's overall health and potential risk factors for stroke.

In addition, we determined the accuracy of the proposed structure to be 0.9231, and its log loss to be 0.4351. These values were calculated to facilitate comparisons with the existing models, as shown in Fig. 8. The models used in our ensemble Voting Classifier (RandomForestClassifier, XGBClassifier, and GradientBoostingClassifier) are traditional machine learning algorithms and do not rely on epoch-based training. These ensemble methods

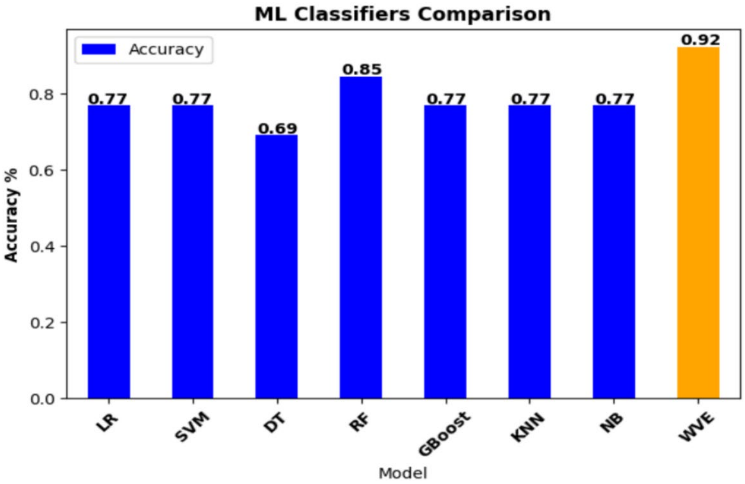


Fig. 6. Comparison of the proposed construction with alternative models.

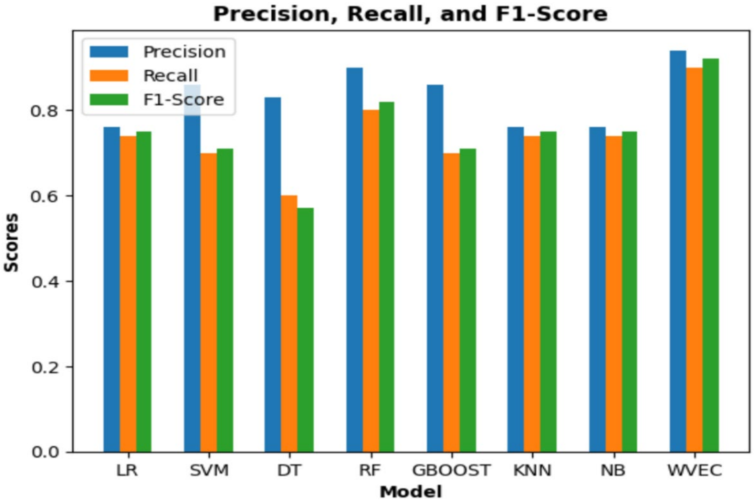


Fig. 7. An analysis of the proposed model's precision, recall, and F1-score.

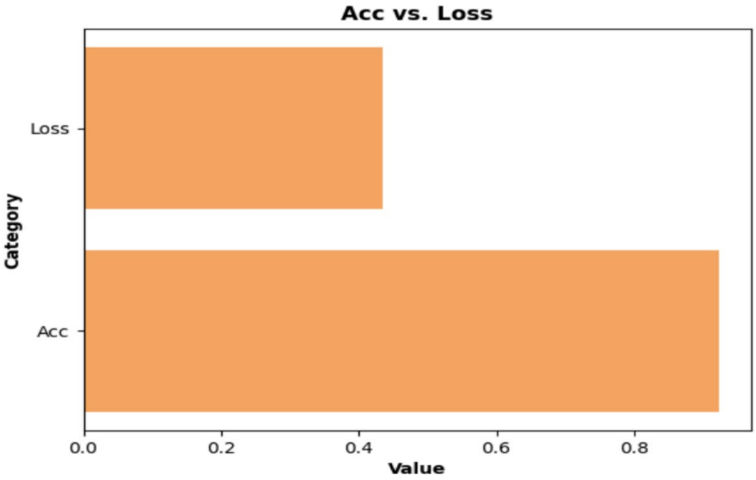


Fig. 8. Accuracy and log loss interpretation.

do not use epochs, as neural networks. Instead, they built decision trees based on bootstrapped samples. The accuracy (0.9231) and log loss (0.4351) metrics shown in Fig. 8 were not achieved through epoch-based training.

