Neural Computing and Applications (2020) 32:817–828
https://doi.org/10.1007/s00521-019-04041-y

INTELLIGENT BIOMEDICAL DATA ANALYSIS AND PROCESSING

Classification of stroke disease using machine learning algorithms

Priya Govindarajan[1] · Ravichandran Kattur Soundarapandian[2] · Amir H. Gandomi[3] · Rizwan Patan[4] · Premaladha Jayaraman[2] · Ramachandran Manikandan[2]

Received: 14 November 2018 / Accepted: 14 January 2019 / Published online: 25 January 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract
This paper presents a prototype to classify stroke that combines text mining tools and machine learning algorithms. Machine learning can be portrayed as a significant tracker in areas like surveillance, medicine, data management with the aid of suitably trained machine learning algorithms. Data mining techniques applied in this work give an overall review about the tracking of information with respect to semantic as well as syntactic perspectives. The proposed idea is to mine patients' symptoms from the case sheets and train the system with the acquired data. In the data collection phase, the case sheets of 507 patients were collected from Sugam Multispecialty Hospital, Kumbakonam, Tamil Nadu, India. Next, the case sheets were mined using tagging and maximum entropy methodologies, and the proposed stemmer extracts the common and unique set of attributes to classify the strokes. Then, the processed data were fed into various machine learning algorithms such as artificial neural networks, support vector machine, boosting and bagging and random forests. Among these algorithms, artificial neural networks trained with a stochastic gradient descent algorithm outperformed the other algorithms with a higher classification accuracy of 95% and a smaller standard deviation of 14.69.

Keywords Stroke · Tagging · Maximum entropy · Data pre-processing · Classification · Machine learning

1 Introduction

Health is considered as an essential aspect of everyone's life, and there is a need for a recording system which tracks data on diseases and the relationship between them. Most of the information pertaining to diseases could be found in the case summaries of patients, medical records found in clinics and other records that are maintained manually. The sentences in them could be deciphered through various methodologies of text mining and machine learning (ML). Machine learning is a tool which can disseminate the content as a part of information retrieval in which semantic and syntactic parts of the content are given prevalence. Various ML and text mining methodologies are proposed and implemented for feature extraction and classification.

Stroke is a term used by most of the healthcare practitioners to describe injuries in the brain and spinal cord resulting from abnormalities in the supply of blood. Stroke projects its meaning based on different perspectives; however, globally, stroke evokes an explicit visceral response. A brain comprises 100 billion and a trillion neurons and glia, respectively, wrapped into more than three pounds of tissue, which contains every memory and

✉ Amir H. Gandomi
a.h.gandomi@stevens.edu

Priya Govindarajan
priya@src.sastra.edu

Ravichandran Kattur Soundarapandian
raviks@it.sastra.edu

Rizwan Patan
patan.rizwan@galgotiasuniversity.edu.in

Premaladha Jayaraman
premaladha@ict.sastra.edu

Ramachandran Manikandan
manikandan75@core.sastra.edu

[1] Department of Computer Science, SASTRA Deemed University, Kumbakonam, India

[2] Department of Information and Communication Technology, SASTRA Deemed University, Thanjavur, India

[3] School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA

[4] School of Computing Science and Engineering, Galgotias University, Greater Noida, India

encodes and stores them in a network. Brain activity supports each and every individual's breath and movement. The number of people who lose their life due to stroke is ten times greater in developing countries for more than the past five decades (i.e., from 1970), and it is projected to double globally by 2030. Generally, stroke is classified into the following three types: *ischemic stroke* (*IS*), *hemorrhagic stroke* (*HE*), and *transient ischemic attack* (*TIA*). Ischemic stroke is the most common type of stroke. The American Heart Association (AHA) has predicted that 87% of strokes are ischemic stroke [1], which occur if a clot or an obstacle persist in a blood vessel of the brain. Ischemic stroke has two categories: embolic stroke and thrombotic stroke [2]. Embolic stroke occurs if a block/clot forms in any part of the body and moves toward the brain and blocks blood flow. Thrombotic stroke is due to a clot that weakens blood flow in an artery, which carries blood to the brain. Hemorrhagic stroke occurs from a split/burst of weakened blood vessels. Only 10–15% of strokes are predicted to be a hemorrhagic stroke, but the rate of mortality is high when compared with ischemic stroke [3–5]. Hemorrhagic stroke is classified into two types: subarachnoid hemorrhage and intracerebral hemorrhage. Transient ischemic attack is described as a "mini-stroke," which is due to a clot. TIA is a temporary blockage relative to other types of stroke, and it lasts only for a short period of time (an average of 1 min), and symptoms disappear within 24 h. TIA does not cause permanent injury to the brain or its tissues [6]; however, TIA is taken as a warning for the occurrence of an additional stroke in the near future. Stroke, irrespective of the type, is mostly considered to be a fatal disease. This study focuses on developing methodologies to extract the base form of the text from the patients' case summaries or medical records, to retrieve the root word or stem from the base form through stemmers and to classify the type of stroke as ischemic and hemorrhagic stroke (based on their common symptoms) from the retrieved root words. Beyond all these detections, stroke patients are always in need of intensive care, which can be provided by an interdisciplinary team.

