RISK PREDICTION OF STROKE USING DATA MINING CLASSIFICATION TECHNIQUES

Rahatara Ferdousi [1][0000-1111-2222-3333], Masum Mohammad Jubayel[2][1111-2222-3333-4444]
Asima Akter Chowdhury[3][2222-3333-4444-5555] and M M Faniqul Islam [4][3333-4444-5555-6666]

[1] Metropolitan University, Sylhet 3100 Bangladesh
[2]Metropolitan University, Sylhet 3100 Bangladesh
[3]Metropolitan University, Sylhet 3100 Bangladesh
[4] Queen Mary University of London, UK

rahatara@metrouni.edu.bd
masumjubayel@gmail.com
chowdhuryasima@gmail.com
m.islam@smd17.qmul.ac.uk

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, SVM, ANN, J48, Random Forest, Decision Tree.

1 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 [1]. 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. [2]. 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 [2].

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 [3].

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 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.

2.1 STROKE RISK PREDICTION THROUGH NONLINEAR SUPPORT VECTOR CLASSIFICATION MODELS [4]

In this research paper Sabibullah Mohammaed 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. The authors used the SVMLight software for implementation. Their data set was extracted from cohort of population set of various hospital situated at Tiruchirappali city, Tamilnadu, India. They used 100 patient's data with 8 attribute called Hypertension, Diabetes Mellitus, Obesity, Cigarette Smoking, Heart Disease, Prior Stroke, High Cholesterol, and Physical Activity. 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 correctness of prediction by RBF is 98% where by 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.

2.2 DIABETES PREDICTION USING MEDICAL DATA [5]

In this research paper the author presented a diabetes prediction system to diagnosis diabetes. They tried to improve the accuracy in diabetes prediction using medical data with various supervised machine learning algorithms namely Naïve Bayes (NB), Mul-

tilayer-perceptron (MLP), Random Forest (RF) and the accuracy is noted with different test methods such as 10-fold cross validation (FCV), use percentage split with 66%(PS), and using training dataset (UTD) with pre-processing method and without pre-processing methods. For this work the author used WEKA software tool. They used PID dataset and collected dataset with the medical report of 768 persons include 8 features. However, the concluded that the pre-processing methods increase accuracy for Naive Bayes algorithm.

2.3 A REVIEW ON PREDICTION OF MULTIPLE DISEASES AND PERFORMANCE ANALYSIS USING DATA MINING AND VISUALAIZATION TECHINIQUES [6]

The paper was motivated to construct a basic prototype model which can determine unknown knowledge related with multiple disease from past database records of specified multiple diseases. In this research work, at first the authors experimented on the dataset consisting of 1000 records and 14 attribute and the Naive Bayes algorithm gave better accuracy than others. Then the authors experimented again on the dataset of 1000 records within 76 attributes and back-propagation algorithm gave the best accuracy (100%) where the accuracy is given by Naive Bayes algorithms is 90.74%. However, the authors told in their paper that the accuracy result can be changed if we increase the number of datasets and the number of attributes.

2.4 BREAST CANCER CLASSIFICATION USING SUPPORT VECTOR MACHINE AND NEURAL NETWORK [7]

In this work, the author tried to carry out using Wisconsin Diagnosis breast Cancer database to classify the breast cancer as either benign or malignant. They used dataset consists of 400 observations of patients. Among them 300 are benign and 100 are malignant status. Each instance has 20 features. They used two classes for classifications named cancerous cell and non-cancerous cell and the experiment was carried out by using SVM and Neural Network. The result both SVM and Neural Network were compared on the basis of accuracy and precision. For this experiment, NN technique is more efficient compare to SVM technique. However, the author observed that NN technique is more efficient but the difference is very low.

2.5 DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES [8]

The authors have employed two algorithms namely J48 and Naïve Bayes to create the model for diagnosis in determining diabetes in women. They used PID (Pima Indians Diabetes) database of National Institute of diabetes and Digestive and Kidney Diseases dataset pre-processed in CSV format as their input. The data was divided into training set and test set by 10-fold cross validation and percentage split techniques. They pre-processed their dataset by using Weka software tools. The J48 algorithm is used

on the dataset using WEKA after which data are divided into “tested-positive” or “tested-negative” depending on the final result of the decision tree that is constructed.

2.6 APPLICATIONS OF MACHINE LEARNING IN CANCER PREDICTION AND PROGNOSIS [9]

In this paper, the author have described how the machine learning works well in cancer prediction and prognosis. They have been showed two histograms. The first one is about the steady increase in published papers using machine learning to predict cancer risk, recurrence and outcome. The another one is about the frequency with which different types of machine methods namely: Naïve Bayes, Genetic algorithm, Fuzzy logic, Clustering, SVM, ANN, Decision Tree are used to predict different types of cancer such as bladder, breast, colorectal, liver, lung, lymphoma, Prostate, Skin, Throat etc. They discussed in their paper about 3 case study which was about prediction of cancer risk or susceptibility, survivability, recurrence. They worked very well. However, all machine learning studies are not conducted with the same attention by them.

