Detection of Stroke Disease using Machine Learning Algorithms

Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat, Sharmin Akter Tanna,Taslima Mostafa Alif, Romana Rahman Ema.

Department of Computer Science and Engineering
North Western University, Khulna, Bangladesh
Email:{shoilytasfia,tajulkuet09,sumaiyajannat707,tannasharmin,mostafaalif14,romanacsejstu}@gmail.com

Abstract—A stroke is a medical condition in which poor blood flow to the brain results in cell death. It is now a day a leading cause of death all over the world. Several risk factors believe to be related to the cause of stroke has been found by inspecting the affected individuals. Using these risk factors, a number of works have been carried out for predicting and classifying stroke diseases. Most of the models are based on data mining and machine learning algorithms. In this work, we have used four machine learning algorithms to detect the type of stroke that can possibly occur or occurred form a person's physical state and medical report data. We have collected a good number of entries from the hospitals and use them to solve our problem. The classification result shows that the result is satisfactory and can be used in real time medical report. We believe that machine learning algorithms can help better understanding of diseases and can be a good healthcare companion. *Index Terms*—Stroke, machine learning, WEKA, Naive Bayes, J48, k-NN, Random Forest.

I. Introduction

A stroke occurs due to poor blood flow to the brain which results in cell death. Two main types of stroke are ischemic stroke and hemorrhagic stroke. Ischemic stroke occurs due to lack of blood flow and hemorrhagic stroke occurs due to bleeding [1]. Another type of stroke is transient ischemic attack. Ischemic stroke has two categories- embolic stroke and thrombotic stroke. An embolic stroke occurs by forming a clot in any part of the body and moves toward the brain and blocks blood flow. A thrombotic stroke caused by a clot that weakens blood flow in an artery. Hemorrhagic stroke is classified into two types- subarachnoid hemorrhage and intracerebral hemorrhage. Transient ischemic attack is also known as "mini-stroke" [2].

A large number of people lose their life due to stroke and it is increasing in developing countries [3]. There are several stroke risk factors that regulate different types of stroke. Predictive algorithms help to understand the relation between these risk factors to types of strokes. The machine learning algorithm can improve patients' health through early detection and treatment. We have used several machine learning algorithms to detect the type of stroke that can occur in a patient or already occurred from their clinical report and statistical data. We have built a stroke dataset by collecting data from various sources validated by medical experts. Then the dataset was processed to be used with the machine learning algorithms. We have built several models of classification. The result of the models is satisfactory and can be used in a realtime patient's stroke classification.

Our main contribution in this paper is as follows:

- We have collected a entries and prepared a dataset
- Cleaned and prepared the data for using in WEKA
- We have trained four different models (eg. Naive Bayes, J48, k-NN, Random Forest)
- We have tested and calculated statistical accuracy measures to validate the models.

The organization of this paper is as follows: Section II highlights an overview of related work. Problem statement and the methodology is described in section III and section IV. Section V delineate the experimentation and results. Finally, Section VI concludes this paper with discussions regarding possible future research on this topic.

II. Related Works

In recent years, there were published different works based on Machine Learning algorithms. Some of them are discussed in here:

Govindarajan et al. used Artificial Neural Networks (ANN), Support Vector Machine (SVM), Decision Tree, Logistic Regression, and ensemble methods (Bagging and Boosting) to classify the stroke disease [2]. They have collected the data from Sugam Multispeciality Hospital, India which contains information about 507 stroke patients ranging from 35 to 90 years of age. The novelty of their work is in the data processing phase, where an algorithm called novel stemmer was used to attain the dataset. In their collected dataset, 91.52% of patients were affected by ischemic stroke and only 8.48% of patients were affected by hemorrhagic stroke. Among the mentioned algorithms, Artificial Neural Networks with stochastic gradient descent learning algorithm have the highest accuracy with 95.3% for classifying stroke.

