Measurement: Sensors 29 (2023) 100871

Contents lists available at ScienceDirect

Measurement: Sensors

journal homepage: www.sciencedirect.com/journal/measurement-sensors

A brain stroke detection model using soft voting based ensemble machine learning classifier

A. Srinivas [a,b,*], Joseph Prakash Mosiganti [c]

[a] Department of Computer Science and Engineering, Faculty of Informatics, University College of Engineering, Osmania University, Hyderabad, Telangana, India
[b] SRR Govt. Arts & Science College, Karimnagar, Telangana, India
[c] Department of Computer Science and Engineering, MCET, Hyderabad, Telangana, India

ARTICLE INFO

Keywords:
Brain stroke detection
Machine learning
Soft voting
Ensemble classifier
Random forest
Extremely randomized trees
Histogram-based gradient boosting

ABSTRACT

A brain stroke is a medical emergency that occurs when the blood supply to a part of the brain is disturbed or reduced, which causes the brain cells in that area to die. The process involves training a machine learning model on a large labelled dataset to recognize patterns and anomalies associated with strokes. The proposed model is an ensemble machine learning algorithm, which integrates predictions obtained from several individual classifiers like Random Forest, Extremely Randomized Trees and Histogram-Based Gradient Boosting to make a final prediction. Each classifier provides a probability estimate for each class, final prediction is based on weighted average of these probabilities. The weights assigned to each classifier can be based on their performance on a validation set or can be set uniformly. The proposed soft voting model improved accuracy and robustness of final prediction compared to a single classifier. The limitation of study classification of stroke type can lead to appropriate use of resources and help to reduce healthcare costs. To solve this issue, future introduced a swarm intelligence-based optimization for improve classification accuracy. The proposed model obtained an accuracy of 96.88%. In order to carry out the investigation, the stroke prediction dataset is collected from UCI machine learning repository.

1. Introduction

In the world, stroke is a top reason behind the death and also a prominent reason behind high morbidity, which may lead to disability [1]. Before beginning therapy for a stroke, it is critical to get an accurate diagnosis, since the course of treatment for a stroke is determined by the kind of stroke that was experienced. Stroke is a neurological disorder that may develop from ischemia or bleeding of the brain arteries, and it often results in diverse motor and cognitive deficits, which affect functions [2]. Stroke is also known as cerebrovascular accident. Around the globe, around 16 million people are affected by strokes every year, which is connected with tremendous expenses to society.

There are a variety of perspectives that influence how a stroke is understood; yet, in most cases, a stroke will provoke an intuitive response that is unmistakable. Each memory is encoded and kept in a network in a human brain, which weighs more than three pounds and consists of more than 100 billion neurons and one trillion glia. The functioning of a person's brain is essential to the maintenance of their respiration and movement. Since 1970, or over the course of more than half a century, 10 times more people have died from strokes in developing countries than in industrialized nations, and it is anticipated that the number will quadruple worldwide by the year 2030. In general, strokes may be broken down into one of the following three categories:

Ischemic strokes, hemorrhagic strokes, and transient ischemic attacks are all kinds of strokes (TIA). Ischemic strokes are far and by the most prevalent kind of stroke [3]. Ischemic stroke, which occurs when a clot or other blockage remains in a brain blood channel [4], is responsible for 87% of all strokes, as per the American Heart Association (AHA). Ischemic strokes may also be caused by high blood pressure. Ischemic strokes may be divided into two categories, namely thrombotic strokes and embolic strokes. An embolic stroke is what happens when a clot or blockage develops in any place of the body and then shifts to the brain, where it obstructs the bloodflow to the brain. A thrombotic stroke results if a clot exists in an artery, which provides blood to the brain, which restricts the flow of blood.

A hemorrhagic stroke occurs when a blood vessel that has been compromised either splits or bursts [5]. Although hemorrhagic strokes are believed to account for only around 10%–15% of all strokes, they

* Corresponding author. Department of Informatics, University College of Engineering, Osmania University, Hyderabad, Telangana.
E-mail addresses: andhe.srinivas2020@gmail.com (A. Srinivas), mjosephp7@gmail.com (J.P. Mosiganti).

https://doi.org/10.1016/j.measen.2023.100871
Received 31 March 2023; Received in revised form 19 June 2023; Accepted 7 August 2023
Available online 7 August 2023
2665-9174/

have a greater mortality rate than ischemic strokes. This is despite the fact that the number of people who suffer from hemorrhagic strokes is considered to be relatively low. There are two different types of hemorrhagic strokes. These include subarachnoid hemorrhage and intracerebral hemorrhage. A transient ischemic attack, sometimes referred as a "mini-stroke," is brought on by a clot, and the condition is described as follows: In contrast to other forms of stroke, a transient ischemic attack (TIA) is a temporary blockage that only lasts for a short period of time [6] (on average, 1 min), with symptoms disappearing within 24 h. Even though a TIA does not cause any kind of permanent damage to the brain or any of its components doctors nevertheless consider it to be a warning indication that another stroke is going to happen soon [7]. It is generally accepted that stroke is a disease that results in death, regardless of the kind.

