IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BRAIN STROKE DETECTION USING MACHINE LEARNING

B.Mamatha, R.Sreelatha, Dr M.Saravanamuthu
Madanapalle Institute of Technology and Science,Madanapalle,India.

Abstract

This paper provides a prototype of a text mining and machine learning-based stroke classification system. With the use of properly trained machine learning algorithms, machine learning may be portrayed as a significant tracker in fields like surveillance, medicine, and data management. This work uses data mining techniques to provide an overview of the monitoring of information from both a semantic and a syntactic standpoint. The suggestion is to extract patients' symptoms from case sheets and use the information to train the system. The case sheets of 507 patients from the Sugam Multispecialty Hospital in Kumbakonam, Tamil Nadu, India, were gathered during the data collection phase. The case sheets were then mined using maximum entropy and tagging approaches, and the suggested stemmer extracts the common and distinctive set of properties to categorise the strokes. Following that, a variety of machine learning methods, including artificial neural networks, support vector machines, boosting and bagging, and random forests, were fed the processed data. With a classification accuracy of 95% and a standard deviation of 14.69, artificial neural networks trained using a stochastic gradient descent approach surpassed the other algorithms.

- Introduction

Health is regarded as a vital component of everyone's life, therefore there is a need for a system of storing information about diseases and their connections. The majority of information on diseases can be found in patient case summaries, clinic medical records, and other manually maintained records. Through various text mining and machine learning approaches, the sentences in them may be understood (ML). In information retrieval, when the semantic and syntactic components of the content are given priority, machine learning is a technology that may distribute the content. For feature extraction and classification, many ML and text mining approaches are suggested and put into practice..

The majority of medical professionals refer to lesions to the brain and spinal cord caused by irregularities in blood flow as strokes. Different viewpoints affect how a stroke is perceived; yet, stroke generally elicits a visceral reaction that is clear. In a brain, which weighs more than three pounds and is made up of 100 billion neurons and a trillion glia, respectively, each memory is encoded and stored in a network. Each and every person's breathing and movement are supported by brain activity. Since 1970, or more than 50 years, ten times more people have died from strokes in developing nations than in developed ones, and by 2030, the number is expected to quadruple globally.Generally, stroke is classified into the following three types:

strokes that are ischemic, hemorrhagic, and transient ischemic attacks (TIA). The most typical kind of stroke is an ischemic stroke. According to the American Heart Association (AHA), ischemic stroke, which happens when a clot or other obstruction persists in a brain blood channel, accounts for 87% of all strokes [1]. Thrombotic stroke and embolic stroke are the two types of ischemic strokes [2]. When a block or clot forms in any part of the body and travels to the brain, blocking blood flow, this is known as an embolic stroke. A clot that reduces blood flow in an artery that supplies blood to the brain causes thrombotic stroke. A weakened blood vessel splits or bursts, resulting in a hemorrhagic stroke. Hemorrhagic strokes are expected to account for only around 10% to 15% of all strokes, although they have a higher fatality rate than ischemic strokes [3-5]. Subarachnoid haemorrhage and intracerebral haemorrhage are the two kinds of hemorrhagic stroke. A clot causes a transient ischemic attack, which is referred to as a "mini-stroke." Compared to other types of stroke, a TIA is a temporary blockage that lasts just briefly (on average, 1 minute), with symptoms going away within 24 hours. Although TIA does not result in long-term damage to the brain or its structures [6], it is regarded as a warning sign for the impending occurrence of another stroke. No of the type, stroke is typically thought to be fatal disease. This study focuses on developing methodologies to extract the base form of the text from

patient case summaries or medical records, retrieve the root word or stem from the base form through stemmers, and classify the type of stroke as ischemic and hemorrhagic stroke from the retrieved root words (based on their common symptoms). Stroke victims are constantly in danger despite all these detections. need of intensive care, which can be provided by an interdisciplinary team.

- Related works

Few studies are utilising machine learning (ML) methods to predict strokes. In this section, significant contributions to research are described. In a recent study, the multilayer perceptron (MLP) algorithm-trained artificial neural networks (ANN) technique was used to predict stroke patients' death, and the results showed an accuracy of 80.7% [7]. Another study automated the detection of ischemic stroke using support vector machines (SVM), k-nearest neighbours (kNN), and artificial neural networks (ANN), which revealed that SVM had superior prediction accuracy [8, 9]. When Amini et al. [10] compared the accuracy of C4.5 decision tree methods to k-nearest neighbour approaches for predicting stroke occurrence, they found that C4.5 decision tree methods had a better accuracy rate of 95.42%. Another group [11] employed SVM and machine learning techniques to forecast the outcome of stroke thrombolysis, which showed that SVM was more accurate. Using two ANN models, Cheng et al. [12] predicted ischemic stroke with accuracy rates of 79.2% and 95.1%. One study [13] forecasted the presence of stroke using the knowledge discovery process (ANN and SVM). The findings of this study revealed that ANN outperformed other models in terms of prognostic performance. RDFs and CNNs have higher classification accuracy than other methods, according to Maier et al. [14], who used nine classification techniques to classify ischemic stroke, including generalised linear models, random decision forests (RDFs), and convolutional neural networks (CNNs). In order to predict stroke, Kansadub et al. [15] employed decision trees (DTs), naive Bayes, and ANN and found that DT produced better categorization than the other two approaches. Multiple linear regression was used by Sung et al (MLR), kNN's superior performance over other models was shown when used to predict the stroke severity index (SSI) using both a regression tree model and kNN. A prediction model comprising DT, ANN, SVM, logistic regression (LR), and ensemble generalised boosted model (GBM) was given in another work [17] to predict the transfer of stroke patients to the ICU, and it was found that GBM had the highest accuracy.