Jeena and Kumar proposed a model based on Support Vector Machine for stroke prediction [4]. they have collected data from International Stroke trial Database [5]. The dataset contains 12 risks factors (attributes). They have used 350 samples for their work. For training purpose 300 samples and for testing 50 samples were used. Different kernel functions like polynomial, quadratic, radial basis function and linear functions were applied. The highest accuracy of 91% was found with the linear kernel which gives the balance measure F1-score F-measure 91.7.

Singh and Choudhary developed a model with Artificial Neural Network (ANN) for stroke prediction [6]. They have collected datasets from the Cardiovascular Health Study (CHS) database. Three datasets were constructed which contains 212 strokes (all three) and 52, 69, 79 nonstroke respectively. The final dataset contains 357 attributes and 1824 entities with 212 occurrences of stroke. During feature selection, the C4.5 decision tree algorithm was used and Principle Component Analysis (PCA) for dimension reduction. In ANN implementation they have used Back Propagation learning method. They have got the accuracy as 95%, 95.2% and 97.7% for the three datasets respectively.

Adam et al. have been developed a classification model for ischemic stroke using decision tree algorithm and knearest neighbor (k-NN)[7]. Their dataset was collected from several hospitals and medical centers in Sudan which is the first dataset for ischemic disease in Sudan. It contains 15 features and information about 400 patients. The results of the experiment show that the performance of decision tree classification is higher than the performance of k-NN algorithm.

Sudha et al. used the Decision Tree, Bayesian Classifier, and Neural Network for stroke classification [8]. Their dataset contains 1000 records. PCA algorithm was used for dimensionality reduction. In ten rounds of each algorithm, they have got the highest accuracy as 92%, 91%, and 94% in Neural Network, Naive Bayes classifier, and Decision tree algorithm respectively.

Some of the methods like [4] and [7] use a very small dataset. Govindarajan et al. [2] have predicted only two classes of stroke. Therefore we have proposed a method which uses a large dataset with four classes of stroke.

III. Problem Formulation

A. Data Sources

We have constructed a dataset by collecting stroke data from various sources. Our dataset contains total 1058 individual patient's information of which 412 and are male and 646 are female patient. The type of strokes as: 437 are from ischemic stroke class, 302 from hemorrhagic stroke, 142 of mini-stroke, and 177 reports as brain stem stroke class. Although the dataset is not perfectly symmetrically distributed over all the classes, it has a good ratio to the others. Practically it is very difficult to gather symmetric dataset. The dataset contains total of 28 features. They are summarized in Table I.

B. Classifiers

We used Naive Bayes, J48, k-NN, and Random Forest classifier in WEKA toolkit.

Naive Bayes classifier are a collection of classification algorithms based on Bayes' Theorem. The Bayes theorem finds the conditional probability of an event occurring given the probability of another event that has already occurred. Mathematically it can be stated as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Advantage of using Bayesian classifier includes rapidity of use and it is very simple for handling the dataset containing many attributes. Moreover, it is good for small datasets (i.e. needs less training data), highly scalable, able to handle continuous and discrete data and irrelevant feature insensitive.

J48 The popular decision tree algorithm C4.5 is implemented in WEKA as a classifier and named J48. The features of J48 includes- accounting for missing values, pruning the decision trees, continuous attribute value ranges, classify discrete and continues data using thresholds, derivation of rules, etc [9].

k-Nearest Neighbors (k-NN) is a non-parametric method used for classification and regression. It is among the simplest of all machine learning algorithms. It is an instance-based learning, where the function is approximated locally and all computation is anticipated until classification. Predictions are made for a new instance by searching through the entire training set for the *k* most similar instances called the neighbors. The prediction output depends on those *k* instances. Majority vote is usually used for choosing the class. Different distance metrics can be used with k-NN like, Euclidean distance, Manhattan/Cityblock distance, Minkowski distance, etc. We have used Euclidean distance for our model. In k-NN, no training is necessary. Rather it just takes a test instance and calculates the distance between the test instance and all the dataset instances. Fig. 1 shows the classification idea in k-NN. The red point will be classified in class B for k=3, but for k=6 it will be classified into class A. Therefore, the classification result depends on choosing *k* and that's the difficult task in k-NN.