The final performance metrics were obtained from a single training and evaluation cycle on the test data after fitting the models. The relationship between accuracy and log loss is a result of the probability calibration from the soft voting method, where class probabilities from all base estimators are averaged. The reported metrics represent the ensemble's performance on unseen test data, suggesting that the model generalizes well with minimal overfitting.

This mathematical equation describes the process of weighted voting-based ensemble classification, and the final predicted class is determined by choosing the class with the highest average probability among all models. Figure 9 illustrates the importance of the Random Forest model, which shows the relative implication of each variable in predicting outcomes. By analyzing the frequency with which features are used to split the data within the decision trees, we can identify the most influential factors. This information is valuable for selecting relevant features, understanding the decision-making process of the model, and exploring the underlying relationships between variables. However, potential limitations, such as correlation and non-linearity, should be considered.

A voting classifier equation is proposed using a weighted average approach for every prediction model, defined as follows in Eq. 8, where the weight must be a specific value. w_j represents the weight assigned to each classifier. Where m represents the individual classifiers. $Y_k^{(j)}$ are the classifiers. The approximate probabilities p can be calculated for the models as follows in Eq. 9, where w_j = weight assigned to j$^{\text{th}}$ classifier. The final prediction $P(x)$ for input x is determined by the weighted sum of the classifier predictions, as expressed in Eq. 10.

$$\widehat{Y}_k = \sum\nolimits_{j=0}^{n} w_j Y_k^{(j)}, where w_j > 0 \tag{8}$$

$$\widehat{y} = \text{argmax} \sum\nolimits_{j=0}^{m} w_j p_{ij} \tag{9}$$

$$P(x) = \left\{ \begin{array}{c} 1, if \sum_{i=1}^{N} w_i . C_i(x) \geq T \\ 0, otherwise \end{array} \right. \tag{10}$$

where:

- T is the decision-making threshold.
- w_i is the weight of the i$^{\text{th}}$ classifier, based on its performance.
- N is the total number of classifiers in an ensemble.
- $C_i(x)$ be the prediction of the i$^{\text{th}}$ classifier for input x, where $i \in \{1,2,\ldots..,N\}$.
- w_i be the weight allocated to the i$^{\text{th}}$ classifier, where $w_i \geq 0 and \sum_{i=1}^{N} w_i = 1$.

For the Binary Classification problem, which is stroke and no stroke, where 0 indicates no stroke detected and 1 indicates stroke detected.