1.1 Related works

Few researchers are working on stroke prediction with machine learning (ML) algorithms. Significant research contributions are described in this section. A previous study used the artificial neural networks (ANN) method, trained with six different multilayer perceptron (MLP) algorithms to predict mortality of stroke patients which generated an accuracy of 80.7% [7]. Another study used support vector machine (SVM), *k*-nearest neighbor (kNN), and ANN to automate the detection of ischemic stroke, which suggested that SVM has higher prediction accuracy [8, 9]. Amini et al. [10] predicted stroke incidence by using *k*-nearest neighbor and C4.5 decision tree methods to reveal that C4.5 decision tree methods yielded a higher accuracy rate of 95.42%. Another group [11] used machine learning methods and SVM to predict stroke thrombolysis outcome, which showed that SVM was more accurate. Cheng et al. [12] predicted ischemic stroke using two ANN models that provided accuracy rates of 79.2% and 95.1%. One study [13] used the knowledge discovery process (ANN and SVM) to forecast the presence of stroke. The results of this study suggested that ANN had better prognostic performance than other models. Maier et al. [14] applied nine classification methods, including generalized linear models, random decision forests (RDFs), and convolutional neural networks (CNNs), to classify ischemic stroke and concluded that RDFs and CNNs provide better classification accuracy than other methods. Kansadub et al. [15] used decision trees (DTs), naive Bayes, and ANN to predict stroke and concluded that DT yielded better classification than other methods. Sung et al. [16] used kNN, multiple linear regression (MLR), and a regression tree model to predict the stroke severity index (SSI) and demonstrated that kNN has better accuracy than other models. Another study [17] presented a prediction model with DT, ANN, SVM, logistic regression (LR), and ensemble approach generalized boosted model (GBM) to predict ICU transfer of stroke patients and concluded that GBM provided the highest accuracy.

Classification of stroke through machine learning techniques is discussed in the research work by Adam et al. [18], and they have reviewed many works with the perspective of classification. Their work discussed two algorithmic approaches, decision tree and *k*-nearest neighbor (KNN). It concluded that decision tree performed better than KNN algorithm. A recent study states that the etiology of the stroke patients remains unclear, even though there are many diagnostic techniques available for ischemic stroke [19]. The study concluded the importance of the phenotypic form of classification of stroke; it also describes the lack of its reliability in performance and accuracy.

Chantamit-O-Pas et al. [20] propose stroke prediction through deep learning. The knowledge of medical domain problems could not be traced accurately by the traditional predictive models. The outcome of the study was more accurate than a scoring system in the medical domain in the prediction of stroke.

A survey was conducted in different perspectives by the Asian Stroke Advisory Panel in 12 different countries in 13 Asian regions. The report states that in Asian countries, a higher proportion of people were found to be more prone to ischemic stroke. The number of stroke patients in Asia ranges between 116 and 483/100,000 per year. They also

observed a three times rise in the count of neurologists in all the countries [21].

In this paper, 22 attributes, derived from real-time data collected from the Multispecialty Hospital (Table 1), were analyzed with text mining and machine learning algorithms to improve classification accuracy.

1.2 Motivation and objective of the study

The major pitfalls identified in the literature survey are: Most of the research works have its contribution only to ischemic stroke (IS) type; the impact of risk factors for stroke and its classification are not given due importance in the research; and most of the research works have classified stroke with the aid of 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 study, mining techniques are proposed to overcome the above-mentioned drawbacks and to classify the type of strokes precisely.

In the perspective of many neurologists, there is no medicine available till date to cure stroke completely. Rather we do have supportive palliative treatment which would probably prolong the lifespan of an individual. The number of people who lose their life due to stroke was ten times more in developing countries, and it tends to increase doubly throughout the Universe by 2030 [22]. Canadian Institutes of Health Research and Heart and Stroke foundation conducted a study in two phases that indicated the impact of the risk factor of a patient in the occurrence of stroke [23, 24]. As per the report given by Asian Stroke Advisory Panel, a higher proportion of people were found to be affected by ischemic stroke [21]. Hence, to bring down the level of symptoms in such diseases, a clear classification is needed. The study proposes methodologies to mine the data from case sheets and medical reports in order to classify this fatal disease. The outcome of the research work could possibly help the practitioners in the medical field to know the intensity level of the disease and to make decisions accordingly.

The primary objective of this research work is to acquire stroke dataset and classify the type of stroke by using mining and machine learning algorithms. To achieve the primary objective, tagging, stemming, and classification of the stroke are done. Based on these, the sub-objectives are formulated as follows:

1. To mine the relevant information from the raw data using tagging and maximum entropy methodologies.
2. To fetch the processed dataset using a novel stemming algorithm and avoid the discrepancies related to the size of the words and stemming errors found in the earlier studies.
3. To classify the type of stroke with reasonable accuracy by providing importance to the variation in the dataset.

The paper is structured as follows: Sect. 2 presents the proposed prototype, Sect. 3 illustrates the results and discussion, and Sect. 4 summarizes the results obtained from the proposed prototype.

2 Proposed prototype

The proposed prototype is comprised of three main stages: data acquisition, data pre-processing, and classification. The proposed workflow diagram is shown in Fig. 1. Data were acquired from the case sheets collected from the hospital using tools equipped with tagging and maximum entropy algorithms. The data acquired from the acquisition phase are then pre-processed using correlation analysis to

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		

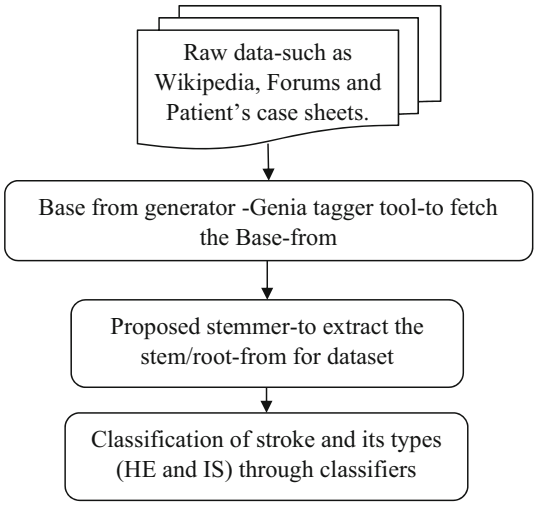


Fig. 1 Workflow diagram of the proposed prototype

remove redundancies, which is technically termed as data duplication or repetition of data. Then, the pre-processed data are fed to different machine learning algorithms for classification.