1.1. Brain stroke classification is important for several reasons

- **Treatment:** As mentioned earlier, the treatment for ischemic and hemorrhagic strokes is different. Therefore, accurate classification of the form of stroke is very important for ensuring that patients receive the appropriate treatment.
- **Prognosis:** The type of stroke can also affect the patient's prognosis. For example, the outcome for patients with ischemic stroke is generally better than for those with hemorrhagic stroke.
- **Research:** Classifying strokes by type can also aid in research studies, allowing scientists to better understand the causes and risk factors for each type of stroke.
- **Patient Care:** Accurate classification of stroke can help healthcare providers to make informed decisions about patient care and management.
- **Time Management:** Quickly classifying the stroke type can help in quickly administering the treatment and hence can save precious time which is critical for the stroke patients.
- **Cost Management:** Proper classification of the stroke type can lead to the appropriate use of resources and help to reduce healthcare costs.

Overall, brain stroke classification is important for ensuring that patients receive the best possible care and for improving our understanding of stroke.

In recent years, machine learning has seen tremendous growth and advancement in a number of applications across an extensive array of health care systems. Machine learning (ML) is a cutting-edge technology that may provide medical practitioners with assistance in making clinical judgements and forecasts. Throughout the course of the last several decades, a number of researches have been carried out on the enhancement of stroke diagnosis via the use of ML according to speed and accuracy. The present study categorizes few of those researches depending on their similarities, evaluates every classification in a systematic manner, and gives useful perspectives into the use of ML-based approaches in brain stroke.

The aim of this research work is to introduce a model for detecting strokes applying a soft voting classifier. The proposed model utilizes an ensemble machine learning algorithm that incorporates the predictions of various individual classifiers, such as Random Forest, Extremely Randomized Trees and Histogram-Based Gradient Boosting, to produce a final prediction. This approach employs a probability estimate for each class generated by each classifier, and the final prediction is obtained through the weighted average of these probabilities. The weights assigned to each classifier can be either uniform or based on their respective performances on a validation set. By utilizing this soft voting classifier-based approach, the proposed model can improve the accuracy of stroke detection, providing a faster and more reliable method for identifying strokes.

2. Literature review

Stroke is a prominent reason behind the death and ailments worldwide. Early and accurate classification of stroke is crucial for providing appropriate treatment and improving outcomes. Machine learning (ML) methods have been applied to classify brain strokes using several imaging modalities, like computed tomography (CT) and magnetic resonance imaging (MRI). One way of the methodology to stroke classification using ML is to extract features from imaging data, such as texture, shape, and intensity, and then use these features as input to a classifier. Various feature extraction techniques have been introduced, which includes gray-level co-occurrence matrix (GLCM), gray-level run length matrix (GLRLM), and wavelet transform. Different classifiers, like support vector machine (SVM), k-nearest neighbors (k-NN), and decision trees, have been used to classify the stroke type depending on the features' extracted.

One more approach is to use deep learning (DL) methods, like convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to classify brain strokes directly from imaging data. These models have been shown to achieve high accuracy in classifying stroke type, and they have the advantage of being capable of learning the features automatically from the data.

Many studies have been carried out for evaluating the performance achieved with ML-based methods for classifying brain strokes, and they have reported promising results. For example, a study using SVM with GLCM features and CT images achieved an accuracy of 85.7% for classifying ischemic and hemorrhagic strokes, and a study using a CNN with MRI images achieved an accuracy of 94.2% for classifying infarction and edema.

The study effort that was offered by Garg et al. [8], aimed to automate Ischemic Stroke (IS) subtyping by using natural language processing techniques to electronic health records (EHR) in conjunction with machine learning methodology. Techniques: authors used natural language processing to analyze unordered text-based EHR data, which included neurology progress records and neuroradiology reports. These data came from an observational registry with TOAST subtyping that was resolved by board-certified vascular neurologists. Among the patients in this registry were people with IS. The authors used a few different strategies for feature selection to bring down the features' high dimensionality, and 5-fold cross validation used to assess the generalizability and limit the amount of overfitting that occurred. Authors made use of a number of different machine learning approaches and computed the kappa values to determine the degree of agreement that existed between each machine learning technique and human adjudication. Following that, we conducted an experiment in which the best algorithm was tested anonymously against a held-out subset of fifty instances.