Random Forest is an ensemble learning method for classification and regression. Each classifier in the ensemble is a decision tree classifier (i.e. ID3, C4.5, CART, etc.) so that the collection of classifiers is a forest. It constructs decision trees at training phase and outputs the class that is the mode of the classes or mean prediction of the individual trees. Several works have been carried out to predict the life-threatening diseases using decision tree and proven to be more efficient. The idea behind the random forest is shown in Fig. VII.

IV. Methodology

A. Data pre-processing

Our database contains string values which cannot be processed by WEKA. Therefore we had to integer encoding for string values. For example, we have replaced the string "Male"

TABLE I: List of attributes of the dataset.

Sl.	Attributes	Description
1	Age	Age of the patient
2	Sex	Sex of the patient
3	Confusion	Health confusion
4	Vision Loss	Decreasing ability to see
5	Dizziness	A range of sensations, such as feeling faint, woozy, weak or unsteady
6	Headache	Symptom of pain anywhere in the region of the head or neck
7	Weaknessnausea	Feeling queasy or queasy in the stomach
8	Nausea	A sensation of unease and urge to vomit
9	Vomiting	Vomiting is the involuntary emptying of stomach contents through the mouth
10	Seizures	A seizure is a sudden, uncontrolled anxiety in the brain.
11	Loss of Balance	Loss the balancing sensation
12	Irregular Heartbeat	A situation when the heart beats too fast, slow, or irregularly.
13	Chest Discomfort	Feeling pressure or squeezing in the chest.
14	Fainting	Fainting is loss of consciousness caused by decreased blood flow to the brain
15	Fatigue	Fatigue is a feeling of constant tiredness or weakness
16	Difficulty Breathing	Feeling difficulty in breathing
17	Difficulty Speaking	Feeling difficulty in speaking
18	Hearing Loss	Reducing ability to hear.
19	Paralysis	Paralysis is the loss of function of muscle in any part of the body
20	Sensation Loss	Being unable to feel pain, heat, or cold
21	CT	Computed Tomography result of the patient
22	CTA	Computed Tomography Angiography result of the patient
23	MRI	Magnetic Resonance Imaging result of the patient
24	CTP	Computer-To-Plate result of the patient
25	MRA	Magnetic Resonance Angiogram result of the patient
26	X-RAY	X-RAY result of the patient
27	ECG	Electrocardiogram result of the patient
28	ECO	Echocardiogram result of the patient

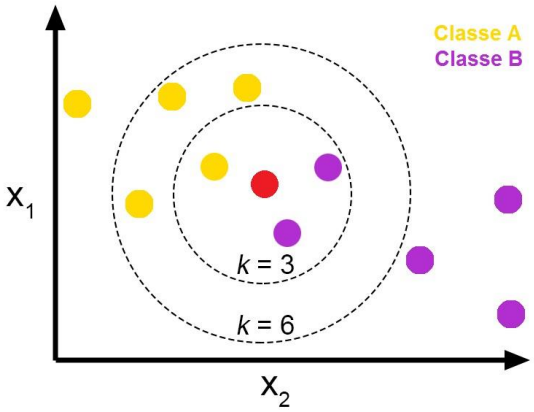

Fig. 1: k-NN.

with 0 and ”Female” with 1 and so on. Some attributes are missing in the dataset. Some of the attributes do not apply for the individuals i.e. N/A. We replaced them with zero ”0” for avoiding the null value exception. We also removed unnecessary information like ”3 times” used with the frequency of vomiting replaced by only 3 etc. Data preprocessing example is shown in Table II.