$$C_i(x) \in \{0,1\}$$

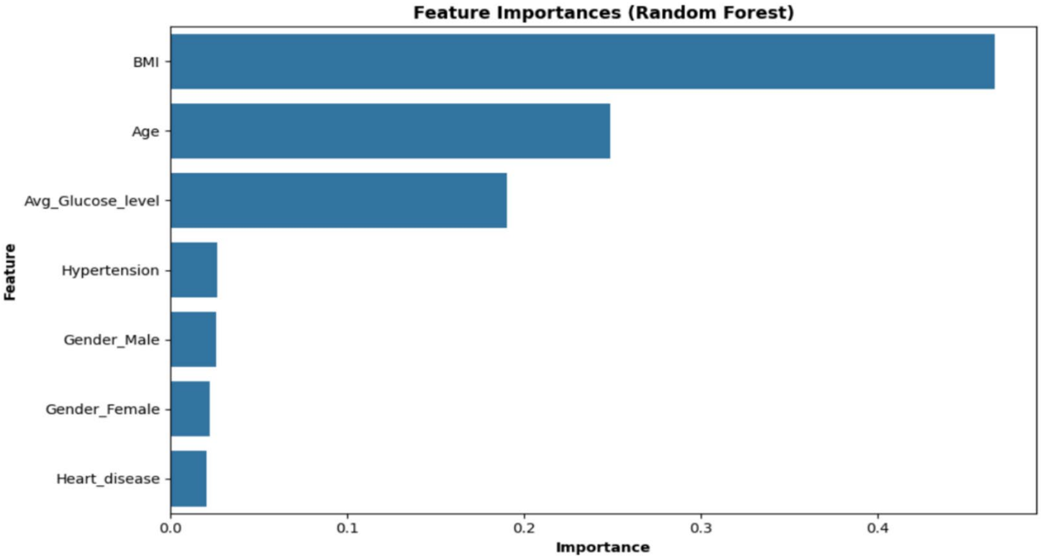
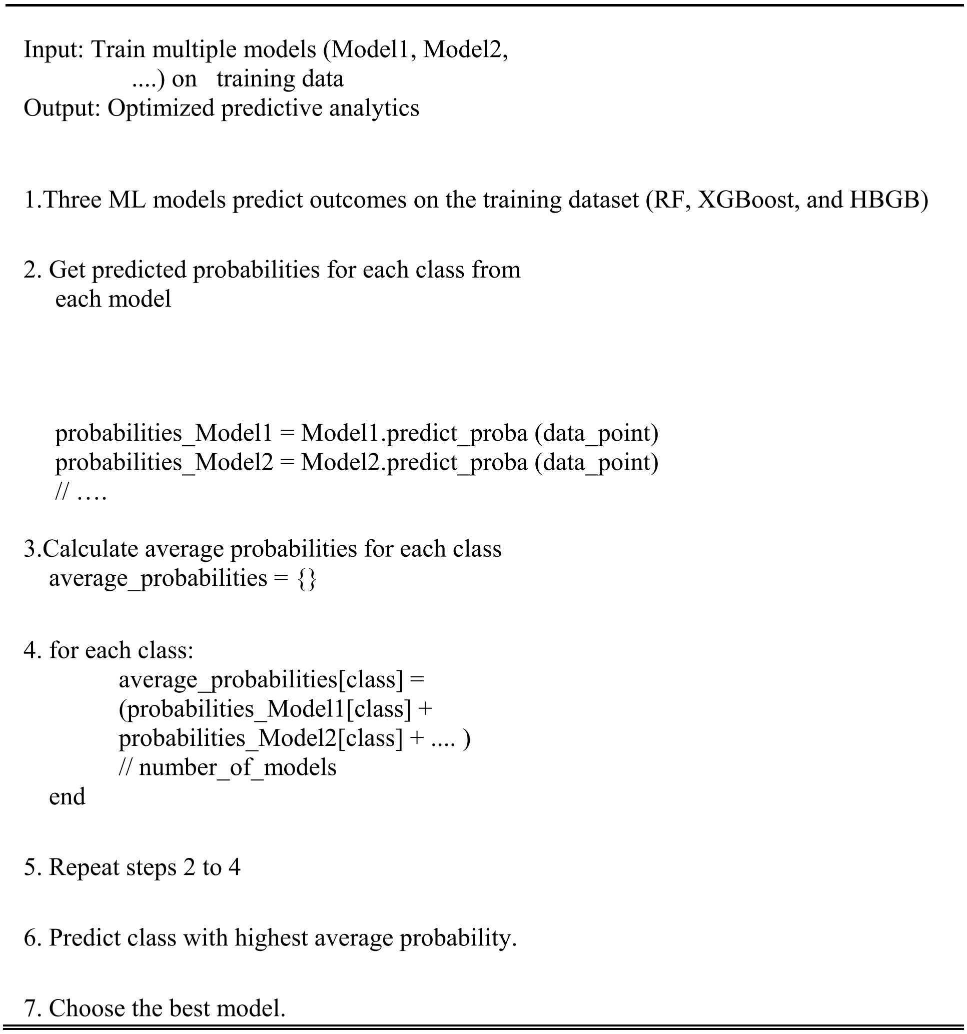


Fig. 9. Feature importance ranking.

Input: The function takes a data point, a list of trained models, and a list of corresponding weights as input.

Weighted Probabilities: It iterates through the models, obtains the predicted probabilities for each class, and multiplies them by the corresponding weights. The weighted probabilities were accumulated for each class.

Prediction: The function returns the class with the highest accumulated weighted probability as the predicted class.



```
Input: Train multiple models (Model1, Model2,
            ....) on   training data
Output: Optimized predictive analytics


1.Three ML models predict outcomes on the training dataset (RF, XGBoost, and HBGB)


2. Get predicted probabilities for each class from
    each model



    probabilities_Model1 = Model1.predict_proba (data_point)
    probabilities_Model2 = Model2.predict_proba (data_point)
    // ….

3.Calculate average probabilities for each class
   average_probabilities = {}


4. for each class:
            average_probabilities[class] =
            (probabilities_Model1[class] +
            probabilities_Model2[class] + .... )
            // number_of_models
   end


5. Repeat steps 2 to 4


6. Predict class with highest average probability.


7. Choose the best model.
```

Algorithm 1 Proposed model

Algorithm 1 combines the predictions from multiple machine learning models (Random Forest (RF), XGBoost, and Histogram-Based Gradient Boosting (HBGB)) using an ensemble averaging approach. Key steps include:

1. Training of multiple models.
2. Extracting the predicted probabilities from each model for every class.
3. Average probabilities for each class across all models.
4. The class with the highest average probability was selected for final prediction.
5. Finally, the optimal model was identified based on the performance metrics.

where n is the number of samples, d is the number of features, and t is the number of trees.