2.7 LIVER DISEASE PREDICTION USING SVM AND NAÏVE BAYES ALGORITHMS [10]

The author have been used classification algorithms namely Naïve Bayes and SVM for liver diseases prediction in this research work and they implemented their work in Mat lab 2013 tool. The authors described the classification of liver diseases such as cirrhosis, bile duct, chronic hepatitis, liver cancer and acute hepatitis. In this work paper, they showed that SVM algorithms works better than Naïve Bayes algorithms on the basis of performance accuracy. But if we think about execution time, then Naïve Bayes classifier needs minimum execution time.

2.8 A BRIEF SURVEY ON THE TECHNIQUES USED FOR TECHNIQUIES THE DIAGNOSIS OF DIABETES MELLITUS [11]

This paper described about different data mining methods such as k-ford cross validation and classification, Class wise K Nearest Neighbor, SVM, LDA-Support Vector Machine and Feed Forward Neural Network, Artificial Neural Network, Statistical Normalization, Back Propagation. At first the authors have been collected 768 cases but after deleting the missing values they had 460 cases for their experiment. They had compared the accuracy of the performance of those classification algorithms and observed that the SVM gave best accuracy as 81.77% compare to others.

2.9 PREDICTION OF STROKE THROUGH STACKED TOPOLOGY OF ANN MODELS [12]

This paper was motivated to predict the stroke risk by proposing the stacked ANN topology model with higher prediction accuracy. They have collected 300 data for their paper from different hospitals at Tiruchirappalli city, Tamilnadu, India. They analysed their data by using back propagation algorithm and implemented those data through MATLAB 7.3.0-Neural Network Toolbox. They divided their output in three categories namely: High risk, Moderate risk and Low risk. They used ANN model consist of three layers, respectively input layer, output layer and one internal layer. They used Confusion matrix as a method for finding error. They also presented a graphical structure of their predicted and actual outputs of all network. They have been presented a good paper. However, sometimes they made a big difference between their predicted outputs and actual outputs.

2.10 PREDICTION AND CONTROL OF STROKE BY DATA MINING [13]

This paper described about stroke and the risk factors of stroke. The authors collected 807 data sets within 50 risk factors for stroke by using a standard checklist during year 2010-2011 in Iran. After pre-processing and cleaning data they have used WEKA Software tool for implementation and data mining techniques such as K-nearest neighbor and C4.5 decision tree for analysing the data sets. In this work they have found the performance accuracy of the C4.5 decision tree was 95.42% and the K-nearest neighbor was 94.18%. So they told C4.5 decision tree and K-nearest neighbor can be use for prediction of stroke. However, the authors presented a description about C4.5 decision tree and K-nearest neighbors only.

2.11 ASSESSMENT OF STROKE RISK BASED ON MORPHOLOGICAL ULTRASOUND IMAGE ANALYSIS WITH CONFORMAL PREDICTION [14]

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 prediction 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 of 98% of its variance. They have used 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 consist 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.

2.12 PREDICTION OF STROKE USING DATA MINING CLASSIFICATION OF STROKE [15]

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 multi layers perceptron (MLP). They collected 969 data sets from National Guard hospitals in three different cities in Kingdom of Saudi Arabia. They collected their data in 2016 from 2nd January to 31st September. They divided their data sets into two classes. First one includes with the patient who have stroke and the 2nd class includes the patient who has mimic stroke but they have diagnosis as they have stroke. The authors have made a train data set to build their model with 10-fold cross validation and have made a test data set to evaluate the model. They have used WEKA Software tool for applying their data mining techniques. 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%).

2.13 A MODEL FOR PREDICTING ISCHEMIC STROKE USING DATA MINING ALGORITHMS. [16]

Balar Khalid and Naji Abdelwahab have done a thesis paper to provide a model for predicting Ischemic stroke using data mining algorithms. They have emphasised 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 a 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 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.