A unique optimized fuzzy level segmentation approach is introduced by Jayachitra and Prasanth [9] to identify the ischemic stroke lesions. This technique was designed to detect the ischemic stroke lesions. Once the data has been segmented, the multi-textural features that will comprise the feature set are extracted. In order to differentiate between normal and pathological stroke lesion classes, these characteristics are assumed as the input to the weighted Gaussian Naive Bayes classifier that was suggested. The findings of the experiment show that the suggested approach is capable of achieving a greater level of accuracy than the current methods that are considered to be state-of-the-art.

Salucci et al. [10] presents a two-step Learning-by-Examples technique for the real-time categorization of hemorrhagic/ischemic brain strokes and their sequential localization using microwave scattering data gathered around the human head. This approach was developed for the purpose of determining where strokes occurred after they had already occurred. In order to evaluate the practicability of the suggested strategy in the direction of a reliable monitoring and instantaneous diagnostic clinic procedure, an experimental evaluation is carried out and compared to data that have been controlled in a laboratory.

An automated approach for detecting an ischemic stroke using a diffusion-weighted image (DWI) series of MR images is presented by Subudhi et al. [11]. This method is based on a computer-aided decision system and is automated. The segmentation and categorization of brain strokes into three distinct categories, as outlined by the Oxford shire Community Stroke Project (OCSP) plan, are both components of the system. The primary categories for the stroke are the partial anterior circulation syndrome (PACS), the lacunar syndrome (LACS), and the whole anterior circulation stroke (TACS). To improve the detection accurateness, a technique known as fractional-order Darwinian particle swarm optimization (FODPSO) was used in the brain region that had been segmented using the expectation-maximization (EM) algorithm after the disrupted portion of the brain caused by the stroke had been identified.

Adam et al. [12], surveyed the most recent research on the categorization of ischemic stroke and provided their findings. In addition to that, the research came out with a classification model for ischemic stroke by making use of the decision tree method and k closest neighbor. The categorization model was developed using a dataset consisting of 400 instances that were gathered from several hospitals in Sudan. Ischemic stroke patients may be classified and diagnosed accurately with the help of medical specialists when they apply the findings of the decision tree algorithm. In addition, the findings of the research demonstrated that some characteristics may be used for the purpose of directly determining the type of ischemic stroke. These outcomes provide assistance to Healthcare experts throughout the process of ischemic stroke categorization. In addition to this, the findings showed that the majority of people who suffer from ischemic stroke in Sudan had thrombotic ischemic stroke.

Badriyah et al. [13], presented a research article in which the preprocessing of the data helps to enhance the image standard of stroke patients' CT scans. This was accomplished by optimizing the image quality to improve results and reduce noise. Additionally, Tessy Badriyah used machine learning algorithms for classifying the patients' images into two sub-categories of stroke disease, known as ischemic stroke and stroke hemorrhage. In this study, the classification of stroke diseases is accomplished through the application of eight different machine learning algorithms. These algorithms are referred to as Naive Bayes, K-Nearest Neighbors, Decision Tree, Logistic Regression, Random Forest, Deep Learning, Multi-layer Perceptron (MLP-NN), and Support Vector Machine.

Mariano et al. [14] proposed a method that is both effective and quick for the creation of huge datasets for using in machine learning algorithms to the categorization of brain strokes using microwave imaging devices. To shorten the amount of time necessary to establish the massive datasets required for training the machine learning algorithms, the suggested technique is founded on the distorted Born approximation and the linearization of the scattering operator. This was done to reduce the amount of effort required to generate these datasets. Every classifier has the capability of determining whether or not a stroke occurred, what kind of stroke it was (ischemic or hemorrhagic), and where exactly it was located inside the brain. The trained procedures were evaluated using datasets that were created by full-wave simulations of the whole system.

3. Proposed model

In this section a Soft Voting machine learning algorithm is proposed. In order to make comparisons regarding classification performance, eleven different machine learning algorithms are utilized. These algorithms are as follows: Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Native Bayes, Decision Tree, Extremely Randomized Trees, AdaBoost, Gradient Tree Boosting, and Histogram-Based Gradient Boosting.

In this work, voting based classifier is used. A voting classifier is a sort of classification model, whose training is done on an ensemble of many models and an output (class) is predicted depending on the class having the highest probability of being picked as the output. This type of classifier is often referred to as a voting classifier. It is used to make forecasts about the results of votes. Fig. 1 shows the flowchart of the voting classifier model for your viewing convenience.

The voting process gives a concise summary of the methods that will be used to compare various training models. Two methods are available for voting: soft voting and hard voting. Both have their advantages and disadvantages. In this work, the soft voting classifier is proposed.