In their research, Adam et al. [18] explore the categorization of stroke using machine learning approaches and analyse a number of previous studies from the standpoint of classification. Their research examined the decision tree and k-nearest neighbour algorithms (KNN). It was found that the decision tree algorithm outperformed the KNN algorithm. Despite the availability of numerous ischemic stroke diagnostic tools, a new study claims that the aetiology of the stroke patients is still unknown [19]. The study came to the conclusion that the phenotypic type of stroke categorization was crucial, but it also discussed how inaccurate and unreliable it was.

Deep learning is a method Chantamit-O-Pas et al. [20] suggest for predicting strokes. Traditional predictive algorithms could not accurately trace the knowledge of medical domain issues. The study's findings showed that predicting a stroke was more accurate than using a grading system from the medical field. The Asian Stroke Advisory Panel conducted a survey from several angles in 12 different nations in 13 Asian regions. According to the survey, a bigger percentage of persons were discovered to be more susceptible to ischemic stroke in Asian countries. In Asia, there are between 116 and 483/100,000 stroke victims annually. Additionally, they noticed a threefold increase in the number of neurologists across all of the nations [21].

To increase classification accuracy, 22 attributes derived from real-time data gathered from the Multispecialty Hospital (Table 1) were examined in this study using text mining and machine learning techniques.

- Motivation and objective of the study

The most common errors in the literature were: the majority of research works only contribute to ischemic stroke (IS) type; the impact of risk factors for stroke and its classification are not given adequate weight in the research; the majority of research works classified stroke using only two or three ML algorithms; and classification of stroke using a collated data obtained from case sheets and case summaries is not attempted. In this work, mining techniques are suggested as a way to get over the problems outlined above and accurately distinguish the different types of strokes.

The majority of research works classified stroke using only two or three ML algorithms; classification of stroke using a collated data obtained from case sheets and case summaries is not attempted; the impact of risk factors for stroke and its classification are not given adequate weight in the research. These were the most frequent errors in the literature. The use of mining techniques is recommended in this work as a means of overcoming the issues mentioned above and accurately differentiating the various sorts of strokes. According to many neurologists, there is currently no treatment that can totally heal a stroke. Instead, we have supportive palliative care, which is likely to increase a person's lifespan. In developing nations, there were ten times as many people who lost their lives to strokes, and by 2030, this number is expected to double globally [22]. A study that was conducted in two phases by the Canadian Institutes of Health Research and the Heart and Stroke Foundation showed the influence of a patient's risk factor in the development of stroke [23, 24].. As per the report given A higher percentage of people were reported to be impacted by ischemic stroke by the Asian Stroke Advisory Panel [21]. Consequently, a precise

classification is required to reduce the severity of symptoms in such disorders. In order to classify this lethal disease, the study suggests strategies for extracting data from case sheets and medical records. The findings of the study may assist medical professionals in understanding the severity of the disease and making appropriate decisions.

- The main goal of this research project is to collect stroke datasets and categorise different types of strokes using machine learning and mining methods. The stroke is tagged, stemmed, and classified in order to accomplish the main goal. Based on them, the following sub-objectives are developed:
- 1. To use tagging and maximum entropy techniques to extract the pertinent information from the raw data.
- 2. Using a novel method, retrieve the processed dataset stemming algorithm and avoid the discrepancies related to the size of the words and stemming errors found in the earlier studies.

3. To assign weight to the variation in the dataset in order to reasonably accurately categorise the type of stroke.The essay is set up like follows: Sect. 2 introduces the suggested prototype, Sect. 3 details the findings and discussion, and Sect. 4 compiles the findings.

from the proposed prototype.

- Proposed prototype

Table 1 Parameters considered in the study as suggested by stroke specialists

Variable name (features)	Extracted feature from the dataset	Variable name (features)	Extracted feature from the dataset
X1	Patient number	X13	Patient with severe headache
X2	Age of the patient	X14	Patient with vomiting
X3	Gender of the patient	X15	Patient with weakness
X4	Patient with numbness	X16	Patient with giddiness
X5	Patient with loss of consciousness	X17	Patient with facial palsy
X6	Patient with diplopia	X18	Patient with nausea
X7	Patient with dysarthria	X19	Patient with aphasia
X8	Patient with difficulty in walking	X20	Patient with altered sensorium
X9	Patient with difficulty in speaking	X21	Patient with hypertension (HT)
X10	Patient with loss of memory	X22	Patient with diabetes mellitus (DM)
X11	Patient with swallowing difficulties	X23	Class of stroke {ischemic (IS), hemorrhage (HE)}
X12	Patient with paralysis		

The three key phases of the suggested prototype are data collection, pre-processing, and categorization. Fig. 1 depicts the proposed workflow diagram. Using tools with tagging and maximum entropy methods, data were gathered from the case sheets that were obtained from the hospital. Correlation analysis is then used to pre-process the data from the acquisition phase.

Fig. 1 Workflow diagram of the proposed prototype

Redundancies, often known as data duplication or repetition, should be eliminated. Next, various machine learning algorithms are fed the preprocessed data to do classification.

