Risk Prediction of Stroke at The Early Stage Using Data Mining Techniques

1st Masum Mohammad Jubayel
Computer Science and Engineering
Metropolitan University, Sylhet 3100
Bangladesh
Sylhet 3100, Bangladesh
masumjubayel@gmail.com

2nd Asima Akter Chowdhury
Computer Science and Engineering
Metropolitan University, Sylhet 3100
Bangladesh
Sylhet 3100, Bangladesh
chowdhuryasima@gmail.com

3rd Mohammad Golam Sabbir
Computer Science and Engineering
Metropolitan University, Sylhet 3100
Bangladesh
Sylhet 3100, Bangladesh
sabbirsabbu@gmail.com

***Abstract*— Stroke, a non-communicable fatal disease causing death to the mass of Bangladesh each year increasingly. If the risk factors of stroke are under control, the treatment would be beneficial. Thus, predicting the risk of developing stroke can aid the individuals to keep their risk factors under control through maintaining precautions, which would delay the progress of developing the risk of stroke. As Data Mining has multiple classification techniques, which are able to predict risk, to analyze the hidden patterns of patients data, contemporary researches have been carried out to predict disease risks from sign, symptoms or risk factors. In this work firstly, we have prepared a dataset of common risk factors of stroke of 635 patients from a direct questionnaire. Secondly, a detailed analysis of this dataset with Data Mining classification algorithms namely Naïve Bayes (NB), Random Forest (RF), Random Tree (RT), Support Vector Machine (SVM) have been carried out. In addition, the performance of these algorithms has been evaluated through 10 fold Cross-Validation and 80:20 Percentage Split to identify the most accurate one for this dataset. As we found Random Forest Decision Tree as the best algorithm for this dataset, finally we generated the decision tree using Jupyter a Python IDE and proposed a knowledge-based system prototype to predict the risk of developing stroke and to provide essential suggestions to the mass.**

Keywords—Stroke, Data Mining, Risk Factor, Risk Prediction, Naïve Bayes, SVM, Random Forest, Random Tree, Decision Tree.

I. INTRODUCTION

A stroke is a sudden interruption in the blood supply of the brain. The effects of a stroke depend on which part of the brain is injured, and how severely it is injured. Strokes may cause sudden weakness, loss of sensation, or difficulty with speaking, seeing, or walking [5]. If the prediction of risk factors is possible, it might be possible to lower risk factors and prevent or delay a stroke [6]. 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 [3]. However, in a third world developing country like Bangladesh, the treatment procedures after stroke result in an economic burden.

In the healthcare industry, Data Mining plays an important role in predicting diseases or chance of developing diseases [1]. Data mining techniques have great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis [2].

Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Random Tree (RT) Support Vector Machine (SVM), etc. supervised classification DM algorithms have been applied to predict disease or its risk mostly. However, the performance varies due to the type of dataset as well as a pattern of information. As the focus of disease or risk of disease prediction systems is the mass of a country, an algorithm with better accuracy should be identified before developing and deploying a system to the mass [4].

The rest of the paper is organized as, in section 2 we have provided literature studies, in section 3 we have provided our proposed system architecture, in section 4 we have provided our dataset description, in section 5 we have demonstrated our experimental analysis with result, and finally in section 6 we have concluded this paper.

II. LITERATURE REVIEW

In [6], the authors used two different data mining techniques named Naïve Bayes and J48 decision tree for the prediction of various diseases such as heart disease, diabetes, and breast cancer and compared their performance for evaluating the best classifier. They found the highest accuracy for diabetes using J48. However, they didn't provide the details of their dataset. They used WEKA tool but they did not show the results which were obtained by using the percentage split and cross-validation separately.

In [7], authors analyzed the application of the most popular data mining techniques in the medical domain and they used some of the algorithms for disease prediction. They showed that by using various tools and techniques on different disease diagnosis a variety of results can be gained. However, they did not show the result which was obtained by using different data mining techniques.

In [8], the authors presented a study about the model of logistic regression and obtained the result with "XLSTAT" software. They showed several steps of the logistic regression model in their study. They found the sensitivity rate was 77.58%, the specificity rate was 83.03% and the error rate was 19.7%.

In [9], the authors presented the prediction of risk of stroke within 10 years and their dataset was within 150,0000 men and 120,0000 women had used data from the national health examination of the entire nation and they classified the total population in five ranges such as normal, slightly high, high, risky and very risky.

In [10], author presented a study to find the possible risk of stroke by subjecting the risk factors to SVM. They used 100 patient's data with 8 attributes. They used SVM model parameters through its kernel function named polynomial kernel and Gaussian (RBF) kernel. They evaluated the result through the Confusion matrix and showed that the rate of the correctness of prediction by RBF was 98% whereby

polynomial was 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.

