# RISK PREDICTION OF STROKE USING DATA MINING CLASSIFICATION TECHNIQUES

1st Rahatara Ferdousi[a], 2nd M M Faniqul Islam[b,] 3rd Masum Mohammad Jubayel[c], 4th Asima Akter Chowdhury[d]

[a] *Metropolitan University, Sylhet 3100 Bangladesh.*

*rahatara@metrouni.edu.bd*

[b] Queen Mary University of London, Bangladesh

*m.islam@smd17.qmul.ac.uk*

[c] *Metropolitan University, Sylhet 3100 Bangladesh.*

*masumjubayel@gmail.com*

[d] *Metropolitan University, Sylhet 3100 Bangladesh.*

*chowdhuryasima@gmail.com*

## Abstract

Stroke, a fatal non-communicable disease of any age, kills more people than AIDS, Tuberculosis and Malaria put together in each year. WHO estimated around 6.2 million deaths because of stroke in 2008. As the incidence, prevalence, mortality, and disability rates are increasing, overall stroke burden has increased globally. Almost 70% of patients are unaware of their mild stroke, 30% seek medical attention lately and another 30% suffer from recurrent stroke, before seeking attention. Data mining, with its several techniques for classification and regression, plays a leading role in developing an effective model of risk prediction in the context of healthcare. Even though stroke prevention is a complex medical issue, primary prevention could be feasible by using data mining classification techniques that will assess risk factors to predict the likelihood of the disease among mass people. This work is aimed at providing an analysis of different data mining classification algorithms like Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), etc. on a newly created dataset of 435 patient's risk factors to find the algorithm with the best accuracy to propose a tool for the end users to check stroke risk prediction.

*Keywords*: Stroke, Data mining, risk factor, risk prediction, Naïve Bayes.

## 1.0 Introduction

Stroke, a disruption of blood supply to the brain due to blockage (blood clot) or rupture of blood vessels that causes the death of some brain cells as a result of lack of oxygen supply to them (Walter Johnson, 2016) Stroke is one of the leading causes of death, physical disability, dementia, and depression around the world. The incidence of stroke increases with age and the presence of atherosclerosis along with smoking tobacco, obesity, physical inactivity, hypertension, alcohol intake, diabetes, high blood lipid level, kidney disease, unhealthy diet, stress, male gender, genetic factor etc. (Organization, n.d.) The overall incidence of stroke had exceeded 20% in low to middle-income countries than that of high-income countries from the year 2000 to 2008. It is estimated that globally 1 in 6 people will be affected by stroke in their lifetime. (Organization, n.d.)

About 70% of patients suffering from stroke fail to recognize the initial symptoms and 30% seek medical attention lately regardless of their demographic groups. An interactive and easily accessible end-user tool will be very effective to make the people acquainted with stroke-related risk factors and common sign symptoms for acute prevention and minimize further long- term physical problems as well as financial burden. (Chandratheva, 2010)

In this thesis paper, we are aimed at providing a model for the risk prediction of Stroke. At present, data mining techniques will help us a lot to predict risk. Data mining techniques which include classifications, clustering, association rule mining for finding risk prediction. In this research work, the Naïve Bayes, Support Vector Machine (SVM) classifier algorithm is also used for stroke risk prediction.

## 2.0 Related Work

In this section different research works that were envisioned to predict the risk of stroke and other diseases. We have read a sizeable number of papers that are related to our works but here we are giving a short description of few papers.

### *A.* **STROKE RISK PREDICTION THROUGH NONLINEAR SUPPORT VECTOR CLASSIFICATIONS MODELS** (S.V., 2010)

In this research paper, Sabibullah Mohamed Hanifa and Kasmir Raja S.V presented a study to find the possible risk of stroke by subjecting the risk factors to Support Vector Machines. They used 100 patient's data with 8 attributes. They used Support Vector Classification model parameters through its kernel function named polynomial kernel and Gaussian (RBF) kernel. The authors evaluated the result through Confusion matrix and show that the rate of the correctness of prediction by RBF is 98% whereby polynomial is 92%. So, the author told in this paper that the application of SVM models can be used for the processing of stroke-related risk factor data.