2.1 Data acquisition

The data were collected in the form of patient case sheets from Sugam Multispecialty Hospital, India. The case sheets contained information from over 507 stroke patients ranging from 35 to 90 years of age. A total of 22 unique class labels related to stroke were identified that fell under two major stroke types: ischemic stroke and hemorrhagic stroke. Risk factors such as hypertension and diabetes mellitus were also taken into account during classification. The case sheets were pre-processed with the base-form generator and novel stemmer algorithms to acquire the dataset for classification of the type of stroke.

Base-form generator Genia tagger [25] is used as the tool, which processes English sentences and provides the base forms, chunk tags, part-of-speech (POS) tag, and named entity (NE). This tool was developed for analyzing biomedical text such as the abstract of MEDLINE entries [26]. Toutanova et al. [27] developed a bidirectional dependency network which was used by Tsuruoka et al. [25], to define an algorithm for part-of-speech tagging called POS tagging algorithm. The POS part of the GENIA tagger 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):

$$p(t|c) = \frac{1}{Z(C)} \exp \sum \lambda_i f_i(C,t) \tag{1}$$

where $f_i(C,t)$ represents the features of content C used to find the tag t. $\sum$ is used for summing n items, which takes the value from $i = 1$ to n in which λ_i denotes the weight and $Z(C)$ represents the constants used for normalization; as a whole, it (1) is used as a classifier for text classification. Maximum entropy methodology makes assumptions only from the given details of the data. Tagging is carried out using a rule-based algorithm that correlates distinct terms and POS with its tags. Maximum entropy classifier is one of the machine learning methodologies that act as an aid for the problem of POS tagging, with an attainable accuracy of 95% [1].

Novel stemmer Keeping in perspective many of the affix removal stemmers, novel stemming algorithm is proposed with few rule sets, which covers larger counts of words to be stemmed. This algorithm provides the stem or root form by removing the suffixes and prefixes. A set of modified rules are given below:

- The suffix of any given word which ends with "ies" is replaced with "y"
- The suffix of any given word which ends with "er" is removed and is replaced with "null"
- The suffix of any given word which ends with "es" is replaced with "e"
- The suffix of any given word which ends with "ment," "ing," "ed" is removed, and it is replaced with "null."

Thus, with the help of the above rules 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.

2.2 Data pre-processing

Data pre-processing is vital to enhance the quality of the data for their usage. Accuracy, interoperability, and reliability are the factors that reflect the quality of the data. Data pre-processing involves many stages, such as data cleaning, data integration, data reduction, and data transformation. Data cleaning resolves many inconsistencies, even noisy data, and provides missing terms. Data integration means amalgamating data from various sources into a single unified data. Redundant data are one of the most significant issues found in the pre-processing step which is carried out in data integration whereas, data reduction reduces voluminous data. As stated earlier, one of the most significant problems in data integration is redundancy. Some of the reasons for redundancy are attribute naming, inconsistency, and whether the attribute is taken from another set of attributes. It can be detected from correlation analysis, which manipulates the Chi-square test for nominal data and the correlation coefficient for numerical data. To overcome data duplication, online tools such as Data Cleanser and Merge/PurgeLibrary (Sagent/QMSoftware), which have user-specific matching rules for integration, are used [29]. The transformation of data into a form appropriate for mining is data transformation. These are some of the stages involved in pre-processing the data.

The data are pre-processed to obtain the symptoms and factors of the patients as predefined parameters for classification and prediction of stroke and its type. Tables 1 and 2 represent the features extracted from the case sheets and data samples, respectively.

Pre-processed data of 507 samples with 22 features (excluding X1, the patient number) are given to the different machine learning algorithms, and the implementation details are given in the next section.

2.3 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 To represent classification accuracy, receiving operating characteristic (ROC) curves are used in this work. In dichotomic classification, instances are predicted using a continuous variable X that is "value" computed for each instance. Based on a threshold T, instances are classified if $X > T$, then it is positive; otherwise, it is negative. $f_1(x)$ is the probability density function that would be followed by X if the instance is positive. $f_0(x)$ is the probability density function that would be followed by X if the instance is negative. Hence, true positive rates and false positive rates are given by:

$$\text{True positive rate } (T) = \int_T^{\infty} f_1(x)\mathrm{d}x \tag{2}$$

$$\text{False positive rate } (T) = \int_T^{\infty} f_0(x)\mathrm{d}x \tag{3}$$

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

2.3.1 ANN

ANN are constructed with 22 inputs and one hidden layer with ten neurons and two outputs (IS and HE). The network is trained with the stochastic gradient descent algorithm. Stochastic gradient descent or incremental gradient descent is a stochastic approximation of the gradient descent optimization. It also helps in optimizing differentiable objective function through an iterative method. It finds maxima or minima by iteration [31]. The following is the pseudocode of stochastic gradient descent:

1. Initial vector of parameters w and learning rate η is chosen
2. Repeat the following steps until an approximate minimum is obtained:
   - Examples are randomly shuffled in the training set
   - For $i = 1$ to n, do

     $w = w - \eta \nabla Q_i(w)$

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

From the samples, 300 samples were used for training and the remaining 207 samples were used for testing. Implementing the neural networks achieved 95% classification accuracy with a standard deviation (calculated between the performance measures) of 14.69.