- Data acquisition

Patient case sheets from Sugam Multispecialty Hospital in India were used to collect the data. The case files included data from 507 stroke patients, ranging in age from 35 to 90. There were discovered to be 22 distinct class labels for stroke that belonged to the two main forms of stroke: ischemic stroke and hemorrhagic stroke. During classification, risk variables like diabetes mellitus and hypertension were also considered. To obtain the dataset for the categorization of the kind of stroke, the case sheets were pre-processed using the base-form generator and cutting-edge stemmer algorithms.

generator of base-forms The programme used to analyse English phrases and provide base forms, chunk tags, part-of-speech (POS) tags, and named entities is called Genia tagger [25]. (NE). This programme was created for the analysis of biomedical language, such as MEDLINE entry abstracts [26]. A bidirectional dependency network was created by Toutanova et al. [27] and used by Tsuruoka et al. [25] to design the POS tagging method, a part-of-speech tagging technique. The GENIA tagger's POS component is trained using Wall Street Journal corpus, GENIA corpus, GENIA POS corpus [28], and PennBioIE corpus [26].

GENIA tagger is implemented in the UNIX operating system. The following is its output format:

Word1 Base1 POS1 Chunk1 NE1

Word2 Base2 POS2 Chunk2 NE2

The input provided in this tagger is an original plain text (one sentence/line/paragraph). The output is projected with a single token separated by tabs. The output contains surface, base form, POS, chunk tags, and the information about NE.

The significant perspectives of Genia are tagging and maximum entropy methodologies, which take the form shown in Eq. (1):

$$p(tjc) = \frac{1}{Z(C)} \exp\left(\sum_i k_i f_i(C;t)\right) \quad (1)$$

where $f_i(C;t)$ represe is used to add up n items, where ki stands for weight and Z(C) for the constants used for normalisation; as a whole, it (1) is employed as a classifier for text classification. nts are the characteristics of content C that were used to P discover the tag t. Only from the information provided about the data are assumptions made using the maximum entropy method. A rule-based algorithm is used for tagging, correlating various phrases and POS with the tags. One machine learning approach that helps with the POS tagging issue is the maximum entropy classifier, which can achieve an accuracy of 95% [1].
fresh stemmer A novel stemming algorithm that covers higher numbers of words to be stemmed is proposed while taking into consideration several of the affix removal stemmers. By deleting the suffixes and prefixes, this technique yields the stem or root form. The following is a list of amended rules:

- Any word that ends in "ies" is changed to "y" as the suffix.
- Any word that ends in "er" gets stripped of its suffix and given the suffix "null."
- Any given word that ends in "es" has its suffix changed to "e."
- The prefix of any word that ends in "ment,"

"ing,' "ed," is deleted, and "null" is added in its place.

Therefore, using the aforementioned guidelines that remove and replace terms, more word content could be stemmed. A collection of data is obtained through the base-form generator and novel stemmer methodology consisting of defined parameters that must be pre-processed to obtain the dataset.

- Data pre-processing

To improve the quality of the data for use, pre-processing is essential. The characteristics that indicate the data quality are accuracy, interoperability, and reliability. Data cleansing, data integration, data reduction, and data transformation are just a few of the several steps that make up data pre-processing. Data cleaning fills in missing terms and corrects many discrepancies, even in noisy data. Data integration is the process of combining information from many sources into a single, cohesive data set. One of the biggest problems in the pre-processing stage of data integration is redundant data, whereas data reduction lowers bulky data. Redundancy is a major issue in data integration, as was previously mentioned. Redundancy can occur for a variety of reasons, including attribute naming, consistency, and whether the attribute is taken from another set of attributes. Correlation analysis, which manipulates the correlation coefficient for numerical data and the Chi-square test for nominal data, can identify it. Online solutions like Data Cleanser and Merge/PurgeLibrary (Sagent/QMSoftware), which contain user-specific matching rules for integration, are used to prevent data duplication [29]. Data transformation is the process of transforming data into a format suitable for mining. These are a few of the steps in the data pre-processing process.

The data are pre-processed to extract the patient's symptoms and risk variables as pre-defined parameters for stroke classification and type prediction. The features that were taken from the case sheets and data samples are shown in Tables 1 and 2, respectively.

The various machine learning methods are given pre-processed data of 507 samples with 22 features (excluding X1, the patient number), and the implementation details are provided in the next section.

- Classification

The classification method accurately predicts the target class of each tuple in the given data. Pre-processed data are fed into different classification algorithms to estimate the accuracy of each classification method.

ROC Receiving operating characteristic (ROC) curves are employed in this work to depict classification accuracy. A continuous variable X that is "value" computed for each instance is used in dichotomous classification to predict the outcomes of cases. Instances are categorised according to a threshold T if X[T, then it is positive; otherwise, it is negative. If the instance is positive, X would be followed by the probability density function f1x. If the instance is negative, X would

be followed by the probability density function f0x. As a result, the following provides the genuine positive rates and false positive rates:

True positive rate ðTÞ ¼ $\int_T^{Z_1} f_1(x)dx$	ð2Þ
False positive rate ðTÞ ¼ $\int_T^{Z_1} f_0(x)dx$	ð3Þ

Hence, ROC curves plot the true positive rate versus the false positive rate with threshold T as the varying parameter [30].

- ANN

- 22 inputs and one hidden layer with ten neurons and two outputs make up an artificial neural network (ANN) (IS and HE). The stochastic gradient descent approach is used to train the network. A stochastic approximation of the gradient descent optimization is stochastic gradient descent, often known as incremental gradient descent. Additionally, it aids in differentiable objective function optimization through an iterative process. Iteration is used to discover maxima or minima [31]. The pseudocode for stochastic gradient descent is as follows:Initial vector of parameters w and learning rate g is chosen
- Repeat the following steps until an approximate minimum is obtained:
  - Examples are randomly shuffled in the training set
  - For i = 1 to n, do w ¼ w grQ$_i$(w)

example.