### ***B.*** **PREDICTION OF STROKE USING DATA MINING CLASSIFICATION OF STROKE** (Alshammari, 2018)

This research was carried out by Ohoud Almadami and Riyad Alshammari to predict patient at risk of developing stroke by using data mining techniques and to find the patient with who has higher chances to develop stroke. They used three classifier algorithms namely: C4.5, Jrip and multilayers perceptron (MLP). They collected 969 data sets and divided the dataset into two classes. They have used WEKA Software tool for applying their data mining techniques and for evaluation performance they used 10-fold Cross validation and percentage split. In this research work it is observed that with the comparison of 10-fold cross validation the Jrip classifier algorithm gave the best performance accuracy (92.60%) but after applying PCA on stroke data C4.5 algorithm gave the best performance on test data set (95.25%).

### ***C.*** **ASSESSMENT OF STROKE RISK BASED ON MORPHOLOGICAL ULTRASOUND IMAGE ANALYSIS WITH CONFORMAL PREDICTION** (Antonis Lambour, 2010)

Antonis Lambrou, Harris Papadopoulos et.al have been presented a research paper to provide reliable confidence measures for the assessment of stroke by using the Conformal Prediction framework. They evaluated their results of four different conformal predictions respectively: ANN, NB, SVM, and K-NN. For experimentation, the authors applied Principal Component Analysis (PCA) on the dataset and have selected its 6 features which accounted for 98% of its variance. They have used the Leave-One-Out (LOO) method for evaluating. The classifier algorithm ANN-CP was structured with one hidden layer consist of 3 units and the output layer consisting of 2 units. In this paper, the authors observed that SVM classifier has the best accuracy. However, when the authors increase the percentage of confidence the certainty rates start to decrease.

### ***D.*** **AN INTEGRATED MACHINE LEARNING APPROACH TO STROKE PREDICTION** (Anon., 2010)

Aditya Khosla, Yu Cao et.al. published a paper to present an integrated machine learning approach combining the elements of data imputation, feature selection and prediction. They used different metrics for evaluating their methods such as Notation, Area under the ROC curve (AUC), Concordance Index. They handled their missing data by using mean, median, imputation through linear regression. They evaluated their data imputation quality by using 10-fold cross-validation and used SVM for stroke prediction. They showed in their paper that the combination of Conservative Mean feature selection and Margin-based censored regression gave the best performance.

### ***E.*** **A MODEL FOR PREDICTING ISCHEMIC STROKE USING DATA MINING ALGORITHMS** (Abdelwahab, 2015)

Balar Khalid and Naji Abdelwahab have done a thesis paper to provide a model for predicting Ischemic stroke using data mining algorithms. They have emphasized on prediction and finding risk factors for ischemic stroke so that they could provide a model. They analysed their data sets through C4.5 DT algorithm and Logistic regression and WEKA 3.6 Software tool. They analysed their data by using software of Microsoft "XLSTAT". This software based on Visual Basic language. They have found the rates of sensitivity was 77.58% and specificity was 83.03% and error rate was 19.7% with this "XLSTAT" software and AUC (Area Under Curve) area under ROC (Receiver Operating Curve) equals 0.89.

## 3.0 Question:

Is it true that most people do not recognize the first symptoms presented by stroke?

**Answer:** Correct. Approximately 70% of patients do not correctly recognize their TIA or minor stroke, 30% delay seeking medical attention for >24 hours, regardless of age, sex, social class, or educational level, and approximately 30% of early recurrent strokes occur before seeking attention. Without more effective public education of all demographic groups, the full potential of acute prevention will not be realized.

*Source:* Chandratheva, A., Lasserson, D.S. et al. (June 2010). Population-Based Study of Behavior Immediately After Transient Ischemic Attack and Minor Stroke in 1000 Consecutive Patients: Lessons for Public Education. Stroke, 41, 1108-1114.