Model	Time complexity	Space complexity	Training time (s)	Memory usage (Bytes)
Random forest	$O(n.d.t)$	$O(n.d)$	2	4848
XGBoost	$O(n.d.t)$	$O(n.d)$	2	96

Table 6. Complexity analysis of model.

Model	Accuracy	Study
Logistic regression	78.40%	[39]
Support vector machines	78.40%	[40]
Naïve bayes	77.40%	[41]
K-Nearest neighbor	91.72%	[42]
RF + XGBoost + HGB	92.31%	Current study

Table 7. Performance comparison between proposed method and previous study.

Table 6 shows the complex analysis of the random forest and XGBoost Models. The algorithm balances the predictions from RF, XGBoost, and HBGB to optimize the final performance. Although RF is computationally simpler, XGBoost offers superior accuracy for certain tasks at the expense of higher complexity. The ensemble approach ensures that the strengths of both models are utilized, and the optimal complexity depends on the performance-complexity trade-off demonstrated by the dataset.

The WVE classifier is an ensemble learning technique that combines the predictions of several base models. Each base model was assigned a weight based on its performance on the validation set. During prediction, the weighted votes from all base models were combined, and the class with the highest weighted vote was selected. The WVE can improve generalization, reduce overfitting, and increase robustness compared with individual models. We chose to use [Random Forest, eXtreme gradient boosting, and histogram-based gradient boosting] for WVE, as they have been shown to be effective in similar tasks.

A comparison of the accuracies of the proposed WVE classifier model and other models is presented in Table 7. The proposed method showed a significant improvement in accuracy compared with previous studies that used machine learning techniques for stroke detection.

Conclusion and future work

Preventive detection is essential for reducing brain damage and improving patient prognosis after stroke. Machine learning can play a key role in early stroke detection by analyzing medical data. In this study, a stroke recognition approach based on the WVE classifier was presented. The proposed model is a composite machine learning method. The integration of the results of several different classifiers, including Random Forest, eXtreme Gradient Boosting, and histogram-based gradient boosting, produces a definitive prediction. These probabilities, which all classifiers deliver as a valuation of the chance that they belong, are added to create the final forecast using a weighted average. Finally, our research provides a hybrid framework that integrates machine learning techniques. The improved weighted voting ensemble model classified brain stroke with high accuracy (92.31%). Our research provides a comprehensive understanding of patients' overall health and potential risk factors.

Limitation

One of the limitations of this study is the small size of our private dataset, which may affect the generalizability of the results. Furthermore, the complexity of combining multiple machine learning models in a weighted voting ensemble classifier to optimize accuracy is a challenge. Another drawback is the difficulty in understanding the hybrid RF + XGBoost + HGB models, which may make it difficult for healthcare providers to trust the model's predictions. The stroke dataset used contained many samples labeled as"unknown,"and although data cleaning was performed, testing the models on different datasets might yield different results.

Future work

Future research could develop an app for stroke diagnosis using CT scan image data to enhance patient outcomes and individualized treatment. The effectiveness of the framework should be assessed using larger datasets and for clinical utility in healthcare settings. The integration of advanced neuroimaging modalities and comprehensive clinical data could improve stroke localization and diagnostic accuracy. Privacy-preserving techniques, such as federated learning and centralized deep learning security exploration, can be adopted to maintain patient data confidentiality.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to that permission has not been granted by the hospital to share the dataset publicly, but are available from the corresponding author on reasonable request.