In machine learning, soft voting is an approach for merging the predictions of several models so that the overall performance of the ensemble can be improved. Soft voting is used when the base classifiers in the ensemble are probability-based classifiers, such as logistic regression or a neural network with a SoftMax activation function, which output probability values for every class. In soft voting, the ensemble classifies a new input by taking a majority vote of the class predictions from each base classifier, where the vote weight of each base classifier is proportional to the predicted class probabilities for that input. This means that a base classifier with a higher probability for a particular class will have a higher vote weight for that class than a classifier with a lower probability.

Soft voting is generally considered to be more robust and accurate than hard voting, which just counts the number of times every class is predicted by the base classifiers, because it takes into account the confidence of each base classifier in its predictions. It can be applied to both classification and regression problem and it is a simple yet powerful technique to increase the performance achieved with machine learning model. The suggested model is depicted in Fig. 2.

In the proposed soft voting classifier, 3 base classifiers are considered: Random Forest, Extremely Randomized Trees and Histogram-Based Gradient Boosting. Each base classifier is discussed as follows:

Random Forest: Random Forest is an ensemble learning method used for classification and regression in machine learning. It is a fusion of decision trees, in which a group of decision trees are created and their outputs are combined to make a final prediction. The random forest

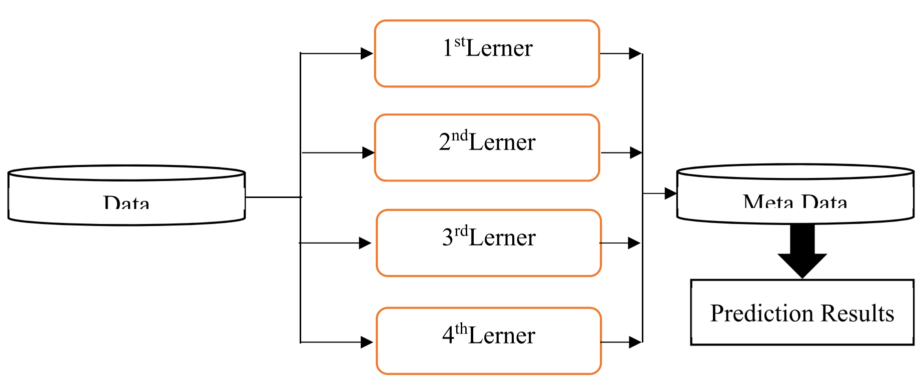

Fig. 1. Voting classifier.

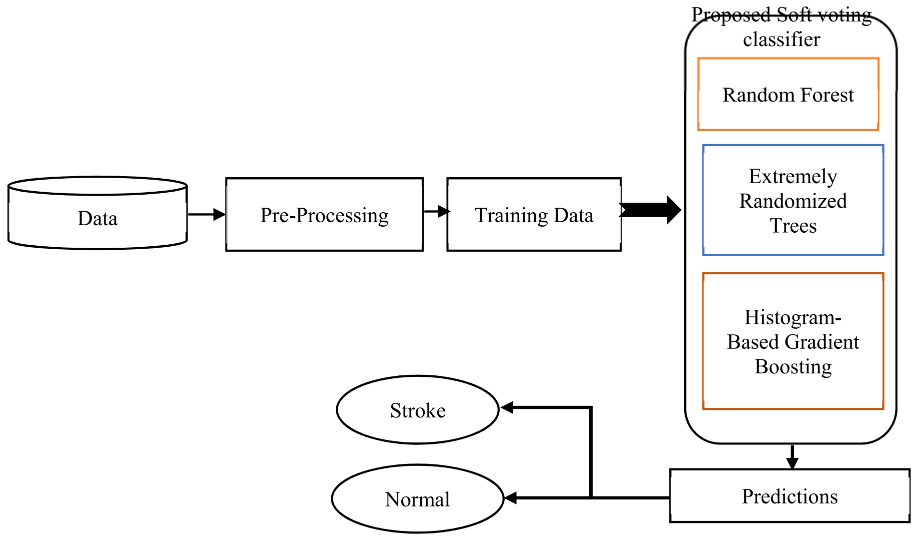

Fig. 2. Proposed soft voting classifier method.