Table 2 Sample dataset

X2	X3	X4	X5	X6	X7	X8	X9	X_{10}	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23
96	M	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	IS
75	M	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	IS
45	F	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	HE

where Q(w) represents the empirical risk and Q$_i$(w) represents the value of the loss function at ith

An artificial neural network (ANN) is composed of one hidden layer with ten neurons and two outputs and 22 inputs (IS and HE). The network is trained using the stochastic gradient descent method. Stochastic gradient descent, also referred to as incremental gradient descent, is a stochastic approximation of the gradient descent optimization. It also supports the iterative process of differentiable objective function optimization. Maxima or minima are found via iteration [31]. The following is the stochastic gradient descent pseudocode:14.69.

- SVM

Another classification technique for predicting strokes is called SVM. It was created from statistical learning theory and is used extensively in a variety of fields, including bioinformatics and image recognition. A prototype is built by the SVM training algorithm that adds new entities to an existing group or forms a new group.

SVM, also known as supervised machine learning, is employed for both classification and regression problems. Each entity from the dataset is plotted as a point in n-dimensional space (n is the number of characteristics) by this approach. The value of a certain coordinate is taken into consideration for each feature. To classify something, one must identify the hyperplane that separates the two classes.

The separating function in SVM is described as a linear combination of kernels linked with support vectors

$$f(x) = \sum_{x_j \in S} a_j y_j K(x_j; x) + b \qquad (4)$$

In Eq. (4), x_j indicates the patterns of the training set, y_j [{? 1, - 1} indicates the respective class labels, and S indicates a set of support vectors [32, 33].

Several kernels, including linear SVM, quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, and coarse Gaussian SVM, are used to classify the pre-processed samples. The approaches with the highest accuracy, 91.5%, were linear SVM, medium Gaussian SVM, and coarse Gaussian SVM. With a training time of 2.28 seconds, the different SVM kernels' ROC curves showed the highest accuracy in Figs. 2, 3, and 4.

- Decision tree

A decision tree builds a tree structure using regression or classification models. The dataset is divided into more manageable subsets, and a decision tree is created. Tree organises the dataset, but it is unable to learn about itself [34, 35] using the patient as an example. Every dataset belongs to one of the designated classes. As a result, it is considered supervised learning as opposed to unsupervised learning. A single pruning technique is used to construct the tree utilising the information gained, and sustained improvements are expected. It categorises all data kinds, including continuous, discrete, brief, and easy to infer ones, and the throughput is always one that is legible by humans.

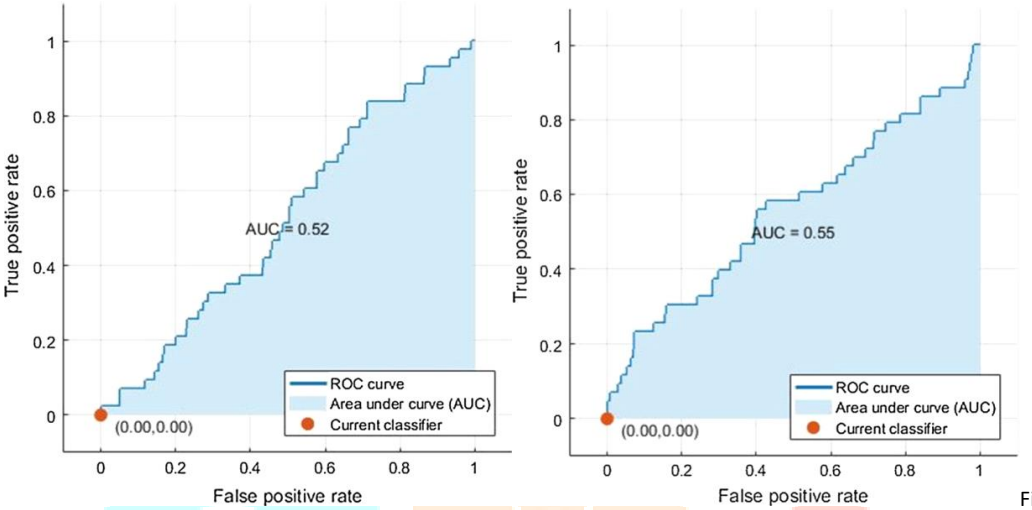

Fig. 2 Linear SVM

Fig. 3 Medium Gaussian SVM Simple example for classification tree:

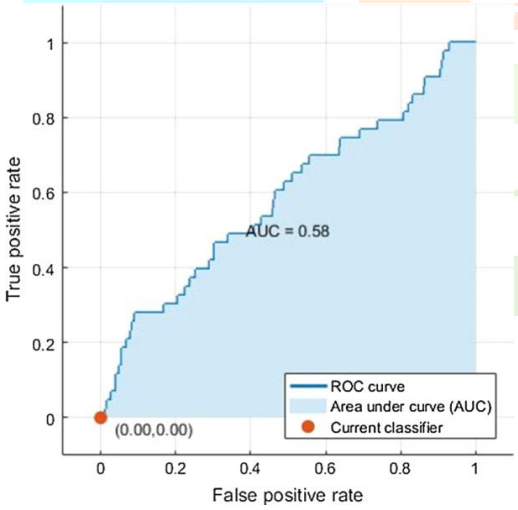

Fig. 4 Coarse Gaussian SVM

The prediction begins at the tree's root node (D) and uses the first predictor (X1) to verify the outcome: It follows the left branch and the tree classifies its forecast as type 0 if its value is less than 1.0; otherwise, it follows the right branch and makes another prediction based on the second predictor, X2. If the X2 value is less than 1.0, the prediction is classified as type 0 and follows the left branch; otherwise, it follows the right branch and is classified as type 1. as type 1 [34].