Received: 24 January 2025; Accepted: 30 July 2025
Published online: 25 August 2025

References

1. Yüksel, M. M. et al. Low-intensity focused ultrasound neuromodulation for stroke recovery: A novel deep brain stimulation approach for neurorehabilitation?. *IEEE Open J. Eng. Med. Biol.* **4**, 300–318. https://doi.org/10.1109/ojemb.2023.3263690 (2023).
2. Polat, Ö., Dokur, Z. & Ölmez, T. Classification of brain strokes using divergence-based convolutional neural networks. *Biomed. Signal Process. Control* **93**, 106193. https://doi.org/10.1016/j.bspc.2024.106193 (2024).
3. Goni, M. R., Ruhaiyem, N. I. R., Mustapha, M., Achuthan, A. & Che Mohd Nassir, C. M. N. Brain vessel segmentation using deep learning—a review. *IEEE Access* **10**, 111322–111336. https://doi.org/10.1109/access.2022.3214987 (2022).

4. Baek, H. et al. Deep cerebellar low-intensity focused ultrasound stimulation restores interhemispheric balance after ischemic stroke in mice. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28**(9), 2073–2079. https://doi.org/10.1109/tnsre.2020.3002207 (2020).
5. Guo, T. et al. Pulsed transcranial ultrasound stimulation immediately after the ischemic brain injury is neuroprotective. *IEEE Trans. Biomed. Eng.* **62**(10), 2352–2357. https://doi.org/10.1109/tbme.2015.2427339 (2015).
6. Sarmento, R. M., Vasconcelos, F. F. X., Filho, P. P. R., Wu, W. & de Albuquerque, V. H. C. Automatic neuroimage processing and analysis in stroke—a systematic review. *IEEE Rev. Biomed. Eng.* **13**, 130–155. https://doi.org/10.1109/rbme.2019.2934500 (2020).
7. Sirsat, M. S., Fermé, E. & Câmara, J. Machine learning for brain stroke: A review. *J. Stroke Cerebrovasc. Dis.* **29**(10), 105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162 (2020).
8. Amin, J., Sharif, M., Haldorai, A., Yasmin, M. & Nayak, R. S. Brain tumor detection and classification using machine learning: A comprehensive survey. *Complex. Intell. Syst.* **8**(4), 3161–3183. https://doi.org/10.1007/s40747-021-00563-y (2021).
9. Zhu, W. et al. Global and regional burden of ischemic stroke disease from 1990 to 2021: an age-period cohort analysis. *Cold Spring. Harbor Lab.* https://doi.org/10.1101/2024.08.29.24312683 (2024).
10. Efremov, V., Lakshmanan, R. S., O'Donnell, J. & Killard, A. J. Rapid whole blood clot retraction assay on quartz crystal microbalance. *IEEE Sens. Lett.* **5**(1), 1–4. https://doi.org/10.1109/lsens.2020.3049063 (2021).
11. Boudihi, A., EL-azrak, M., Doudouh, O., Ismaili, N. & El Ouafi, N. Acute myocardial infarction complicated by severe left ventricle systolic dysfunction in a young patient after Covid 19 vaccination: A case report. *Radiol. Case Rep.* https://doi.org/10.1016/j.radcr.2022.09.053 (2023).
12. Lee, K. B. et al. Six-month functional recovery of stroke patients. *Int. J. Rehabil. Res.* **38**(2), 173–180. https://doi.org/10.1097/mrr.0000000000000108 (2015).
13. Park, K., Lonsberry, G. E., Gearing, M., Levey, A. I. & Desai, J. P. Viscoelastic properties of human autopsy brain tissues as biomarkers for Alzheimer's diseases. *IEEE Trans. Biomed. Eng.* **66**(6), 1705–1713. https://doi.org/10.1109/tbme.2018.2878555 (2019).
14. Krag, A. et al. Uniting to defeat steatotic liver disease: A global mission to promote healthy livers and healthy lives. *J. Hepatol.* https://doi.org/10.1016/j.jhep.2023.07.029(2023 (2023).
15. Wiltshire, T. J. & Fiore, S. M. Social cognitive and affective neuroscience in human-machine systems: A roadmap for improving training, human-robot interaction, and team performance. *IEEE Trans. Human-Mach. Syst.* **44**(6), 779–787. https://doi.org/10.1109/thms.2014.2343996 (2014).
16. Aspberg, S., Cheng, D., von Heijne, A., Gigante, B. & Singer, D. E. Brain MRI microbleeds and risk of intracranial hemorrhage in atrial fibrillation patients: A Swedish case-control study. *J. Stroke Cerebrovasc. Dis.* **33**(4), 107629. https://doi.org/10.1016/j.jstrokecerebrovasdis.2024.107629 (2024).
17. Wang, H.-L. et al. Automatic machine-learning-based outcome prediction in patients with primary intracerebral hemorrhage. *Front. Neurol.* https://doi.org/10.3389/fneur.2019.00910 (2019).
18. Srinivas, A. & Mosiganti, J. P. A brain stroke detection model using soft voting based ensemble machine learning classifier. *Meas.: Sens.* https://doi.org/10.1016/j.measen.2023.100871 (2023).