## 4.0 System Architecture

Our proposed system architecture has been depicted in fig.1 Initially, an original dataset including the risk factors of 435 people has been used for selecting the best prediction algorithm. The pre-processing stage was associated following the missing tuple handling method. Then the processed dataset was feed to the database (which will be used as a trained dataset for the end-user tool) and to the classification algorithms for simulation. The performance accuracy has been evaluated using 10-Fold Cross Validation and Percentage Split techniques. Finally, according to the best accuracy, the best algorithm will be chosen for enabling the risk prediction feature of the tool.

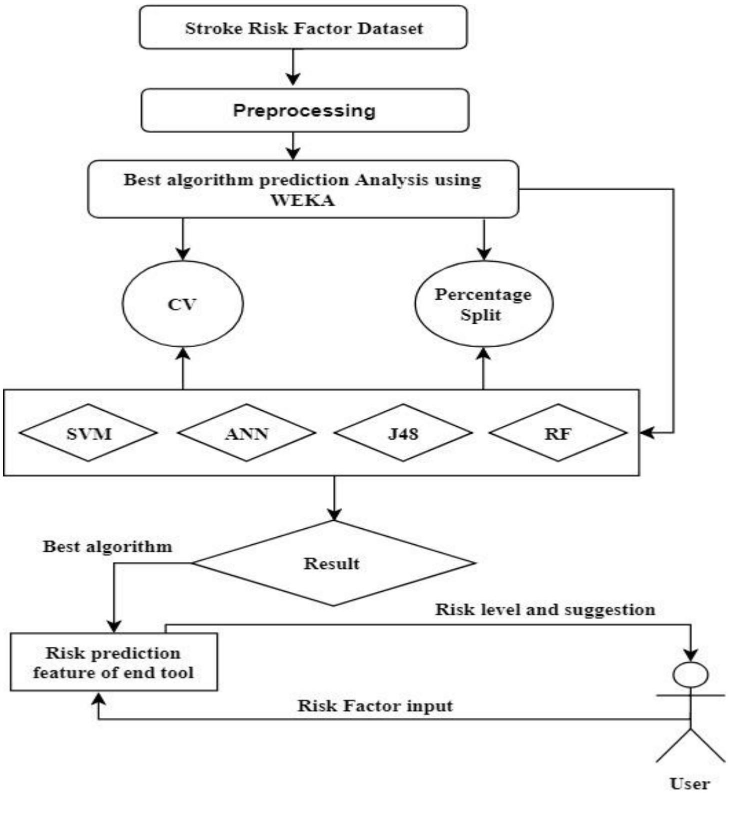

Stroke Risk Factor Dataset
Preprocessing
Best algorithm prediction Analysis using WEKA
CV
Percentage Split
SVM
ANN
J48
RF
Best algorithm
Result
Risk level and suggestion
Risk prediction feature of end tool
Risk Factor input
User

Fig. 1. Proposed System architecture

## 5.0 Methodology

In this section we gave a description about our methodology.

**Input**: Risk factor test dataset.

**Output**: Predicted risk level, suggestions and tips.

Let,

Risk_algorithm, A;

Performance_accuracy, Acc;

Now,

$$T_{train} = T_n - T_\varphi \quad (1)$$

Where, $T_{train}$ represents train data, $T_n$ represents the total number of instances and $T_\varphi$ represents missing tuple.

$$Acc = A\,(T_{train}, T_{test}, R_n) \quad (2)$$

Where, $T_{test}$ represents test data and $R_n$ represents the total number of risk factors.

$$best\_accuracy = \max\,(Acc_i) \quad (3)$$

Here, i is the number of algorithms used for prediction and simulation process.

### 5.1 Flow chart

In this section we are presenting a flow chart of our methodology. Our methodology works according to this flow chart. Here in the first step dataset will be collected and gone to the preprocessing stage to be as binary data from binary nominal data. Then seven most popular classification techniques will apply on the dataset. The performance of all algorithms will be evaluated by 10-Fold Cross Validation and 80:20 percentage split techniques. With the best performed algorithm will be selected and also be used for developing a reliable tool with risk prediction feature for the end user. Then end user will give the answer of all questions from questionnaire form as input. Here the train dataset will be updated every time.