The tree generates accuracy with the help of kernels such as a simple tree, medium tree, and complex tree, which are shown in Figs. 5, 6, and 7. The simple tree produced the highest accuracy (90.7%) as compared to the others with a training time of 1.45 s.

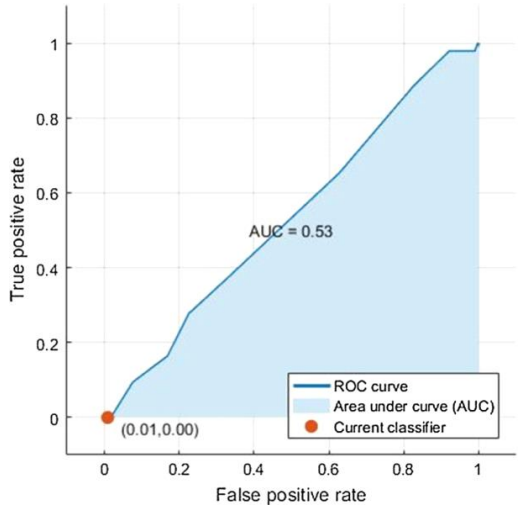

Fig. 5 Simple tree (20 splits)

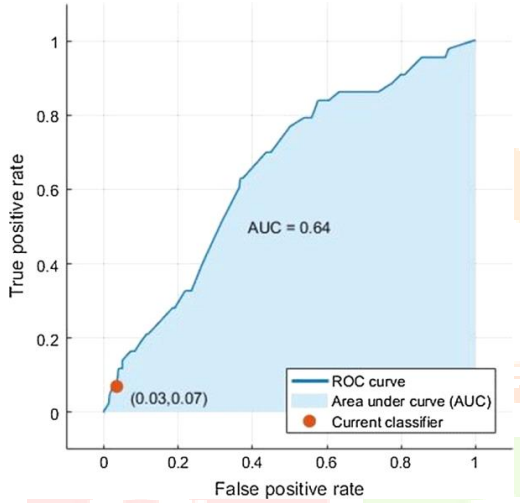

Fig. 6 Medium tree (60 splits)

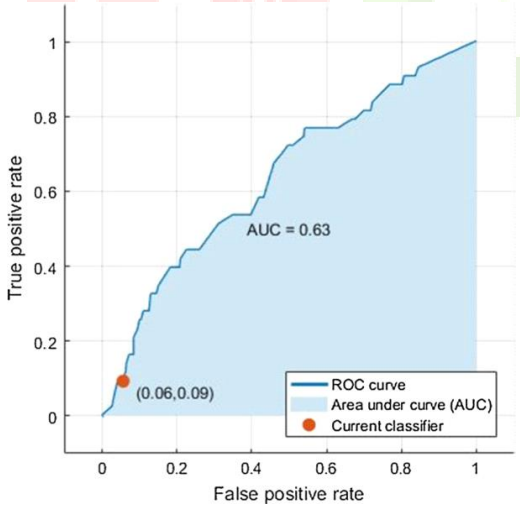

Fig. 7 Complex tree (100 splits)

- Logistic regression

Predictive analysis, which summarises the data and shows how independent variables and dependent (binary) variables relate to one another, is the foundation of logistic regression (LR). Does the patient's age, degree of hypertension, and diabetes mellitus affect the stroke patient, for instance? Other predictors are treated as variables and the outcome of the process is either 0 or 1, which is labelled as dependent [36]. This concept is utilised in many different sectors, such as marketing and machine learning to anticipate the existence of disease (based on characteristics). In this work, 10-fold cross-validation was performed after the

sample data were put in, and the LR approach yielded an accuracy of 90.6% with a training duration of 8.51 s. The ROC curve for this kind is displayed in Fig. 8.

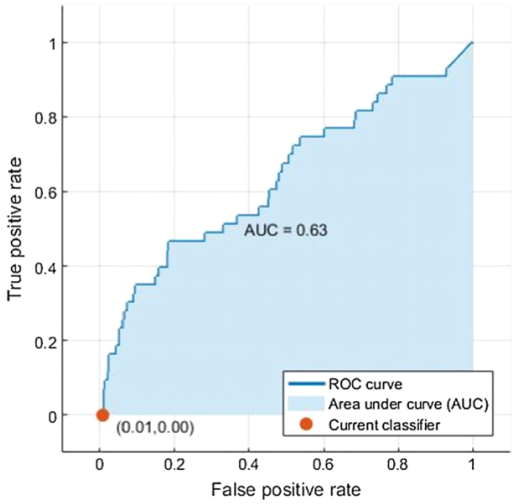

Fig. 8 Logistic regression

- Bagging and boosting

A approach known as the ensemble method combines predictions from various machine learning algorithms to produce predictions that are more accurate than those from any one model. An ensemble of models (classifiers or predictors), which is a learning strategy where each model provides an equally-weighted prediction, is created by the bagging method. The bagging classifier type employs random forest as its methodology.