B. Data mining process

Waikato Environment for Knowledge Analysis (WEKA) is a machine learning toolkit, developed and maintained by the University of Waikato, New Zealand [10]. Previous studies show that WEKA is a very reliable suite for machine learning. A large number of similar works has been carried out using

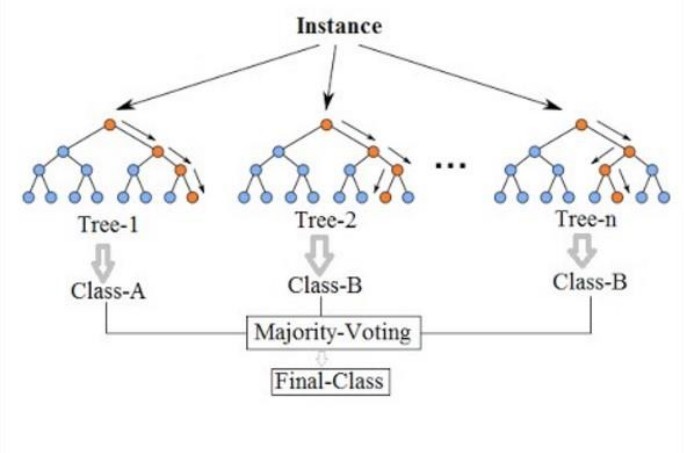

Fig. 2: Random Forest.

weka and they have found it advantageous [11] [12] [13] [14]. We have used the built-in algorithms in WEKA for stroke disease detection like Naive Bayes, Random Forest, and J48. These algorithms are described previously. First, we import the data from the stroke database. After pre-processing and integer encoding we apply WEKA to classify the strokes. The following steps have been performed for stroke detection in WEKA:

- Data pre-processing and visualization
- Attribute selection
- Test set and train set splitting
- Classification using different algorithms
- Model evaluation

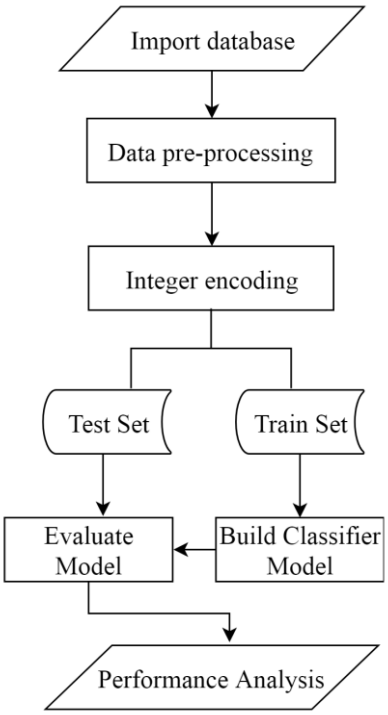

Fig. 3: Work-flow of data mining .

The work-flow of data mining is given in Fig. 3.

V. Experimental Results

To evaluate the performance, we have used Accuracy, Precision, Recall, and F1-score. Classification accuracy is the ratio of correct predictions to total number of predictions made by the model. Precision is the ratio of true positive to the true positive and false positive prediction. Recall is defined as the ratio of true positives to the true positive and false negative. F1-score or F-measure is the balance measure to express the performance in a single quantity. It is the harmonic mean of precision and recall They are formulated as follows:

$$Accuracy = \frac{TP + TN}{P + N} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

Where, TP: correct positive prediction, FP: incorrect positive prediction, TN: correct negative prediction, FN: incorrect negative prediction, P: TP+FP, N: TN+FN. The confusion matrix for calculating TP, FP, TN, FN is given in Fig. 4.

TABLE II: Data pre-processing.