2.3.2 SVM

SVM is another classification method used for predicting strokes, which was developed from statistical learning theory and is widely used in many domains, from image recognition to bioinformatics. SVM training algorithm constructs a prototype that assigns new entities to the existing group or creates a new group.

SVM which is tagged as supervised machine learning is used for both regression challenges and classification. This algorithm plots each entity from the dataset in n-dimensional space (n—number of features) as a point. Each feature's value is considered as the value of specific coordinate. Classification is performed by discovering the hyperplane which distinguishes both the classes.

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

$$f(x) = \sum_{x_j \in S} \alpha_j y_j K(x_j, x) + b \tag{4}$$

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

The pre-processed samples are classified using various kernels such as linear SVM, quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, and coarse Gaussian SVM. Of these methods, linear SVM, medium Gaussian SVM, and coarse Gaussian SVM produced the highest accuracy of 91.5%. ROC curves obtained for the different SVM kernels achieved the higher accuracy with a training time of 2.28 s, which are depicted in Figs. 2, 3, and 4.

2.3.3 Decision tree

Decision tree constructs a tree structure using classification or regression models. A decision tree is developed by splitting the dataset into smaller subsets. Tree classifies the dataset, but it does not know to learn on itself [34, 35] through the example of the patient. Each and every dataset comes under any of the labeled class. Therefore, it falls under the perspective of supervised learning rather than unsupervised learning. The tree is built using the information gain, and prolonged improvements are projected through a single pruning methodology. It classifies all types of data, such as continuous, discrete, concise, and easy to infer, and the throughput is always a human-readable one.

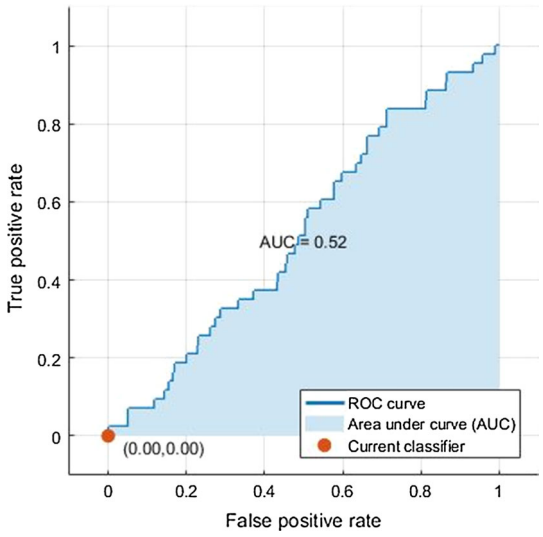


Fig. 2 Linear SVM

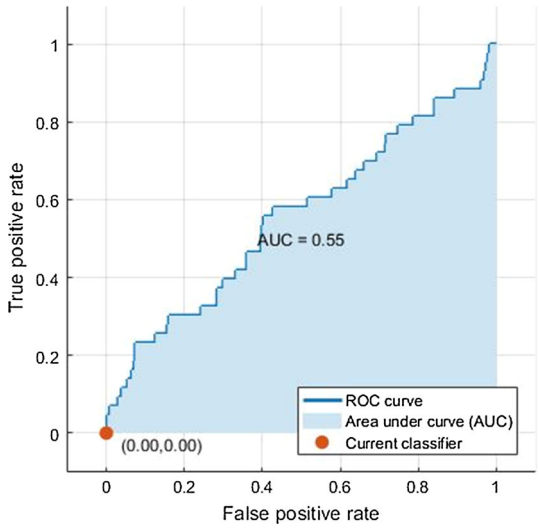
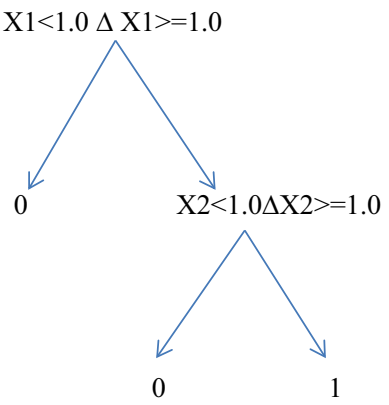


Fig. 3 Medium Gaussian SVM

Simple example for classification tree:



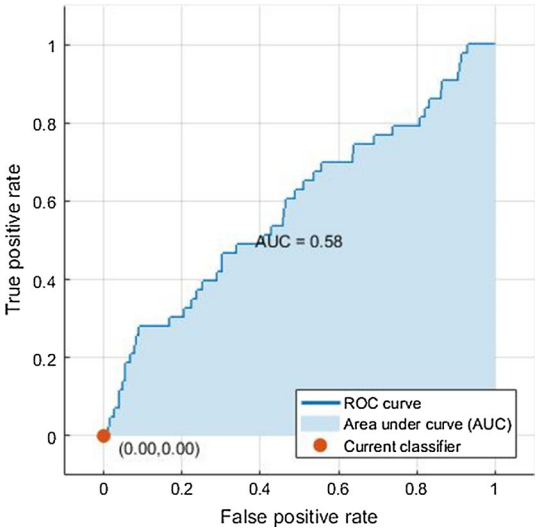


Fig. 4 Coarse Gaussian SVM

The prediction starts at the head node of the tree (Δ) and checks the decision with the first predictor (X1): If its value is less than 1.0, then it follows the left branch and the tree classifies its prediction as type 0; else it follows the right branch where again a prediction is made based on the second predictor X2. If X2 value is smaller than 1.0, then it follows the left branch and the tree classifies its prediction as type 0; else it follows the right branch and the tree classifies its prediction 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.