III. System Architecture

Our system architecture has been delineated in Fig. 1. Initially, an original dataset including the risk factors of 635 people has been used for selecting the best prediction algorithm. After cleaning data, risk factors of 606 peoples has been used for the experiment. Then the processed dataset was fed 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 80:20 Percentage Split techniques. Finally, according to the best accuracy, the best algorithm has been chosen for enabling the risk prediction feature of the tool.

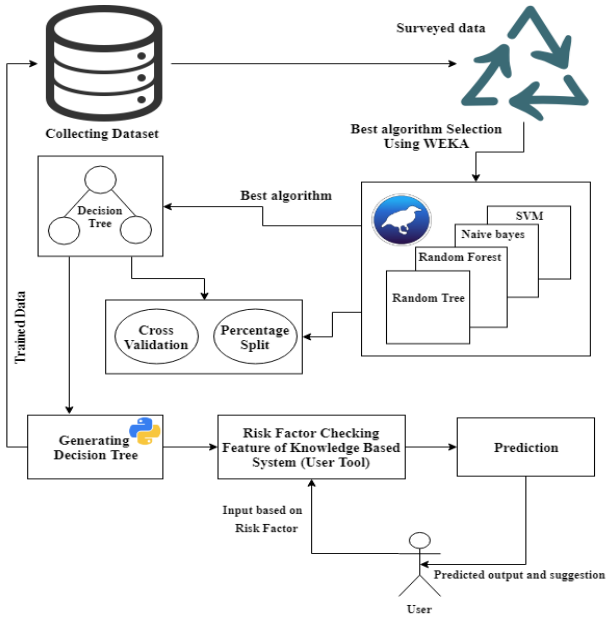

Fig. 1 Proposed System Architecture

IV. Dataset Description

In this section, we represent the details of our dataset and attribute description. This dataset contains the information of 635 persons. This dataset has been created from a direct questionnaire to people who have recently developed stroke, or who are still not developed the stroke but having few or more risk factors of stroke. The data has been collected from the patients using direct questionnaires from the different hospitals in Sylhet, Bangladesh. We have collected the information from Sylhet Women's Medical College & Hospital and Jalalabad Ragib Rabeya Medical College & hospital. The description of the dataset is given below.

TABLE I. Description of the Training Dataset

	Number of Attributes	Number of Instances
Report based dataset	15	635

TABLE II. Description of Attributes

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

The data pre-processing has been conducted by handling the missing values following the technique of ignoring the tuples with incomplete values. After pre-processing, 606 instances have remained in total. Among them, 451 are positive values and 155 are negative values. The detail description of the attributes is shown in Fig 2. Two class variables are used to find whether the patient is having a risk of developing stroke (positive) or not (negative).

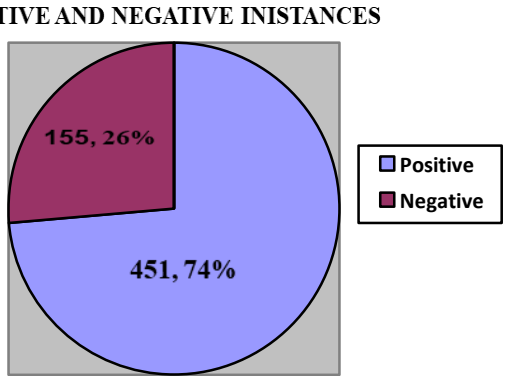

Fig. 2 Class Attributes Distribution

V. Background Study

According to the World Health Organization, 15 million people suffer stroke worldwide each year. Of these, 5 million die and another 5 million are permanently disabled. In developed countries, the incidence of stroke is declining, largely due to efforts to lower blood pressure and reduce smoking. However, the overall rate of stroke remains high due to the aging of the population. [11]

A. Data Mining

Data mining is a process of collecting or gathering hidden data from large datasets where dataset means a collection of

data. Data mining process is used for finding a pattern or similarity and relationship among large data set.

Mining process can be thought as a genuine appraisement of information technologies. There are many different issues in research which can be implemented by using data mining process and techniques. In health care sector, a huge amount of data is adding day by day. So, we preferred to use data mining techniques in our research work.

Knowledge Discovery in Databases (KDD)

Data mining is a part of KDD. In KDD process at first the data that we have collected through database from the user interface, will be stored in data warehouse. In warehouse data will be checked or tested that whether this data is good or not for the user or the operation.

After this KDD complete the rest four steps. They are transformation, mining, interpretation or evaluation and knowledge.