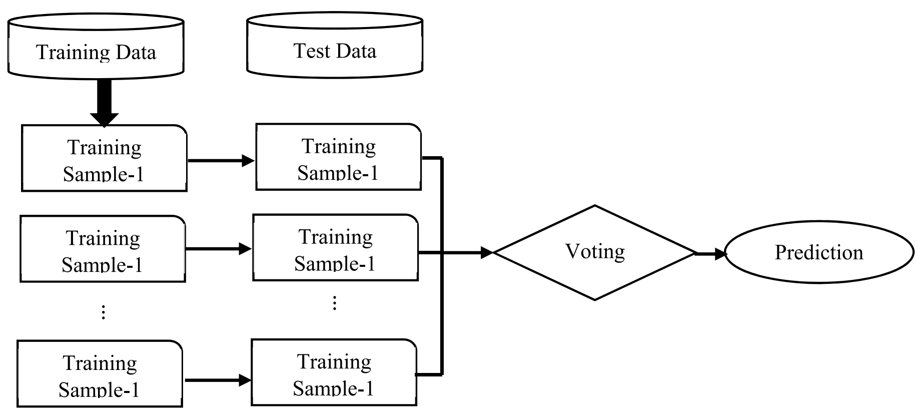

Fig. 3. Random forest classifier.

algorithm creates multiple decision trees (also called "forest") during training and selects a subset of features at every split point in the decision tree, rather than using all the features, like a traditional decision tree algorithm would do. This is done by using a technique called bootstrap aggregating or bagging, which helps to reduce the variance and overfitting that can occur in decision trees. The pictorial representation of Random Forest is depicted in Fig. 3.

A random forest classifier is typically an ensemble made with decision trees, where the training of every tree is performed on another subset of the data, and the final forecast depends on the consensus of all of the forest's trees, who each made a prediction. Each tree in the forest is trained on a different randomized subset of the data, and a randomized subset of the features is used at every split point in the tree. This randomization is useful to decrease the association between the trees present in the forest and makes the overall performance of the ensemble to improve. Random Forest is considered to be a very powerful and robust algorithm, it is easy to use and it can handle a large number of features and categorical variables. It is also less prone to overfitting compared to a single decision tree.

Extremely Randomized Trees: Extremely Randomized Trees (ERT) is a variation of the Random Forest algorithm in machine learning. It uses the same concept of an ensemble of decision trees, but the method for building the trees is slightly different. In Random Forest, the decision trees are built by repeatedly choosing a random subset of the features at every split point, and choosing the feature that gives the best split according to a certain criterion (e.g., information gain or Gini impurity). In contrast, ERT takes randomization one step further and randomly selects the threshold for each feature at each split point, instead of choosing the best threshold based on a criterion. The ERT structure is shown in Fig. 4.

This additional randomization allows ERT to introduce more randomness and diversity into the ensemble, making it less prone to overfitting and more robust to noise in the data. Also, it makes the algorithm faster to train because it doesn't need to find the optimal threshold for each feature. ERT used for classification and regression tasks and it is also relatively fast and easy to use.

Histogram-Based Gradient Boosting: Histogram-Based Gradient Boosting (HBGB) is a variation of the Gradient Boosting algorithm in machine learning. Gradient Boosting is, in fact a powerful ensemble technique, which builds a model by combining the predictions of multiple weak models, where every model is trained for rectifying the errors of the earlier models in the ensemble. HBGB utilizes histograms for approximating the underlying distribution of the data instead of the traditional approach of using a single decision tree. It builds a histogram for each feature, and for each feature, it uses the histogram to split the data into different bins. Then it applies a decision tree model to each bin. This allows the algorithm to approximate the underlying distribution of the data more accurately, which can lead to better performance. HBGB is particularly well-suited for datasets with a large number of features, or when the data is highly skewed or has outliers, since it can better handle

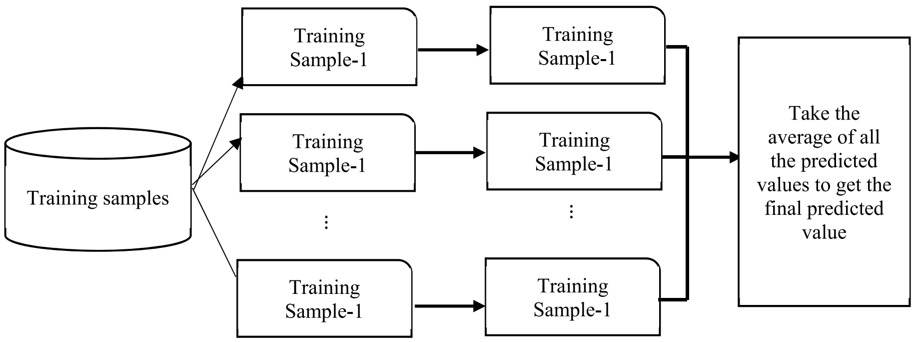

Fig. 4. Extremely randomized trees structure.

these types of distributions than traditional gradient boosting. It is also relatively fast and easy to use, and it can be parallelized to speed up training.