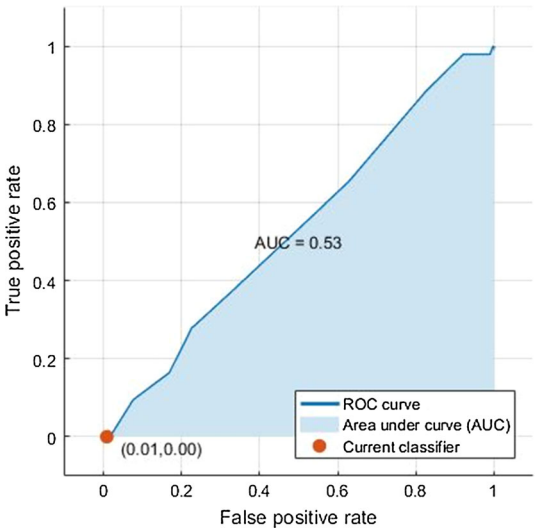


Fig. 5 Simple tree (20 splits)

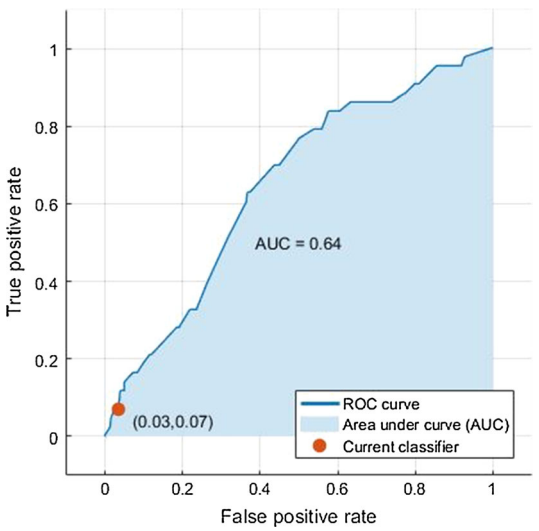


Fig. 6 Medium tree (60 splits)

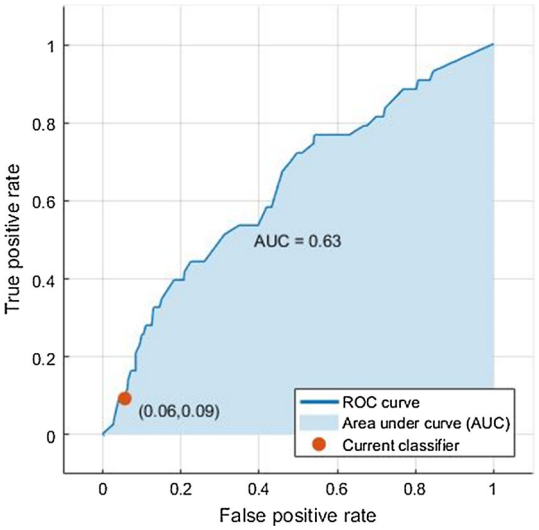


Fig. 7 Complex tree (100 splits)

2.3.4 Logistic regression

Logistic regression (LR) is based on predictive analysis, which describes the data and provides the relationship between independent variables and dependent (binary) variables. For example, does the patient's age, hypertension, and diabetes mellitus level impact the stroke patient (yes or no)? The outcome of the process is either 0 or 1, which is tagged as dependent, and other predictors are taken as covariates [36]. This methodology is used in various fields, including machine learning to predict the presence of disease (based on factors) and in the marketing field. In this work, the LR method produced an accuracy of 90.6% with a training time of 8.51 s; when the sample data were fed in, 10-fold cross-validation is done. ROC curve for this type is shown in Fig. 8.

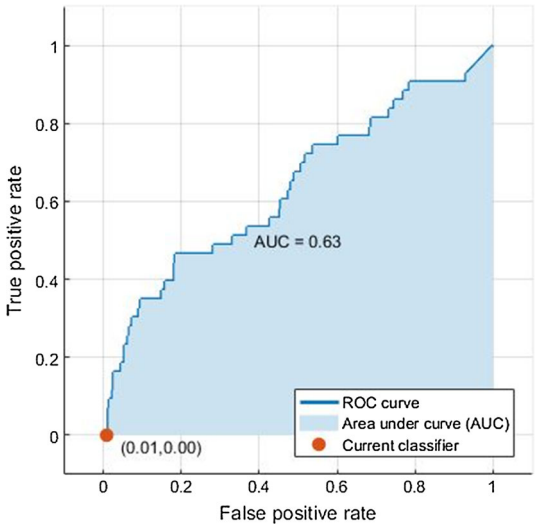


Fig. 8 Logistic regression

2.3.5 Bagging and boosting

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms to make predictions that are more accurate than any individual model. The bagging algorithm creates an ensemble of models (classifiers or predictors), which is a learning scheme where each model gives an equal-weighted prediction. The method used in the bagging classifier type is random forest.