Attributes	Before processing	After processing
Age	30	30
Sex	Male	0
Confusion	POSITIVE	1
Vision Loss	central vision loss	4
Dizziness	POSITIVE	1
Headache	POSITIVE	1
Weaknessnausea	POSITIVE	1
Nausea	NEGATIVE	0
Vomiting	3 times	3
Seizures	NEGATIVE	0
Loss of Balance	POSITIVE	1
Irregular Heartbeat	NEGATIVE	0
Chest Discomfort	NEGATIVE	0
Fainting	NEGATIVE	0
Fatigue	POSITIVE	1
Difficulty Breathing	NEGATIVE	0
Difficulty Speaking	N/A	0
Hearing Loss	N/A	0
Paralysis	N/A	0
Sensation Loss	N/A	0
CT	POSITIVE	1
CTA	POSITIVE	1
MRI	POSITIVE	1

CTP	N/A	0
MRA	N/A	0
X-RAY	deformaties in the skull	2
ECG	N/A	0
ECO	N/A	0

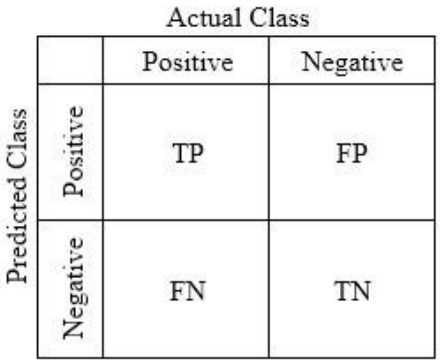

Fig. 4: Confusion matrix.

We have used a 10-fold cross validation for each algorithm. Performance comparison of different algorithms are shown in Table III.

From Table III, we see that the accuracy of Naive Bayes classifier is 85.6%. The accuracy for J48, k-NN and Random Forest is 99.8%. Naive Bayes has got the precision, recall, and f-measure as 88.1%, 85.6%, 86.1%. All of the J48, k-NN and

TABLE III: Performance comparison of different algorithms.

Algorithm	Accuracy	Precision	Recall	F-Measure
Naive Bayes	0.856	0.881	0.856	0.861
J48	0.998	0.998	0.998	0.998
k-NN	0.998	0.998	0.998	0.998
Random Forest	0.998	0.998	0.998	0.998

Random Forest has the precision, recall, and f-measure same as 99.8%, 99.8%, and 99.8% respectively.. Detailed results for each class on every algorithm are shown in Table IV, V, VI, and VII.

TABLE IV: Detailed performance of Naive Bayes algorithm.

Class	Accuracy	Precision	Recall	F-Measure
ISCHEMIC STROKE	0.872	0.995	0.872	0.929
HEMORRHAGIC STROKE	0.801	0.913	0.801	0.854
MINI STROKE	0.803	0.74	0.803	0.77
BRAIN STEAM STROKE	0.955	0.66	0.955	0.781

Table IV shows that the Brain Stem stroke class gets a better classification result for the Naive Bayes classifier in terms of accuracy. In terms of F-measure, it is Ischemic stroke class.

TABLE V: Detailed performance of J48 algorithm.

Class	Accuracy	Precision	Recall	F-Measure
ISCHEMIC STROKE	1	1	1	1
HEMORRHAGIC STROKE	1	0.993	1	0.997
MINI STROKE	0.993	1	0.993	0.996
BRAIN STEAM STROKE	0.993	1	0.994	0.997

From Table V we see that the Ischemic stroke class has the absolute classification result in terms of accuracy, precision, recall, and F-measure. Also, the other classes have higher classification results in J48 than Naive Bayes classifier.

TABLE VI: Detailed performance of k-NN algorithm.

Class	Accuracy	Precision	Recall	F-Measure
ISCHEMIC STROKE	0.998	1	0.998	0.999
HEMORRHAGIC STROKE	0.997	1	0.997	0.998
MINI STROKE	1	0.993	1	0.996
BRAIN STEAM STROKE	1	0.994	1	0.997

Table VI and VII also report that k-NN (with Euclidean distance) and Random Forest classifiers have the highest level of classification results achieved so far in our models.

Confusion matrix for each individual algorithms is shown in Table VIII, IX, X, and XI. The classes are:

TABLE VII: Detailed performance of Random Forest algorithm.