4. Simulation results

Here, the outcomes analysis of the suggested soft voting classifier is described and also discussed comparative analysis.

4.1. Dataset

In order to carry out the investigation, the stroke prediction dataset was used. This particular dataset has 4981 rows and 11 columns in total. There is either a 1 or a 0 for the value of the output column stroke. If you see the number 0, this means that there was no risk of stroke discovered, and if you see the number 1, this means that a risk of stroke was found. In this particular dataset, the chance of having the value 0 in the output column (stroke) is higher than the possibility of having the value 1 in the same column. There are a total of 248 rows in the stroke column that are assigned the value 1, while 4733 rows are only assigned the value 0. In order to get more precision, the data must first undergo preprocessing, which involves balancing the data. The overall number of stroke records and non-stroke records that were in the output column prior to it was preprocessed is shown in Fig. 5. The format of the data sample before preprocessed is reported in Table 1.

As can be seen from Fig. 5, this dataset has a number of imbalances that need to be addressed. The dataset has been made more equitable by the use of the Synthetic Minority Oversampling Technique (SMOTE) method.

4.2. Pre-processing

Before constructing a model, it is necessary to do data preprocessing

Table 2
Performance metrics of the proposed model.

	Precision	Recall	F1-Score
Stroke	0.96	0.98	0.97
Normal	0.98	0.95	0.97
Accuracy			**96.88%**

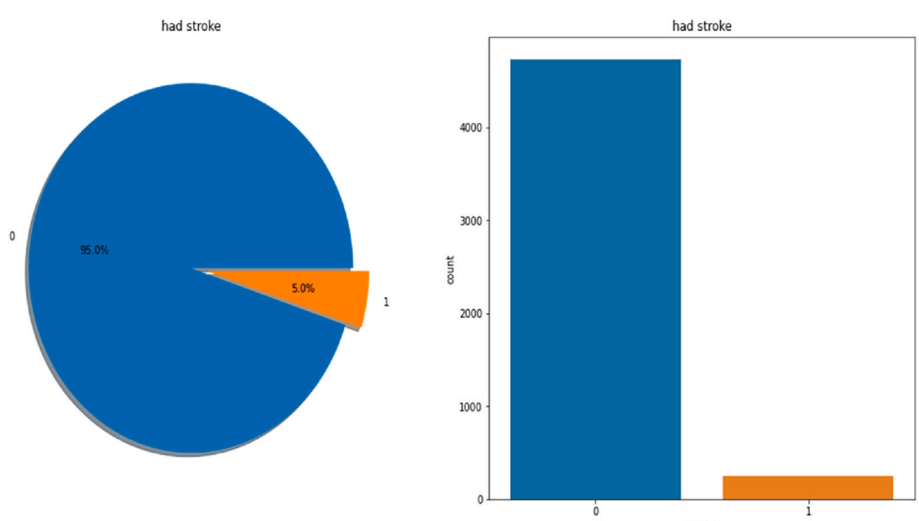

Fig. 5. The dataset before preprocessed.

Table 1
Sample data format.

	age	hypertension	Heart disease	Residence type	Avg Glucose Level	bmi	Smoking Status	stroke
0	67	0	1	Urban	228.69	36.6	formerly smoked	1
1	80	0	1	Rural	105.92	32.5	never smoked	1
2	49	0	0	Urban	171.23	34.4	smokes	1
3	79	1	0	Rural	174.12	24	never smoked	1
4	81	0	0	Urban	186.21	29	formerly smoked	1

Table 3
Performance metrics comparison with proposed model.

		Stroke			Normal		
Model	Accuracy	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Logistic Regression [15]	78.40%	0.81	0.73	0.77	0.76	0.84	0.80
KNN [16]	94.72%	0.96	0.93	0.95	0.94	0.96	0.95
SVM [17]	78.40%	0.83	0.72	0.77	0.75	0.85	0.80
Random Forest [18]	96.83%	0.95	0.98	0.97	0.98	0.95	0.97
Naive Bayes [19]	77.40%	0.77	0.78	0.78	0.78	0.76	0.77
Decision Tree [20]	94.03%	0.95	0.93	0.94	0.93	0.95	0.94
Extremely Randomized Trees	96.67%	0.96	0.98	0.97	0.98	0.96	0.97
AdaBoost	93.34%	0.90	0.97	0.94	0.97	0.90	0.93
Gradient Tree Boosting [21]	94.19%	0.90	0.99	0.94	0.99	0.90	0.94
Histogram-Based Gradient Boosting	96.40%	0.94	0.99	0.96	0.99	0.94	0.96
Proposed Soft Voting	**96.88%**	0.96	0.98	0.97	0.98	0.95	0.97

in order to clean the dataset of any unnecessary noise and outliers that can make the model to move from the training it was supposed to follow. During this stage, all of the problems, which impede the model from performing more efficiently are resolved. After the relevant information has been collected, it is necessary to clean and organize the data before moving on to the creation of the model. The dataset in question has twelve different features, as was previously mentioned. To begin, the column id is not included since its existence does not affect the manner in which the model is constructed. Next, the dataset is verified for any instances of null values, and those that are obtained are filled in. In this particular instance, the null values present in the column BMI are assigned in with the mean of the data column.