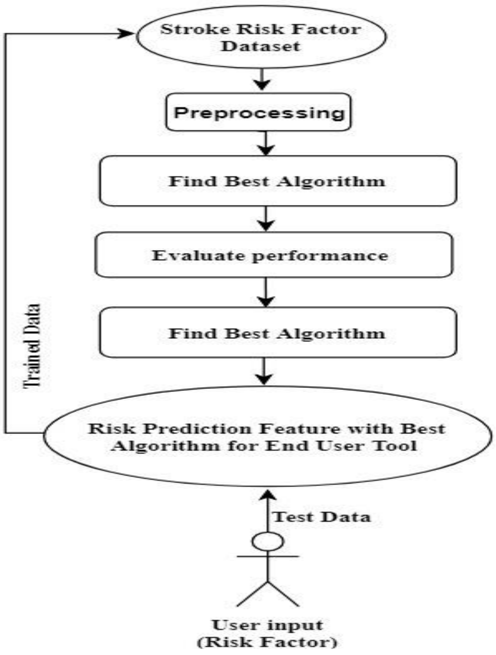


Fig. 2. Flow chart of Methodology

### 6.0 Experimental Analysis

In this section, we represented the details of our dataset and the result analysis.

### 6.1 Details of the training dataset

Our training dataset contains risk factors of 435 persons. We have collected our dataset using a form approved by a medical officer. The details of the dataset are presented in Table 1. Table 2 is representing the attribute description used in the experimental analysis.

Table 1. Description of the training dataset

| | Number of Attributes | Number of Instances |
|---|---|---|
| **Report based dataset** | 15 | 435 |

Table 2: Description of attribute

| **Attributes** | **Values** |
|---|---|
| **Age** | 1.25-34, 2.35-44, 3.45-54,4.55-65,5.65< |
| **Gender** | 1. Male 2. Female |
| **Systolic BP** | 1.120>, 2.120-139, 3. 140-160, 4.160< |
| **Diastolic BP** | 1.180>, 2.80-95, 3.95< |
| **Diabetes** | 1. No, 2. Yes |
| **Ischemic Heart Disease** | 1. No, 2. Yes |
| **Family History of stroke** | 1. No, 2. Yes |
| **Alcoholism** | 1. No, 2. Yes |
| **Less Physically Active** | 1. No, 2. Yes |
| **Smoking** | 1. No, 2. Yes |
| **Stress and depression** | 1. No, 2. Yes |
| **Saturated Fat↑ ()** | 1. No, 2. Yes |
| **Fibre↓ ()** | 1. No, 2. Yes |
| **Chronic Kidney Disease (CKD)** | 1. No, 2. Yes |
| **Class Attribute** | 1. Stroke, 2. Non-stroke |

### 4.2 Details of the result analysis

In our work, we have provided an analysis of different data mining classification algorithms like SVM (Support Vector Machine), NB (Naïve Bayes), RF (Random Forest), J48, ANN (Artificial Neural Network), LR (Logistic Regression), RT (Random Tree). The experimental result has been shown in Table 3.