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:

1. Random selection of "m" features from the total "n" features, where $m \ll n$
2. Among the "m" features, the node "x" calculation is done by means of best split point
3. Node is divided into children node using the best split
4. Repeat the steps 1–3 until "y" number of nodes are reached
5. 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):

1. Store the predicted outcome as the target by using the rules as well as test features of a randomly created decision tree
2. Predicted target's votes are calculated
3. 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):

$$F_T(x) = \sum_{t=1}^{T} f_t(x) \tag{5}$$

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:

1. An initial weight value, $w_i = 1/n$ (n, represents the number of total observations) is assigned to each observation, X_i
2. "Weak" model is trained (often a decision tree)
3. In each observation,
   - 3.1. w_i is increased, for incorrect prediction
   - 3.2. w_i is decreased, for correct prediction
4. A weak model is trained by giving more priority to higher weights in the observations.
5. Steps 3 and 4 are repeated until the observations predict perfectly or until a stipulated number of trees are trained.

Since an (simple) algorithm might classify the objects poorly, one combines many classifiers with the selected training set (in each iteration) and by assigning an appropriate weight (final voting). Through this method, one can achieve good accuracy overall [39].

Both algorithms such as AdaBoost and random forest provide an accuracy of 90.9% with a training time of 7.74 s and 91.5% with a training time of 7.24 s, respectively, which are shown in Figs. 9 and 10.

3 Results and discussion

The study was conducted with the dataset and parameters (patient symptoms) shown in Table 1. The novelty of the work is in the data processing phase, where the proposed algorithm called novel stemmer was used to attain the dataset. The collected data (507 patients) encompassed age of the patients, ranging from 35 to 90 years, with 22 unique class labels (parameters) that fall under either ischemic 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

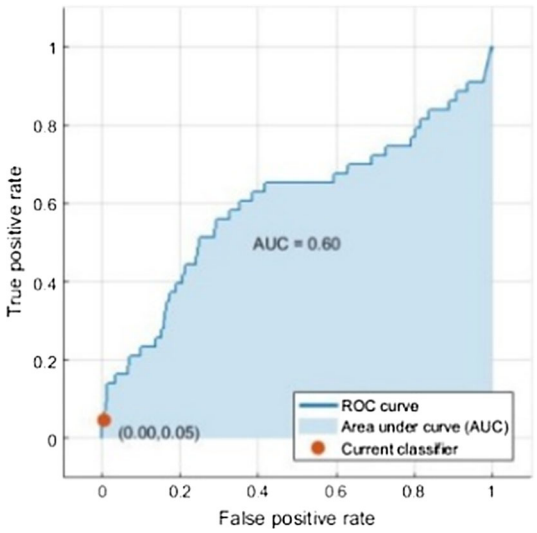


Fig. 9 Bagging

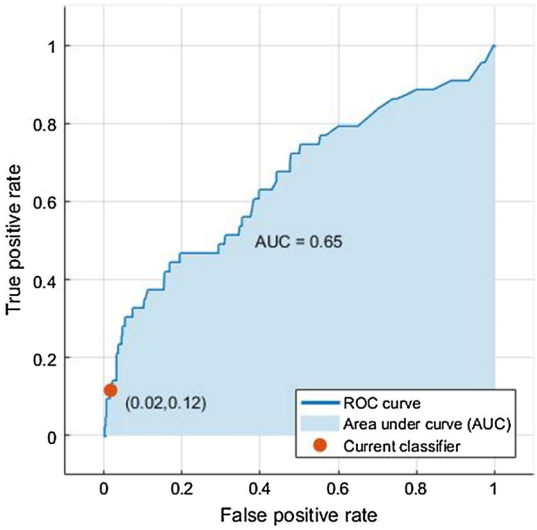


Fig. 10 Boosting

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).

Of the collected dataset, 90% was used for testing the trained data. The developed model produced a minimum predictive error. The classification was based on patient symptoms, along with factors such as age, gender, HT, and DM. The results of the classification methodology are shown in Table 4, which shows the classification evaluation metrics of accuracy, sensitivity, specificity, precision, and recall, which are shown in Table 3. The confusion matrix and associated results for the above classification methodologies are shown in Table 4. The standard deviation between the evaluation metrics was derived to evaluate whether the most accurate classifier with the smallest deviation was statistically significant. The accuracy for all types of classifiers plotted in Fig. 14, which shows 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 illustrates how symptoms along with risk factors play a vital role in not only classifying stroke but also in demonstrating how each individual symptom impacts the presence of either type of stroke (ischemic/hemorrhagic) in a patient, which adds to the novelty of the work. Classification is performed using the data acquired through patient's data sheets.