The textual forms found in the dataset are converted into integer values, which the device has learning via the process of label encoding. The strings need to be converted to integers as numbers are generally learnt by computers. The data set that was acquired comprises five columns that are of the string data type. During the label encoding step, every text is modified into a set of integers, and the whole dataset undergoes this transformation. The dataset that is being utilized for stroke prediction has a lot of inconsistencies. There are a total of 4981rows in the dataset, 248 of which show the possibility of a stroke and 4733 of which show that there was not a stroke. Although there is a possibility that making use of such data for training a machine-level model would yield accuracy, precision and recall values. In the event that such an imbalanced data is not handled with in the appropriate manner, the results will be wrong, and the prediction will not be accurate. As a result, one must first deal with the imbalanced data needed to get an effective model.

4.3. Evaluation of proposed model

In this system that we have proposed, we are making use of a machine learning technique that is based on a soft voting classifier. Other machine learning algorithms, such as Logistic Regression, KNN, SVM, Random Forest, Native Bayes, Decision Tree, Extremely Randomized Trees, AdaBoost, Gradient Tree Boosting, and Histogram-Based Gradient Boosting are used to test the proposed system. The best class models will be chosen for the voting classifier in accordance with their accuracy scores. The proposed model performance is reported in Table 2.

In this particular instance, the overall F1 score that was attained is equal to 97%. The F1-scores of healthy persons individually are 97%, whereas the individual F1-scores of those who have had a stroke are also 97%. After going through many rounds of fine-tuning, this model reached its maximum level of precision. The accuracy achieved with the model was 96.88%.

Table 3 presents the results of a comparison that was carried out using the proposed soft voting classification model for brain stroke classification and other classification models. The accuracy of the suggested model is 96.88%, which is higher than the earlier network models. The accuracy values of the other classification models, which include Logistic Regression, KNN, SVM, Random Forest, Native Bayes, Decision Tree, Extremely Randomized Trees, AdaBoost, Gradient Tree Boosting, and Histogram-Based Gradient Boosting, are as follows: 77.40%, 94.72%, 78.40%, 96.83%, 94.03%, 96.67%, 93.34%, 94.19%, and 96.40% correspondingly. Depending on the findings shown in Tables 3 and it is clear that the proposed soft voting classifier achieves higher levels of accuracy. The accuracy values of other classifiers, such as Random Forest, Extremely Randomized Trees, and Histogram-Based Gradient Boosting, are roughly comparable to those of the suggested technique.

5. Conclusion

Detection in advance of a brain stroke is critical for minimizing the damage to the brain and increasing the chances of recovery for the patient. Machine learning can have a significant part in detecting a stroke early by analyzing medical data. In this article, a model for the identification of strokes that is based on a soft voting classifier is proposed. The suggested model is an ensemble machine learning technique. It generates a final prediction by combining the findings of many independent classifiers, such as Random Forest, Extremely Randomized Trees, and Histogram-Based Gradient Boosting. The final prediction is arrived at by taking a weighted average of all of these probabilities, which are provided by each classifier as an estimate of the likelihood that it belongs to a certain class. It is possible to base the weights that are given to each classifier on how well they performed on a validation set; alternatively, the weights might be set evenly. When compared to a single classifier, the accuracy and resilience of the final prediction were both significantly improved by the suggested Soft voting model. The accuracy of the suggested model was determined to be 96.88.

Funding statement

Authors state no funding involved.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

https://www.kaggle.com/datasets/fedesoriano/stroke- prediction-dataset?resource=download &select = healthcare-dataset-stroke-data.csv