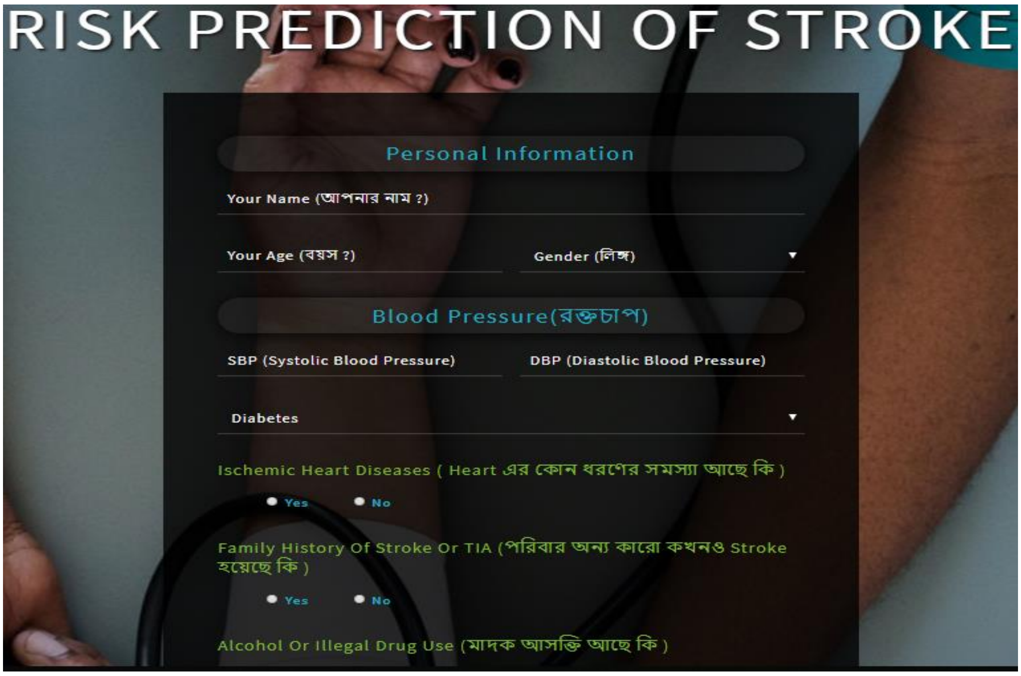
Table 3: Comparison of evaluation metrics using 10fold cross-validation and percentage split (80:20)

| Evaluation Metrics | 10- Fold Cross Validation | | | | | | | Percentage Split (80:20) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NB | SVM | LG | J48 | RT | RF | ANN | NB | SVM | LG | J48 | RT | RF | ANN |
| Total number of the instance | 435 | 435 | 435 | 435 | 435 | 435 | 435 | 87 | 87 | 87 | 87 | 87 | 87 | 87 |
| Correctly Classified Instance % | 345 | 342 | 352 | 365 | 332 | 358 | 353 | 69 | 66 | 67 | 70 | 65 | 68 | 69 |
| | 79.31 | 78.62 | 80.919 | 83.91 | 76.32 | 82.29 | 81.15 | 79.31 | 75.86 | 77.01 | 80.46 | 74.71 | 78.16 | 79.31 |
| Incorrectly classified instance | 90 | 93 | 83 | 70 | 103 | 77 | 82 | 18 | 21 | 20 | 17 | 22 | 19 | 18 |
| | 20.69 | 21.38 | 19.08 | 16.1 | 23.68 | 17.7 | 18.85 | 20.689 | 24.14 | 22.989 | 19.54 | 25.29 | 21.84 | 20.69 |

Although the Random Tree algorithm is a very famous algorithm for classification but for our dataset, the performance accuracy was not about satisfactory. It has been correctly classified only 76.32% in cross-validation evaluation techniques where Random forest algorithm has been correctly classified 82.29% and J48 has been correctly classified 83.91%. In percentage split evaluation, J48 has been correctly classified 80.46% where Random tree has been correctly classified 74.71%. However, the best accuracy has been acquired by the J48 decision tree algorithm for both cases.

### 7.0 Proposed tool for the end user

It is our motto to provide an easily accessible and effective tool for our end users to make them acquainted with the risk factors of having stroke, so that people can seek medical attention even before developing stroke and reduce the stroke related mortality, morbidity and financial burden. At the same time, making people aware about the risk factors and seeking medical attention timely will delay or even prevent the disease development process as well as prevent further attack of stroke.  In this modern era of technology, almost all the people regardless their demography know the use of websites and web technology. So, we have preferred web technology where a simple website could be beneficial to check the risk of developing stroke by using user's risk factors as input. This website will also provide some useful suggestion and tips to the end users to avoid developing stroke and seek medical attention timely. In the fig 2 a demo input page is given below:



RISK PREDICTION OF STROKE

Personal Information

Your Name (আপনার নাম ?)