The random forest algorithm operates in two stages. In the first stage, creation of random forest is carried out, followed by a prediction from random forest classifier that was created in the first stage [37]. The following is the pseudo-code for the creation of random forest:

Random selection of ‘‘m’’ features from the total ‘‘n’’ features, where m n
Among the ‘‘m’’ features, the node ‘‘x’’ calculation is done by means of best split point
Node is divided into children node using the best split
Repeat the steps 1–3 until ‘‘y’’ number of nodes are reached
Repeat the steps 1–4 for ‘‘N’’ number of times to build forest as well as to create ‘‘n’’ number of nodes

The following is the pseudo-code for prediction through random forest classifier (first stage):

Store the predicted outcome as the target by using the rules as well as test features of a randomly created decision tree
Predicted target’s votes are calculated
The predicted target with high votes is considered as the final prediction among the random forest algorithm Boosting algorithm creates an ensemble of classifiers; each one gives a weighted vote. The method used in the boosting classifier type is AdaBoost [38].

AdaBoost indicates a method of training a boosted classifier, and it takes the following form (5):

$$F_T(x) = \sum_{t=1}^{X_T} f_t(x) \quad (5)$$

where f_t is indicated as a weak learner which takes x as the input object and it returns the class of the object as a value. Tth classifier indicates positive if the object is in positive class else it falls under negative class.

The following is the pseudo-code of AdaBoost:

An initial weight value, $w_i = 1/n$ (n, represents the number of total observations) is assigned to each observation, X_i
‘‘Weak’’ model is trained (often a decision tree)
In each observation,

3:1. w_i is increased, for incorrect prediction
3:2. w_i is decreased, for correct prediction

A weak model is trained by giving more priority to higher weights in the observations.

Steps 3 and 4 are repeated until the observations predict perfectly or until a stipulated number of trees are trained.

Since If a (basic) algorithm fails to accurately categorise the objects, one can combine a variety of classifiers with the chosen training set (in each iteration) and by appropriately weighting the classifiers (final voting). One can get overall good accuracy with this strategy [39].

AdaBoost and random forest, two algorithms, respectively offer accuracy of 90.9% with a training time of 7.74 s and 91.5% with a training time of 7.24 which are shown in Figs. 9 and 10.

- Results and discussion

The dataset and parameters (patient symptoms) used for the investigation are provided in Table 1. The data processing stage, where the novel stemmer technique was utilised to obtain the dataset, is where the work's originality lies. The collected information (507 patients) included the patients' ages, which ranged from 35 to 90 years, and 22 distinct class labels (parameters) that are either associated with ischemic stroke or hemorrhagic stroke (Table 2).

The collected dataset demonstrated that 91.52% of patients were affected by ischemic stroke, whereas 8.48% of patients were affected by hemorrhagic stroke (Fig. 11). The manipulated outcome of the dataset showed that

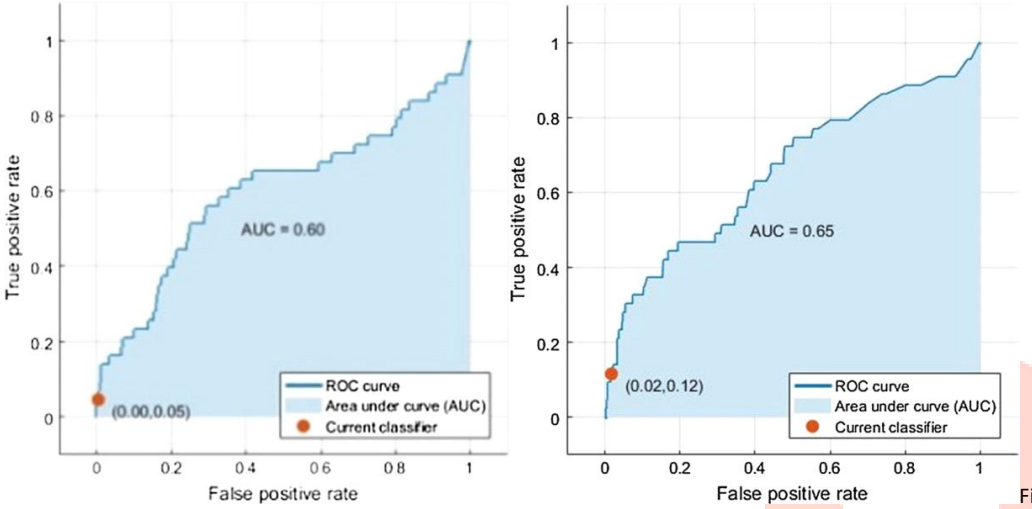

Fig. 9 Bagging

Fig. 10 Boosting

Fig. 11 Prevalence percentages of stroke and its type

weakness was the primary symptom in 51.93% of ischemic stroke patients and 37.20% of hemorrhagic stroke patients. The dataset also showed that 37.50% of ischemic stroke patients and 23.25% of hemorrhagic stroke patients were affected by dysarthria, giddiness (in 36.20% and 13.95% of ischemic and hemorrhagic stroke patients, respectively), difficulty in walking (in 24.56% of ischemic stroke and

23.25% of hemorrhagic stroke patients), and vomiting (in 18.75% of ischemic stroke and 25.58% of hemorrhagic stroke). Other symptoms such as numbness.

Facial palsy, loss of consciousness, decreased responsiveness, diplopia, severe headache, altered sensorium, aphasia, and nausea affected less than or equal to 8% and 28% of ischemic stroke patients and hemorrhagic stroke patients, respectively (Figs. 12, 13).