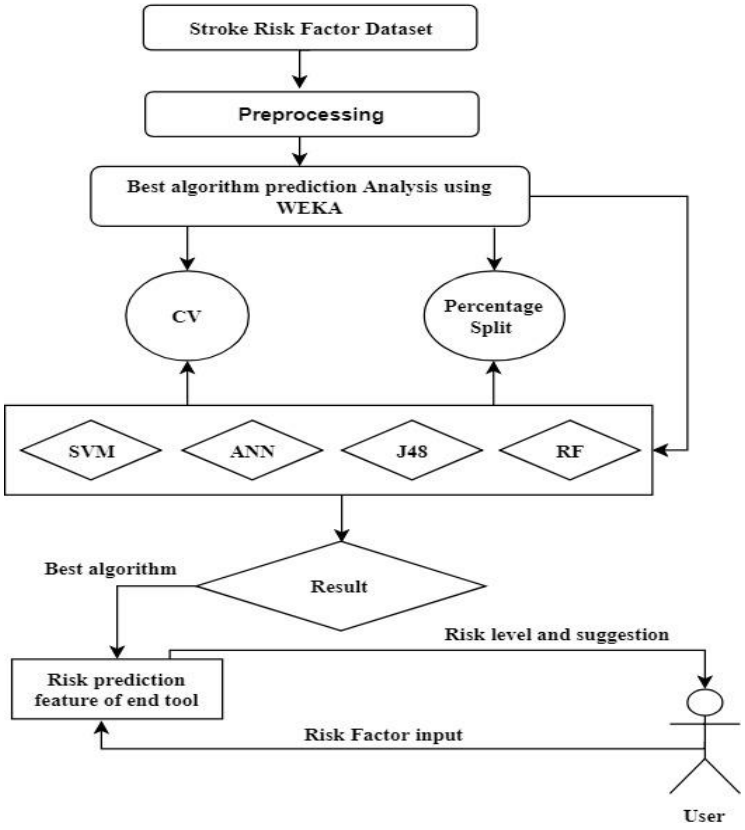

Fig. 1. Proposed System architecture

4 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.

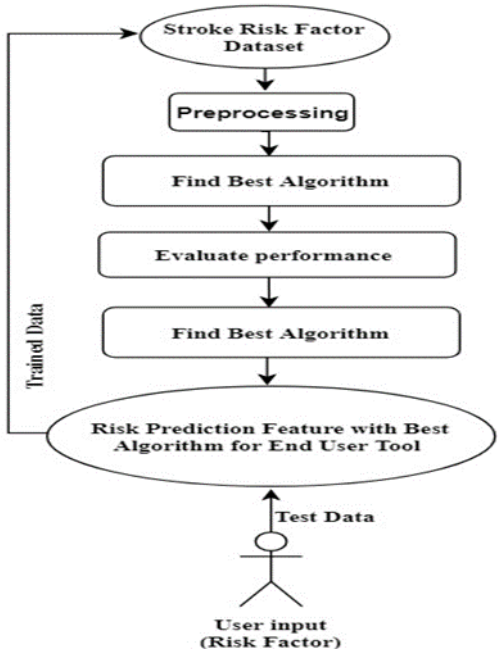

Fig. 2. Flow chart of Methodology

5 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,
$Ttrain = Tn - T\varphi$ (1)
Where, Ttrain represents train data, Tn represents the total number of instances and T φ represents missing tuple.
$Acc = A (Ttrain, Ttest, Rn)$ (2)
Where, Ttest represents test data and Rn represents the total number of risk factors.
$best\_accuracy = \max (Acci)$ (3)
Here, i is the number of algorithms used for prediction and simulation process.

6 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	1.80>, 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

6.2 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	80.92	83.91	76.32	82.29	81.15	79.31	5.86	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 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 dis-

ease 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:

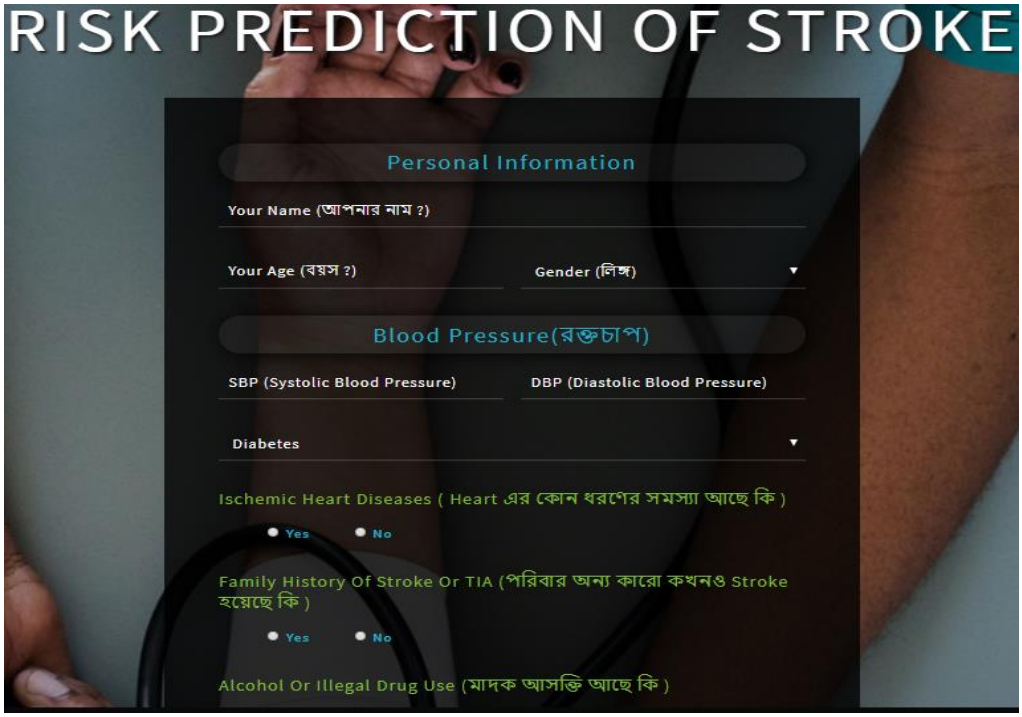

Fig. 3. Proposed Tool for the end user

8 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. [17]. 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.

References

1. O. O. ,. M. O. &. S. S. Walter Johnson, "World Health Organization," 2016. [Online]. Available: https://www.who.int/bulletin/volumes/94/9/16-181636/en/. [Accessed 3 January 2019].
2. W. S. Organization, "World Stroke Organization," [Online]. Available: https://www.world-stroke.org/component/content/article/16-forpatients/84-facts-and-figures-about-stroke.
3. A. L. D. e. a. Chandratheva, "Population-Based Study of Behavior Immediately After Transient Ischemic Attack and Minor Stroke in 1000 Consecutive Patients," *Lessons for Public Education*, p. 41, 2010.
4. K. R. S. Sabibullah Mohamed Hanifa, "Stroke Risk Prediction through Non-linear Support Vector Classification Models," *International Journal of advanced Research in Computer science*, vol. 1, pp. 1-8, 2010.
5. D. E. J. L. B. S. B. Dr. D. Asir Antony Gnana Singh, "Diabetes Prediction Using Medical Data," *Journal of Computational Intelligence in Bioinformatics*, vol. 10, pp. 1-8, 2017.
6. H. S. N. F. S. Ajinkya Kunjir, "A Review on Prediction of Multiple Diseases and Performance Analysis using Data Mining and Visualization Techniques," *International Journal of Computer Applications (0975 – 8887)*, vol. 155, pp. 35-38, 2016.
7. W. Z. F. Ebrahim Edriss Ebrahim Ali, "Breast Cancer Classification using Support Vector Machine and Neural Network," *International Journal of Science and Research (IJSR)*, vol. 5, no. 3, pp. 1-6, 2013.
8. S. J. R. S. Aiswarya Iyer, "Diagnosis of diabetes using classification mining techniques," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 5, pp. 1-14, 2015.
9. D. S. W. Joseph A. Cruz, "Applications of Machine Learning in Cancer Prediction and Prognosis," *SAGE Journals*, pp. 1-19, 2006.
10. M. Dr. S. Vijayarani, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 4, no. 4, pp. 1-5, 2015.
11. A. k. D. Pragati Agrawal, "A BRIEF SURVEY ON THE TECHNIQUES USED FOR THE DIAGNOSIS OF DIABETES-MELLITUS," *International Research Journal of Engineering and Technology (IRJET)*, vol. 2, no. 3, pp. 1-5, 2015.
12. K. R. S. Sabibullah M, "Prediction of Stroke Risk through Stacked Topology of ANN Model," *International Journal of Advanced Research in Computer Science*, vol. 1, pp. 1-10, 2010.

13. R. A. M. T. F. S. A. M. F. J. K. R. N. N. T. Leila Amini, "Prediction and Control of Stroke by Data Mining," *International journal of preventive medicine · ,* vol. 4, pp. 1-5, May 2013.

14. H. P. E. K. S. P. M. S. P. G. a. A. N. Antonis Lambrou, "Assessment of Stroke Risk Based on Morphological Ultrasound Image Analysis with Conformal Prediction," *IFIP International Federation for Information Processing,* vol. 339, pp. 146-153, 2010.

15. M. o. H. I. R. A. Ohoud Almadani, "Prediction of Stroke using Data Mining Classification Techniques," *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 9, no. 1, pp. 1-4, 2018.

16. B. K. a. N. Abdelwahab, "A Model for Predicting Ischemic Stroke Using Data Mining Algorithms," *IJISET - International Journal of Innovative Science, Engineering & Technology,* vol. 2, no. Issue 11, pp. 1-6, November 2015..

17. J. Celko, The Morgan Kaufmann Series in Data Management Systems, 2nd December 2014.