19. Jayachitra, S. & Prasanth, A. Multi-feature analysis for automated brain stroke classification using weighted gaussian naïve bayes classifier. *J. Circ. Syst. Comput.* **30**(10), 2150178. https://doi.org/10.1142/s0218126621501784 (2021).
20. Assam, M. et al. An efficient classification of MRI brain images. *IEEE Access* **9**, 33313–33322. https://doi.org/10.1109/access.2021.3061487 (2021).
21. Hatem, S. M. et al. Rehabilitation of motor function after stroke: A multiple systematic review focused on techniques to stimulate upper extremity recovery. *Front. Human Neurosci.* https://doi.org/10.3389/fnhum.2016.00442 (2016).
22. Sale, P. et al. Predicting motor and cognitive improvement through machine learning algorithm in human subject that underwent a rehabilitation treatment in the early stage of stroke. *J. Stroke Cerebrovasc. Dis.* **27**(11), 2962–2972. https://doi.org/10.1016/j.jstrokecerebrovasdis.2018.06.021 (2018).
23. Orfanoudaki, A. et al. Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score. *PLoS ONE* **15**(5), e0232414. https://doi.org/10.1371/journal.pone.0232414 (2020).
24. Hung, C.-Y., Chen, W.-C., Lai, P.-T., Lin, C.-H., & Lee, C.-C. Comparing deep neural network 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)*, 3110–3113 https://doi.org/10.1109/embc.2017.8037515 (2017).
25. Hegland, S. J., Nielsen, A., Lázaro, A., Bjerknes, A. & Totland, Ø. How does climate warming affect plant-pollinator interactions?. *Ecol. Lett.* **12**(2), 184–195. https://doi.org/10.1111/j.1461-0248.2008.01269.x (2009).
26. Kassner, A. & Merali, Z. Assessment of blood-brain barrier disruption in stroke. *Stroke* **46**(11), 3310–3315. https://doi.org/10.1161/strokeaha.115.008861 (2015).
27. Janyalikit, T. & Ratanamahatana, C. A. time series shapelet-based movement intention detection toward asynchronous BCI for stroke rehabilitation. *IEEE Access* **10**, 41693–41707. https://doi.org/10.1109/access.2022.3167703 (2022).
28. Gong, Z., Ding, Y., Chen, Y. & Cree, M. J. Wearable microwave medical sensing for stroke classification and localization: A space-division-based decision-tree learning method. *IEEE Trans. Antennas Propag.* **71**(8), 6906–6917. https://doi.org/10.1109/tap.2023.3283131 (2023).
29. Movafagh, M. et al. Enhancing radar-based stroke detection: A novel approach integrating huygens' principle and deep learning techniques. *IEEE Sens. J.* **24**(21), 36085–36098. https://doi.org/10.1109/jsen.2024.3459013 (2024).
30. Sadr, H., Salari, A., Ashoobi, M. T. & Nazari, M. Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models. *Eur. J. Med. Res.* https://doi.org/10.1186/s40001-024-02044-7 (2024).
31. Khodaverdian, Z., Sadr, H., & Edalatpanah, S. A. A Shallow Deep Neural Network for Selection of Migration Candidate Virtual Machines to Reduce Energy Consumption.In *2021 7th International Conference on Web Research (ICWR)* 191–196 https://doi.org/10.1109/icwr51868.2021.9443133 (2021).
32. Nazari, M., Emami, H., Rabiei, R., Hosseini, A. & Rahmatizadeh, S. Detection of cardiovascular diseases using data mining approaches: application of an ensemble-based model. *Cogn. Comput.* **16**(5), 2264–2278. https://doi.org/10.1007/s12559-024-10306-z (2024).
33. Saberi, Z. A., Sadr, H., & Yamaghani, M. R. An Intelligent Diagnosis System for Predicting Coronary Heart Disease.In *2024 10th International Conference on Artificial Intelligence and Robotics (QICAR)*, 131–137 https://doi.org/10.1109/qicar61538.2024.10496601 (2024).
34. Sadr, H., Soleimandarabi, M. N. & Khodaverdian, Z. Automatic assessment of short answers based on computational and data mining approaches. *J. Decis. Oper. Res.* **6**(2), 242–255 (2021).
35. Bhowmick, R., Mishra, S. R., Tiwary, S. & Mohapatra, H. Machine learning for brain-stroke prediction: Comparative analysis and evaluation. *Multimed. Tool. Appl.* https://doi.org/10.1007/s11042-024-20057-6 (2024).
36. Xie, S. et al. A comprehensive analysis of stroke risk factors and development of a predictive model using machine learning approaches. *Mol. Genet. Genom.* https://doi.org/10.1007/s00438-024-02217-3 (2025).
37. Sorayaie Azar, A. et al. Predicting stroke severity of patients using interpretable machine learning algorithms. *Eur. J. Med. Res.* https://doi.org/10.1186/s40001-024-02147-1 (2024).
38. Mohamed, A. M., Rabie, A. H., Amer, H. M., Saleh, A. I. & Abo-Elsoud, M. E. A. Real time brain stroke identification using face images based on machine learning and booby bird optimization. *Expert Syst. Appl.* **282**, 127719. https://doi.org/10.1016/j.eswa.2025.127719 (2025).