90% of the dataset that was gathered was used to test the trained data. The generated model yielded the smallest amount of prediction error. The classification took into account the patient's symptoms as well as aspects including age, gender, HT, and DM. The classification methodology's findings are displayed in Table 4, which also displays the classification assessment metrics of recall, accuracy, sensitivity, specificity, and sensitivity that are displayed in Table 3. Table 4 displays the confusion matrix and related findings for the aforementioned classification approaches. To determine whether the best accurate classifier with the least deviation was statistically significant, the standard deviation between the assessment metrics was calculated. Fig. 14 displays the accuracy for all classifier categories and demonstrates that artificial neural networks trained with stochastic gradient descent algorithm have the highest accuracy (95.3%) for classifying stroke as compared to other classifiers.

This work emphasises the critical role that symptoms and risk factors have in classifying stroke as well as showing how each symptom affects the occurrence of either type of stroke (ischemic or hemorrhagic) in a patient, which increases the uniqueness of the work. Utilizing the information obtained from patient data sheets, classification is carried o

Previous studies demonstrated that the decision tree algorithm and kNN had the lowest possible prediction error when classifying ischemic strokes. Even using medical techniques, the study claims that it is challenging to precisely categorise the various ischemic stroke kinds. The conclusion of the whole study

S. no.	Model name	Accuracy	Sensitivity	Specificity	Recall	Precision	Standard deviation
1	Simple tree	90.7	99.1	0	91.4	99.1	38.20
2	Medium tree	89.0	96.5	6.9	91.8	96.5	34.73
3	Complex tree	87.2	94.3	9.3	91.8	94.3	33.14
4	Logistic regression	90.6	99.1	0.0	91.4	99.1	38.19
5	Linear SVM	91.5	1.0	0.0	91.5	1.0	44.50
6	Quadratic SVM	89.5	96.3	11.6	92.1	96.3	32.88
7	Cubic SVM	88.3	95.0	16.2	92.4	95.0	30.68
8	Fine Gaussian SVM	91.1	99.1	4.6	91.8	99.1	36.43
9	Medium Gaussian SVM	91.5	1.0	0.0	91.5	1.0	44.50
10	Coarse Gaussian SVM	91.5	1.0	0.0	91.5	1.0	44.50
11	Ensemble boosted tree	90.9	98.2	11.6	92.3	98.2	33.45
12	Ensemble bagged tree	91.5	99.5	4.6	91.8	99.5	36.55
13	Ensemble RUS boosted tree	63.1	62.9	65.1	95.1	62.9	12.66
14	Artificial neural network	95.3	95.9	60	99.2	95.9	14.69

Table 4 Performance measurements for various classification methods

demonstrates that the decision tree algorithm outperforms the kNN algorithm [10]. In the majority of research studies, at least two to three algorithms were employed to categorise strokes [8, 10, 12, 13, 15, 18, 19], but this paper shows how to categorise strokes utilising a variety of classifiers. The classification process employed patient symptoms and risk variables to categorise the different types of stroke. The ANN classifier classified the kind of stroke with greater than 95% accuracy, zero negative predictive value, and a standard deviation below 14.69. either as ischemic or hemorrhagic, which is higher when compared with the previous research work [7, 12].

Our work classified stroke type utilising more than five classifiers together with their categories as compared to the prior work, which only employed two methods for classification. When compared to earlier work, which only classifies a portion of a stroke, the accuracy and standard deviation of all classifiers were novel.

Even though hemorrhagic stroke has the highest death rate compared to the other form of stroke, the classification of both types of stroke does not receive the attention it deserves in research studies. This study's originality is further enhanced by the classification of both types of stroke using a variety of classifiers and their kernels. To put it simply, the majority of classifications assist medical professionals in identifying the type of stroke [18,40].

In Table 3, TP stands for true positives, i.e., samples identified as being positive by the classifier; TN stands for true negatives, i.e., samples identified as being negative by the classifier. The deviation between these metrics is derived using standard deviation

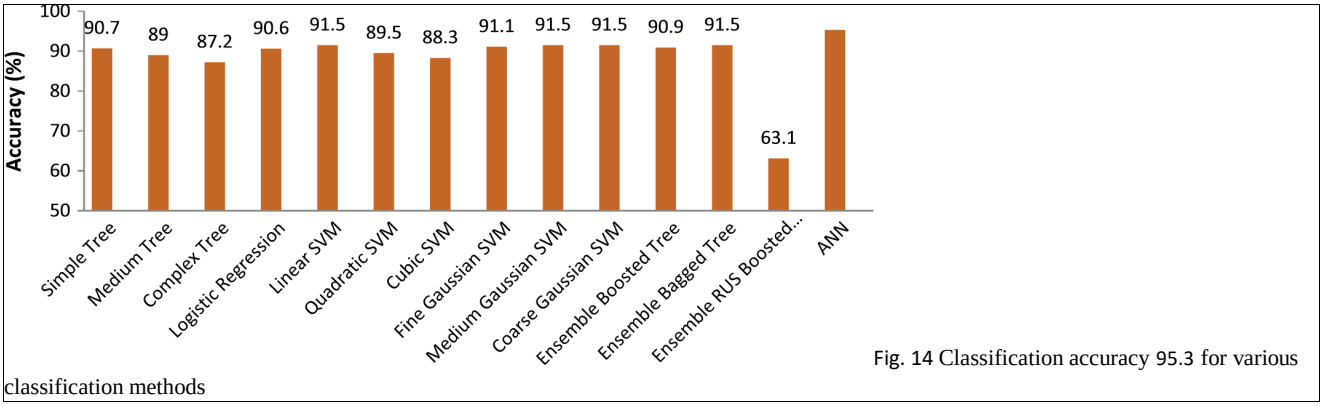

Fig. 14 Classification accuracy 95.3 for various classification methods

- Conclusion

The study highlights the usefulness of categorization techniques for structured entities, such as patient case sheets, in categorising strokes according to specified characteristics (symptoms) and circumstances. Based on classification approaches, this study forecasts the type of stroke that a patient would experience. SVM and ensemble (bagged) categories offered 91%