References

[1] V.L. Feigin, T. Vos, E. Nichols, M.O. Owolabi, W.M. Carroll, M. Dichgans, G. Deuschl, P. Parmar, M. Brainin, C. Murray, The global burden of neurological disorders: translating evidence into policy, Lancet Neurol. 19 (3) (2020) 255–265.
[2] S.K. Feske, Ischemic stroke, Am. J. Med. 134 (12) (2021) 1457–1464.
[3] Y.K. Cetinoglu, I.O. Koska, M.E. Uluc, M.F. Gelal, Detection and vascular territorial classification of stroke on diffusion-weighted MRI by deep learning, Eur. J. Radiol. 145 (2021).
[4] A. Viguier, L. Delamarre, J. Duplantier, J.M. Olivot, F. Bonneville, Acute ischemic stroke complicating common carotid artery thrombosis during a severe COVID-19 infection, J. Neuroradiol. 47 (5) (2020) 393–394.
[5] N.A. Terreros, M.L. Tolhuisen, E. Bennink, H.W. de Jong, L.F. Beenen, C.B. Majoie, E. van Bavel, H.A. Marquering, From perviousness to permeability, modelling and measuring intra-thrombus flow in acute ischemic stroke, J. Biomech. 111 (2020) 1–10.
[6] J.Y. Chen, X.K. Tu, Research progress on postoperative transient neurological dysfunction in pediatric and adult patients with moyamoya disease after revascularization surgery, Clin. Neurol. Neurosurg. 217 (2022).
[7] E.U. Haq, J. Huang, L. Kang, H.U. Haq, T. Zhan, Image-based state-of-the-art techniques for the identification and classification of brain diseases: a review, Med. Biol. Eng. Comput. 58 (2020) 2603–2620.
[8] R. Garg, E. Oh, A. Naidech, K. Kording, S. Prabhakaran, Automating ischemic stroke subtype classification using machine learning and natural language processing, J. Stroke Cerebrovasc. Dis. 28 (7) (2019) 2045–2051.
[9] S. Jayachitra, A. Prasanth, Multi-feature analysis for automated brain stroke classification using weighted Gaussian naïve Bayes classifier, J. Circ. Syst. Comput. 30 (10) (2021).
[10] M. Salucci, A. Gelmini, J. Vrba, I. Merunka, G. Oliveri, P. Rocca, Instantaneous brain stroke classification and localization from real scattering data, Microw. Opt. Technol. Lett. 61 (3) (2019) 805–808.
[11] A. Subudhi, M. Dash, S. Sabut, Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier, Biocybern. Biomed. Eng. 40 (1) (2020) 277–289.
[12] S.Y. Adam, A. Yousif, M.B. Bashir, Classification of ischemic stroke using machine learning algorithms, Int. J. Comput. Appl. 149 (10) (2016) 26–31.
[13] T. Badriyah, N. Sakinah, I. Syarif, D.R. Syarif, Machine learning algorithm for stroke disease classification, in: International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2020, pp. 1–5.
[14] V. Mariano, J.A. Tobon Vasquez, M.R. Casu, F. Vipiana, Brain stroke classification via machine learning algorithms trained with a linearized scattering operator, Diagnostics 13 (1) (2022) 1–19.
[15] V. Krishna, J.S. Kiran, P.P. Rao, G.C. Babu, G.J. Babu, Early Detection of Brain Stroke Using Machine Learning Techniques, 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1489–1495.
[16] L. Kommina, P. Theerthagiri, Y. Payyavula, P.S. Vemula, G.D. Reddy, Post-Stroke readmission prediction model using machine learning algorithms, Emerging trends in data driven computing and communications: proceedings of DDCIoT (2021) 53–65.
[17] U. Raghavendra, T.H. Pham, A. Gudigar, V. Vidhya, B.N. Rao, S. Sabut, J.K.E. Wei, E.J. Ciaccio, U.R. Acharya, Novel and accurate non-linear index for the automated detection of haemorrhagic brain stroke using CT images, Complex & Intelligent Systems 7 (2021) 929–940.
[18] C. Fernandez-Lozano, P. Hervella, V. Mato-Abad, M. Rodríguez-Yáñez, S. Suárez-Garaboa, I. López-Dequidt, A. Estany-Gestal, T. Sobrino, F. Campos, J. Castillo, S. Rodríguez-Yáñez, Random forest-based prediction of stroke outcome, Sci. Rep. 11 (1) (2021) 1–12.
[19] S.H. Rukmawan, F.R. Aszhari, Z. Rustam, J. Pandelaki, Cerebral infarction classification using the k-nearest neighbour and naive Bayes classifier, J. Phys. Conf. 1752 (1) (2021) 1–8.
[20] L.I. Santos, M.O. Camargos, M.F.S.V. D'Angelo, J.B. Mendes, E.E.C. de Medeiros, A. L.S. Guimarães, R.M. Palhares, Decision tree and artificial immune systems for stroke prediction in imbalanced data, Expert Syst. Appl. (2022) 191.
[21] R. Islam, S. Debnath, T.I. Palash, Predictive analysis for risk of stroke using machine learning techniques, IC4ME2, in: International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, 2021, pp. 1–4.