Class	Accuracy	Precision	Recall	F-Measure
ISCHEMIC STROKE	0.998	1	0.998	0.999
HEMORRHAGIC STROKE	0.997	1	0.997	0.999
MINI STROKE	1	0.993	1	0.996
BRAIN STEAM STROKE	1	0.994	1	0.997

- a = ISCHEMIC STROKE
- b = HEMORRHAGIC STROKE
- c = MINI STROKE
- d = BRAIN STEAM STROKE

TABLE VIII: Confusion matrix for Naive Bayes algorithm.

	a	b	c	d
a	381	10	29	17
b	2	242	4	54
c	0	12	114	16
d	0	1	7	169

TABLE IX: Confusion matrix for J48 algorithm.

	a	b	c	d
a	437	0	0	0
b	0	302	0	0
c	0	1	141	0
d	0	1	0	176

TABLE X: Confusion matrix for k-NN algorithm.

	a	b	c	d
a	436	0	0	1
b	0	301	1	0
c	0	0	142	0
d	0	0	0	177

Naive Bayes is extremely simple classifier and we should not expect it to be strong more than that. From the inspection of the classification result, we can say that the J48, k-NN, and Random Forest have done their obligation successfully in detection of stroke diseases.

VI. CONCLUSION

In this paper, a sufficiently large dataset of stroke attacked patients has been classified accurately. Four classifiers such as

TABLE XI: Confusion matrix for Random Forest algorithm.

	a	b	c	d
a	436	0	0	1
b	0	301	1	0
c	0	0	142	0
d	0	0	0	177

Naive Bayes, J48, k-NN, and Random Forest were used for detection of stroke disease. From the performance analysis we see that Naive Bayes performs better than other methods. The novelty and the main contribution of our work are collecting this dataset and preparing them to use with WEKA. The model can help people with a cautionary indication of being affected by stroke. Healthcare industries generate huge amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. Which is very difficult to relate to one another even by a field expert. It will help the clinician to better understand the type of disease. The limitations of our method are that the dataset is not perfectly symmetrical. However, it did not affect the predicted accuracy for the other algorithms. Naive Bayes algorithm didn't work as we expected.

In future work, it is possible to extend the research by using different classification techniques. Moreover, the prediction of stroke can be done by adding some non-stroke data with the existing dataset.

REFERENCES

[1] S. H. Pahus, A. T. Hansen, and A.-M. Hvas, "Thrombophilia testing in young patients with ischemic stroke," *Thrombosis research*, vol. 137, pp. 108–112, 2016.

[2] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," *Neural Computing and Applications*, pp. 1–12.

[3] L. T. Kohn, J. Corrigan, M. S. Donaldson, *et al.*, *To err is human: building a safer health system*, vol. 6. National academy press Washington, DC, 2000.

[4] R. Jeena and S. Kumar, "Stroke prediction using svm," in *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 600–602, IEEE, 2016.

[5] P. A. Sandercock, M. Niewada, and A. Członkowska, "The international stroke trial database," *Trials*, vol. 13, no. 1, pp. 1–1, 2012.

[6] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 158–161, IEEE, 2017.

[7] S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," *Int J Comput Appl*, vol. 149, no. 10, pp. 26–31, 2016.

[8] A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," *International Journal of Computer Applications*, vol. 43, no. 14, pp. 26–31, 2012.

[9] G. Kaur and A. Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.

[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[11] P. Sewaiwar and K. K. Verma, "Comparative study of various decision tree classification algorithm using weka," *International Journal of Emerging Research in Management &Technology*, vol. 4, pp. 2278–9359, 2015.

[12] K. A. Shakil, S. Anis, and M. Alam, "Dengue disease prediction using weka data mining tool," *arXiv preprint arXiv:1502.05167*, 2015.

[13] J. A. Alkrimi, H. A. Jalab, L. E. George, A. R. Ahmad, A. Suliman, and K. Al-Jashamy, "Comparative study using weka for red blood cells classification," *International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering*, vol. 9, no. 1, pp. 19–22, 2015.

[14] M. S. Siddiqui and A. I. Abidi, "Comparative study of different classification techniques using weka tool," *Global Sci-Tech*, vol. 10, no. 4, pp. 200–208, 2018.