B. Decision Tree

A decision tree is like a flowchart includes with a root node, leaf nodes. It can perform very well in KDD and data mining by enabling the model and knowledge extraction from the given dataset. It is able to perform with missing data or value. It can handle many types of input dataset like Numeric, Textual and Nominal.

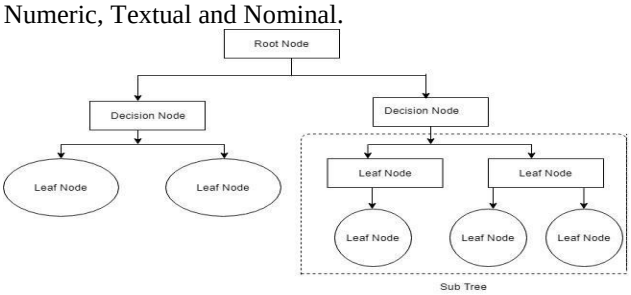

Fig 3. Decision Tree

Decision tree algorithms represents itself by following tree structure. The figure 4 shows how the decision tree algorithm works.

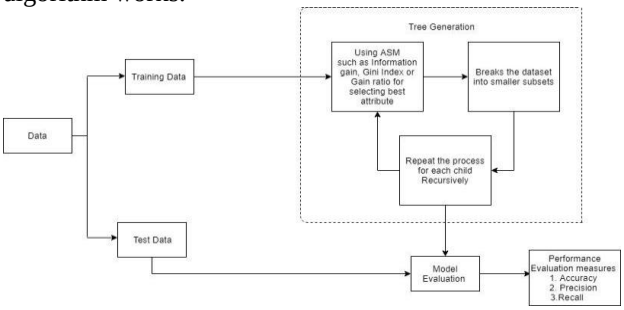

Fig 4. The working process of decision tree algorithm.

For selecting the best attribute, it uses ASM (Attribute Selection Measures) and for continuous-valued it uses some most popular attribute selection measures like Information gain, Gain Ration and Gini index.

Information Gain is one of the most popular attribute selection measure which measures the entropy also. Based on the given dataset, information gain computes the difference between the entropy before split and the average entropy after split. Information gain is used by ID3 decision tree. There are three equations from (1) to (3) are given below. Information gain follows those equation for measures.

$$\text{Info}(S) = -\sum_{i=1}^{k} pi \log_2 pi \quad (1)$$

$$\text{Info}_A(S) = \sum_{j=1}^{v} \left(\frac{|Sj|}{|S|} * \text{Info}(Sj) \right) \quad (2)$$

$$\text{Gain}(A) = \text{Info}(S) - \text{Info}_A(S) \quad (3)$$

Where, Pi is the probability that an attribute tuple S belongs to class Ci.

Where, Info(S) is the required information for identifying the class label within S attribute tuple, Info A(S) is expected information, (|Sj|)/(|S|) is the weight of the jth partition and i is an incremental number .Here, attribute A will be selected within the highest range of value of Gain (A) as the splitting attribute.

We generated a plotted decision tree according to our datasets which is shown in figure 5. For displaying this decision tree, we used Scikit-learn's export_graphviz function and pydotplus also.

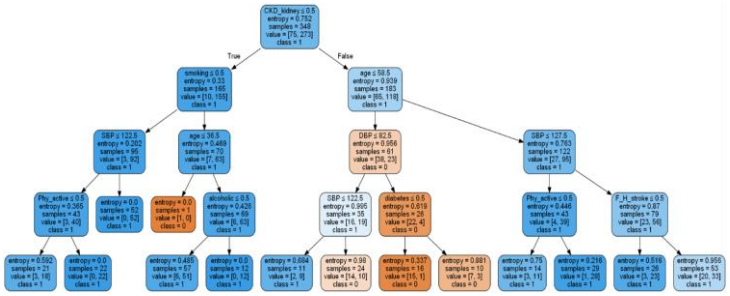

Fig 5. Generated decision tree according to our dataset.

VI. Result Analysis

Performance of different data mining techniques on our dataset with detailed accuracy information is represented in the following tables. Although Support Vector Machine is one of the most popular algorithms for prediction. In the case of our dataset, the accuracy was the lowest for both evaluation techniques those are the Cross-validation methods and the Percentage Split. However, the best result was achieved by using the Random Forest decision tree.10-fold cross-validation classified 84.16% instances correctly while percentage split technique classified 80.99% of the instances correctly. In table 3 to table 18, we have depicted the detailed analysis of the result. We have found the correctly classified instances and incorrectly classified instances for each algorithm.