accuracy with 0.0000 negative predictive value, whereas ANN trained with the stochastic gradient descent approach surpassed other algorithms with a classification accuracy of 95% and a standard deviation of 14.69. According to this study, stroke is more common in males than in women and in people between the ages of 40 and 60. Patients with ischemic stroke were more prevalent in number than patients with hemorrhagic stroke. Determining the type of stroke depends not only on the impact of modifiable and non-modifiable risk factors of the patient but also on individual patient's symptoms.

Acknowledgements We are grateful to Dr. Sundarrajan S, a neurologist at Sugam Multispecialty Hospital, for allowing us access to the patients' real-time data and for his insightful advice on how to categorise the different types of strokes. We also acknowledge the management of Kumbakonam's Sugam Multispecialty Hospital for helping us collect the case sheets. We thank the Department of Science and Technology, India, for funding this research project through an INSPIRE fellowship (No. IF120649). The second author also expresses gratitude to the Department of Science and Technology for grant funding. No.SR/FST/ETI-349/2013.

- Compliance with ethical standards

Conflict of interest There is no conflict of interest among the authors to publish this article.

- References

- Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD,

1. Santos EMM, Yoo AJ, Beenen LF, Majoie CB, Marquering HA(2016) Observer variability of absolute and relative thrombus density measurements in patients with acute ischemic stroke. Neuroradiology 58(2):133–139
- Rebouc¸as ES, Marques RCP, Braga AM, Oliveira SAF, deAlbuquerque VHC, Filho PPR (2018) New level set approach based on Parzen estimation for stroke segmentation in skull CT images. Soft Comput. https://doi.org/10.1007/s00500-018-3491-4
- Shinohara Y, Yanagihara T, Abe K, Yoshimine T, Fujinaka T,Chuma T, Ochi F, Nagayama M, Ogawa A, Suzuki N, Katayama Y, Kimura A, Minematsu K (2011) Cerebral infarction/transient ischemic attack (TIA). J Stroke Cerebrovasc Dis 20(4):S71–S73
- Su¨t N, C¸elik Y (2012) Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turk J Med Sci 42(5):886–893
- Rajini NH, Bhavani R (2013) Computer aided detection ofischemic stroke using segmentation and texture features. Measurement 46(6):1865–1874
- Sundstro¨m C (2014) Machine learning algorithms for stroke diagnostics. Master's thesis in biomedical engineering
- Amini L, Azarpazhouh R, Farzadfar MT, Mousavi SA, Jazaieri F,Khorvash F, Norouzi R, Toghianfar N (2013) Prediction and control of stroke by data mining. Int J Prev Med 4(2):S245
- Bentley P, Ganesalingam J, Jones AL, Mahady K, Epton S, RinneP, Sharma P, Halse O, Mehta A, Rueckert D (2014) Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage Clin 4:635–640
- Cheng CA, Lin YC, Chiu HW (2014) Prediction of the prognosisof ischemic stroke patients after intravenous thrombolysis using artificial neural networks. Stud Health Technol Inform 202:115–118
- Colak C, Karaman E, Turtay MG (2015) Application of knowledge discovery process on the prediction of stroke. Comput Methods Programs Biomed 119(3):181–185
- Maier O, Schro¨der C, Forkert ND, Martinetz T, Handels H (2015) Classifiers for ischemic stroke lesion segmentation: a comparison study. PLoS ONE 10(12):e0145118
- O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, ZhangH, Rao-Melacini P, Zhang X, Pais P, Agapay S, Lopez-Jaramillo P (2016) Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries(INTERSTROKE): a case-control study. Lancet 388(10046): 761–775
- Tsuruoka Y, Tateisi Y, Kim JD, Ohta T, McNaught J, AnaniadouS, Tsujii J (2005) Developing a robust part-of-speech tagger for biomedical text. In: Advances in informatics—10th Panhellenic conference on informatics, pp 382–392
- Kulick S, Bies A, Liberman M, Mandel M, McDonald R, PalmerM, Schein A, Ungar L (2004) Integrated annotation for biomedical information extraction. Linking biological literature, ontologies and databases. In: Proceedings of the HLT/NAACL 2004 workshop: BioLINK, pp 61–68
- Toutanova K, Klein D, Manning CD, Singer Y (2003) Featurerich part-of-speech tagging with a cyclic dependency network. In: Proceedings of NAACL '03, pp 173–180
- Tateisi Y, Tsujii J (2004) Part-of-speech annotation of biologyresearch abstracts. In: Proceedings of 4th international conference on language resource and evaluation (LREC2004), pp 1267–1270
- Pollay M (2012) Overview of the CSF dual outflow system. ActaNeurochir Suppl 113:47–50
- Fan J, Upadhye S, Worster A (2006) Understanding receiveroperating characteristic (ROC) curves. Can J Emergency Med 8(1):19–20
- Dreyfus SE (1990) Artificial neural networks, back propagation,and the Kelley-Bryson gradient procedure. J Guid Control Dyn 13(5):926–928
- Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn20(3):273–297