39. Krishna, V., Sasi Kiran, J., Prasada Rao, P., Charles Babu, G., & John Babu, G. Early Detection of Brain Stroke using Machine Learning Techniques. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, 1489–1495. https://doi.org/10.1109/icosec51865.2021.9591840 (2021).
40. Raghavendra, U. Novel and accurate non-linear index for the automated detection of haemorrhagic brain stroke using CT images. *Complex Intell. Syst.* **7**(2), 929–940. https://doi.org/10.1007/s40747-020-00257-x (2021).
41. Rukmawan, S. H., Aszhari, F. R., Rustam, Z. & Pandelaki, J. Cerebral infarction classification using the k-nearest neighbor and naive bayes classifier. *J. Phys: Conf. Ser.* **1752**(1), 012045. https://doi.org/10.1088/1742-6596/1752/1/012045 (2021).
42. Kommina, L., Theerthagiri, P., Payyavula, Y., Vemula, P. S. & Reddy, G. D. Post-stroke readmission prediction model using machine learning algorithms. *Emerg. Trend. Data Driven Comput. Commun.* https://doi.org/10.1007/978-981-16-3915-9_4 (2021).
43. Ayoub, M. et al. End to end vision transformer architecture for brain stroke assessment based on multi-slice classification and localization using computed tomography. *Comput. Med. Imaging Graph* **109**, 102294. https://doi.org/10.1016/j.compmedimag.2023.102294 (2023).
44. Li, C. et al. Deep-learning-enabled microwave-induced thermoacoustic tomography based on ResAttU-Net for transcranial brain hemorrhage detection. *IEEE Trans. Biomed. Eng.* **70**(8), 2350–2361. https://doi.org/10.1109/tbme.2023.3243491 (2023).
45. Raj, R., Mathew, J., Kannath, S. K. & Rajan, J. StrokeViT with AutoML for brain stroke classification. *Eng. Appl. Artif. Intell.* **119**, 105772. https://doi.org/10.1016/j.engappai.2022.105772 (2023).

Acknowledgements

The authors would like to thank the KC Multispecialty Hospital for providing the dataset, appreciated the insights, and technical support for data analysis.

Author contributions

Reeta Samuel conceived and designed the experiments. Thanapal Pandi analyzed the data. All the authors interpreted the results. Reeta Samuel wrote the first and the revised drafts of the manuscript. All authors contributed to the writing of the manuscript. All authors agree with the results and conclusions of the manuscript. All authors have read and confirmed that they meet the expectations. Thanapal Pandi had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Funding

Open access funding provided by Vellore Institute of Technology. This research was conducted independently and did not receive any specific funding from public, commercial, or non-profit sources.

Declarations

Competing interests

The authors declare no competing interests.

Ethics

A statement to confirm that all the methods and experiments conducted were purely computational and did not involve any human subjects directly. The names and personal details of the patients were highly confidential. All methods were carried out in accordance with relevant guidelines and regulations. All experimental protocols were approved by the KC Multi specialty Hospital in Chennai, India.

Consent for publication

We confirm that informed consent was obtained from all subjects and/or their legal guardian(s).

Additional information

Correspondence and requests for materials should be addressed to T.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.