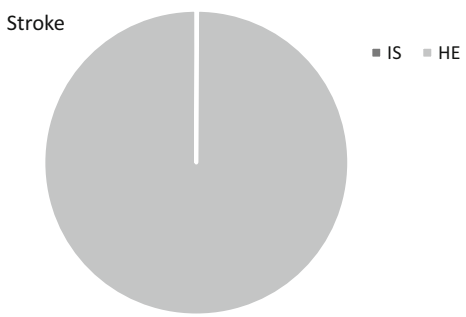


Fig. 11 Prevalence percentages of stroke and its type

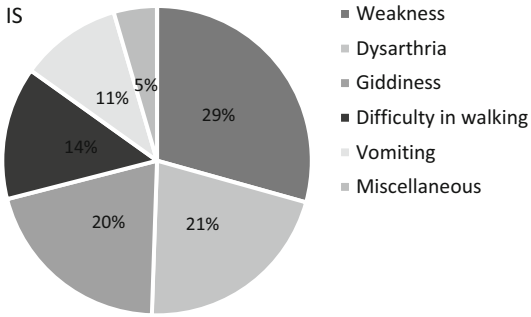


Fig. 12 Prevalence percentage of symptoms in IS patients

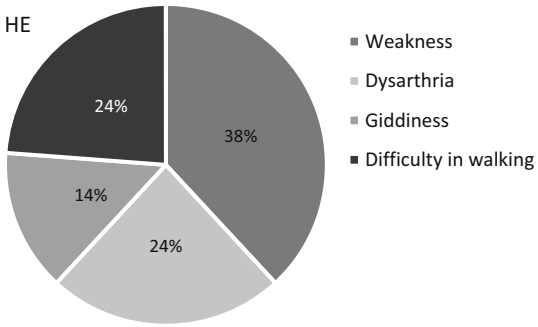


Fig. 13 Prevalence percentages of symptoms in HE patients

Table 3 Various evaluation metrics for classification

S. no.	Metric	Mathematical expression
1	Accuracy	$\frac{(TP+TN)}{\text{Total no. of samples}}$
2	Sensitivity	$\frac{TN}{\text{Actual no. of positive samples}}$
3	Specificity	$\frac{TN}{\text{Actual no. of negative samples}}$
4	Recall	$\frac{TP}{TP+FN}$
5	Precision	$\frac{TP}{TP+FP}$

Previous research showed that classification of ischemic stroke with kNN and decision tree algorithm had minimal prediction error. The study states that it is difficult to accurately classify the types of ischemic stroke, even with medical procedures. The outcome of the entire study indicates that decision tree performs better than kNN algorithm [10]. In most of the research works, a minimum of two to three algorithms were used to classify the stroke [8, 10, 12, 13, 15, 18, 19], whereas this paper demonstrates classification of stroke using various classifiers. Patient symptoms and risk factors were used for classification, which classifies the type of stroke. The ANN classifier provided more than 95% accuracy, zero negative predictive value, and a standard deviation less than 14.69 for classifying stroke type as either as ischemic or hemorrhagic, which is higher when compared with the previous research work [7, 12].

When compared with the previous work that used only two algorithms for classification, our work classified stroke type using more than five classifiers along with their categories. Accuracy and standard deviation of all classifiers were novel when compared to previous work, which classify only part of a stroke.

The classification of both the types of stroke is not given their due importance in the research work even though hemorrhagic stroke has the highest mortality rate than the other type of stroke. In this work, classification of both the types of stroke, with various classifiers with its kernel is illustrated which also adds to the novelty of the study. In short, most of the classification aids the medical specialist to classify 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.

Table 4 Performance measurements for various classification methods

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

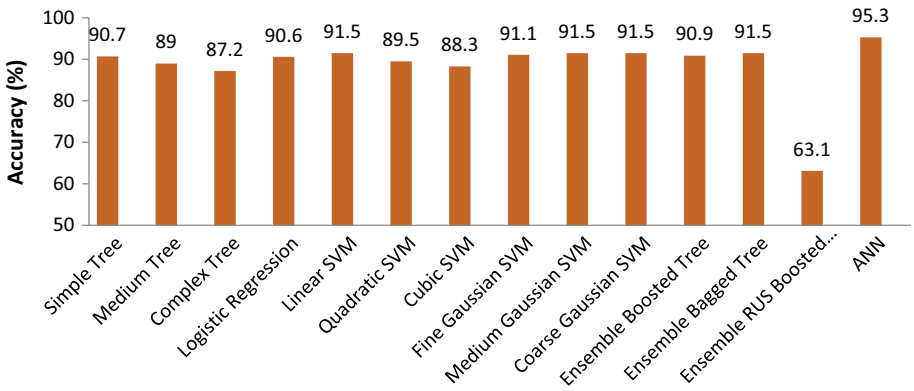


Fig. 14 Classification accuracy for various classification methods

4 Conclusion

The study brings out the effectiveness of the classification methods for structured entities like patient's case sheets to classify strokes based on defined parameters (symptoms) and factors. This study predicts the type of stroke for a patient based on classification methodologies. The categories of SVM and ensemble (bagged) provided 91% accuracy with 0.0000 negative predictive value, while ANN trained with the stochastic gradient descent algorithm outperformed other algorithms, with a higher classification accuracy $> 95\%$ with a lower standard deviation of 14.69. This study indicates that stroke is more prevalent in men than in women and in the age group from 40 to 60 years old. Patients who suffered ischemic stroke were greater 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, Neurologist, Sugam Multispecialty Hospital, for permitting us to access the real-time data of the patients and for his valuable suggestions in classifying the type of strokes. We also thank the management of Sugam Multispecialty Hospital, Kumbakonam, for their assistance in collecting the case sheets. We acknowledge the Department of Science and Technology, India, for providing financial support through INSPIRE fellowship (No. IF120649) to carry out this research work. The second author also thanks Department of Science & Technology for financial aid from grant 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

1. Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, Borden WB, Bravata DM, Dai S, Ford ES, Fox CS, Fullerton HJ, Gillespie C, Hailpern SM, Heit JA, Howard VJ, Kissela BM, Kittner SJ, Lackland DT, Lichtman JH, Lisabeth LD, Makuc DM, Marcus GM, Marelli A, Matchar DB, Moy CS, Mozaffarian D, Mussolino ME, Nichol G, Paynter NP, Soliman EZ, Sorlie PD, Sotoodehnia N, Turan TN, Virani SS, Wong ND, Woo D, Turner MB (2012) Executive summary: heart disease and stroke statistics—2012 update: a report. Circulation 125(1):188–197
2. Pahus SH, Hansen AT, Hvas AM (2016) Thrombophilia testing in young patients with Ischemic stroke. Thromb Res 137:108–112
3. Dupont SA, Wijdicks EF, Lanzino G, Rabinstein AA (2010) Aneurysmal subarachnoid hemorrhage: an overview for the practicing neurologist. Semin Neurol 30(5):45–54
4. 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
5. Rebouças ES, Marques RCP, Braga AM, Oliveira SAF, de Albuquerque 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
6. 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
7. Süt N, Çelik Y (2012) Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turk J Med Sci 42(5):886–893
8. Rajini NH, Bhavani R (2013) Computer aided detection of ischemic stroke using segmentation and texture features. Measurement 46(6):1865–1874
9. Sundström C (2014) Machine learning algorithms for stroke diagnostics. Master's thesis in biomedical engineering
10. 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
11. Bentley P, Ganesalingam J, Jones AL, Mahady K, Epton S, Rinne P, Sharma P, Halse O, Mehta A, Rueckert D (2014) Prediction of stroke thrombolysis outcome using CT brain machine learning. NeuroImage Clin 4:635–640
12. Cheng CA, Lin YC, Chiu HW (2014) Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. Stud Health Technol Inform 202:115–118
13. 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
14. Maier O, Schröder C, Forkert ND, Martinetz T, Handels H (2015) Classifiers for ischemic stroke lesion segmentation: a comparison study. PLoS ONE 10(12):e0145118
15. Kansadub T, Thammaboosadee S, Kiattisin S, Jalayondeja C (2015) Stroke risk prediction model based on demographic data.

In: Biomedical engineering international conference (BMEiCON), pp 1–3
16. Sung SF, Hsieh CY, Yang YH, Lin HJ, Chen CH, Chen YW, Hu YH (2015) Developing a stroke severity index based on administrative data was feasible using data mining techniques. J Clin Epidemiol 68(11):1292–1300
17. Alotaibi NN, Sasi S (2016) Stroke in-patients' transfer to the ICU using ensemble based model. In: IEEE international conference on electrical, electronics, and optimization techniques (ICEEOT), pp 2004–2010
18. Adam SY, Yousif A, Bashir MB (2016) Classification of ischemic stroke using machine learning algorithms. Int J Comput Appl 149(10):26–31
19. Radu RA, Terecoasă EO, Băjenaru OA, Tiu C (2017) Etiologic classification of ischemic stroke: where do we stand. Clin Neurol Neurosurg 159:93–106
20. Chantamit-O-Pas P, Goyal M (2017) Prediction of stroke using deep learning model. In: Liu D., Xie S, Li Y, Zhao D, El-Alfy ES (eds) Neural information processing ICONIP, Lecture notes in computer science 10638
21. Suwanwela NC, Poungvarin N, The Asian Stroke Advisory Panel (2016) Stroke burden and stroke care system in Asia. Neurol India 64:46–51
22. World Health Organization (2004) Global burden of disease (GBD) 2002 estimates. World health report 2004. WHO, Geneva
23. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, Rangarajan S, Islam S, Pais P, McQueen MJ, Mondo C, Damasceno A, Lopez-Jaramillo P, Hankey GJ, Dans AL, Yusoff K, Truelsen T, Diener H-C, Sacco RL, Ryglewicz D, Czlonkowska A, Weimar C, Wang X, Yusuf S (2010) Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet 376:112–123
24. O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, 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
25. Tsuruoka Y, Tateisi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, 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
26. Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, 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
27. Toutanova K, Klein D, Manning CD, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of NAACL '03, pp 173–180
28. Tateisi Y, Tsujii J (2004) Part-of-speech annotation of biology research abstracts. In: Proceedings of 4th international conference on language resource and evaluation (LREC2004), pp 1267–1270
29. Pollay M (2012) Overview of the CSF dual outflow system. Acta Neurochir Suppl 113:47–50
30. Fan J, Upadhye S, Worster A (2006) Understanding receiver operating characteristic (ROC) curves. Can J Emergency Med 8(1):19–20
31. Dreyfus SE (1990) Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. J Guid Control Dyn 13(5):926–928
32. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
33. Vishwanathan SVM, Murty MN (2002) SSVM: a simple SVM algorithm. In: Proceedings of the 2002 international joint conference on neural networks. IJCNN'02, vol 3, pp 2393–2398
34. Utgoff PE (1989) Incremental induction of decision trees. Mach Learn 4(2):161–186
35. Saraee M, Keane J (2007) Using T3, an improved decision tree classifier, for mining stroke-related medical data. Methods Inf Med 46(5):523–529
36. Liu L, Luo G, Ke Q, Zhang X (2017) An algorithm based on logistic regression with data fusion in wireless sensor network. Eurasip J Wirel Commun Netw. https://doi.org/10.1186/s13638-016-0793-z
37. Ho TK (1995) Random decision forests. In: Proceedings of the third international conference on document analysis and recognition, vol 1, pp 278–282
38. Isaac E, Easwarakumar KS, Issac J (2017) Urban landcover classification from multispectral image data using optimized AdaBoosted random forests. Remote Sens Lett 8(4):350–359
39. Freund Y, Schapire R, Abe N (1999) A short introduction to boosting. J Jpn Soc Artif Intell 14(771–780):1612
40. Filho PPR, Rebouças ES, Marinho LB, Sarmento RM, Tavares JMRS, Albuquerque VHC (2017) Analysis of human tissue densities: a new approach to extract features from medical images. Pattern Recognit Lett 2017(94):211–218. https://doi.org/10.1016/j.patrec.2017.02.005

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