Your Age (বয়স ?)

Gender (লিঙ্গ) ▾

Blood Pressure(রক্তচাপ)

SBP (Systolic Blood Pressure)

DBP (Diastolic Blood Pressure)

Diabetes ▾

Ischemic Heart Diseases ( Heart এর কোন ধরণের সমস্যা আছে কি )

○ Yes ○ No

Family History Of Stroke Or TIA (পরিবার অন্য কারো কখনও Stroke হয়েছে কি )

○ Yes ○ No

Alcohol Or Illegal Drug Use (মাদক আসক্তি আছে কি )

Fig. 2. Proposed Tool for the end user

### 8.0 Conclusion

All the statistics are showing that the global prevalence of stroke is rising where people are still unaware of the risk factors of developing stroke. Knowing the risk factors by any means could make them alert to reduce the incidence of stroke and its aftermaths effectively. This research paper presented a system for risk prediction of stroke by using the different data mining techniques. As data mining techniques have some algorithms for predicting many diseases so we used some algorithms of data mining techniques namely SVM, NB, RF, J48, MLP etc. (Pragati Agrawal, 2015) We observed in this work that the RF algorithm gave the best accuracy. We also provide a tool for the end users so that they can know the risk level of their having stroke. With the help of this tool, we can increase awareness about stroke. However, we have collected only 435 data as our train dataset, it can be updated by increasing the number of instances and can be implemented in others data mining techniques for prediction purpose.

## 9.0 **References**

Abdelwahab, B. K. a. N., 2015. A MODEL FOR PREDICTING ISCHEMIC STROKE USING DATA MINING ALGORITHMS. *International Journal of Innovative Science, Engineering & Technology,* 2(11), pp. 1-6.

Alshammari, O. A. a. R., 2018. PREDICTION OF STROKE USING DATA MINING CLASSIFICATION OF STROKE. *International Journal of Advanced Computer Science and Applications.,* Volume 9, pp. 1-4.

Anon., 2010. *AN INTEGRATED MACHINE LEARNING APPROACH TO STROKE PREDICTION.* [Online]
Available at: http://people.csail.mit.edu/khosla/papers/kdd2010.pdf
[Accessed 2 January 2019].

Anon., 2010. C. ASSESSMENT OF STROKE RISK BASED ON MORPHOLOGICAL ULTRASOUND IMAGE ANALYSIS WITH CONFORMAL PREDICTION. *International Federation for Information Processing,* pp. 1-8.

Antonis Lambour, H., 2010. C. ASSESSMENT OF STROKE RISK BASED ON MORPHOLOGICAL ULTRASOUND IMAGE ANALYSIS WITH CONFORMAL PREDICTION. *International Federation for Information Processing,* pp. 1-8.

Chandratheva, A. L. D. e. a., 2010. Population-Based Study of Behavior Immediately After Transient Ischemic Attack and Minor Stroke in 1000 Consecutive Patients. *Lessons for Public Education,* p. 41.

Organization, W. S., n.d. *World Stroke Organization.* [Online]
Available at: https://www.world-stroke.org/component/content/article/16-forpatients/84-facts-and-figures-about-stroke
[Accessed 3 January 2019].

Pragati Agrawal, A. K. D., 2015. A Brief Survey on the Techniques using for the Diagnosis of Diabetes-mellitus. *International Research Journal of Engineering and Technology (IRJET),* 1108-1114(02(03)), p. 1039.

S.V., S. M. H. a. K. R., 2010. STROKE RISK PREDICTION THROUGH NONLINEAR SUPPORT VECTOR CLASSIFICATIONS MODELS. *International Journal of Advanced Research in Computer Science,* Volume 01, pp. 1-7.

Walter Johnson, O. O. ,. M. O. &. S. S., 2016. *World Health Organization.* [Online]
Available at: https://www.who.int/bulletin/volumes/94/9/16-181636/en/
[Accessed 3 January 2019].