TABLE III. Performance Results from Random Forest decision tree using (Cross Validation)

	Number of Instances	**Percentage**
Correctly classified Instances	510	84.16%
Incorrectly Classified Instances	96	15.84%

TABLE IV. DETAILED ACCURACY BY CLASS FROM RANDOM FOREST DECISION TREE USING 10-FOLD CROSS VALIDATION TECHNIQUE

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.665	0.098	0.701	0.665	0.682
	0.902	0.335	0.887	0.902	0.895
Weighted Average	0.842	0.275	0.839	0.842	0.840

TABLE V. PERFORMANCE RESULTS FROM RANDOM FOREST DECISION TREE USING PERCENTAGE SPLIT

	Number of Instances	Percentage
Correctly classified Instances	98	80.99%
Incorrectly Classified Instances	23	19.0083%

TABLE VI. DETAILED ACCURACY BY CLASS FROM RANDOM FOREST DECISION TREE USING PERCENTAGE SPLIT (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.585	0.075	0.800	0.585	0.676
	0.925	0.415	0.813	0.925	0.865
Weighted Average	0.810	0.300	0.809	0.810	0.801

TABLE VII. PERFORMANCE RESULTS FROM NAÏVE BAYES ALGORITHM USING (CROSS VALIDATION)

	Number of Instances	Percentage
Correctly classified Instances	449	74.09 %
Incorrectly Classified Instances	157	25.91%

TABLE VIII. DETAILED ACCURACY FROM CLASS NAÏVE BAYES WITH 10-FOLD CROSS VALIDATION TECHNIQUE

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.303	0.109	0.490	0.303	0.375
	0.891	0.697	0.788	0.891	0.837
Weighted Average	0.741	0.546	0.712	0.741	0.718

TABLE IX. PERFORMANCE RESULTS FROM NAÏVE BAYES - PERCENTAGE SPLIT

	Number of Instances	Percentage
Correctly classified Instances	82	67.77%
Incorrectly Classified Instances	39	32.23%

TABLE X. DETAILED BY CLASS ACCURACY FROM NAÏVE BAYES– PERCENTAGE SPLIT (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.195	0.075	0.571	0.195	0.291
	0.925	0.805	0.692	0.925	0.791
Weighted Average	0.678	0.558	0.651	0.678	0.622

TABLE XI. PERFORMANCE RESULTS FROM RANDOM TREE, A DECISION TREE ALGORITHM USING CROSS VALIDATION)

	Number of Instances	Percentage
Correctly classified Instances	489	80.69%
Incorrectly Classified Instances	117	19.31%

TABLE XII. DETAILED ACCURACY FROM RANDOM TREE USING 10-FOLD CROSS VALIDATION TECHNIQUE

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.677	0.149	0.610	0.677	0.642
	0.851	0.323	0.885	0.851	0.868
Weighted Average	0.807	0.278	0.815	0.807	0.810

TABLE XIII. PERFORMANCE RESULTS FROM RANDOM TREE USING - PERCENTAGE SPLIT

	Number of Instances	Percentage
Correctly classified Instances	95	78.51%
Incorrectly Classified Instances	26	21.49%

TABLE XIV. DETAILED ACCURACY FROM RANDOM TREE USING - PERCENTAGE SPLIT (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.659	0.150	0.692	0.659	0.675
	0.850	0.341	0.829	0.850	0.840
Weighted Average	0.785	0.277	0.783	0.785	0.784

TABLE XV. PERFORMANCE RESULTS FROM SUPPORT VECTOR MACHINE ALGORITHM USING (CROSS VALIDATION)

	Number of Instances	Percentage
Correctly classified Instances	451	74.42%
Incorrectly Classified Instances	155	25.56%

TABLE XVI. DETAILED ACCURACY FROM SUPPORT VECTOR MACHINE USING - CROSS VALIDATION

	TP Rate	FP Rate	Precision	Recall	F-measure
	0	0	0.582	0	0.623
	1	1	0.786	1	0.853
Weighted Average	0.744	0.744	0.029	0.744	0.021

TABLE XVII. PERFORMANCE RESULTS FROM SUPPORT VECTOR MACHINE ALGORITHM – PERCENTAGE SPLIT

	Number of Instances	Percentage
Correctly classified Instances	80	66.12 %
Incorrectly Classified Instances	41	33.88%

TABLE XVIII. DETAILED ACCURACY FROM SUPPORT VECTOR MACHINE USING - PERCENTAGE SPLIT (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0	0	0.366	0	0.223
	1	1	0.661	1	0.796
Weighted Average	0.661	0.661	0.026	0.661	0.013

For more semantic view of the performance of used algorithms for both evaluation techniques two graphs have been depicted. In Fig. 6, the performance of the algorithms using Cross-validation evaluation is depicted and in Fig. 7, the results from percentage split have been shown to represent the comparative accuracy of the used algorithms.

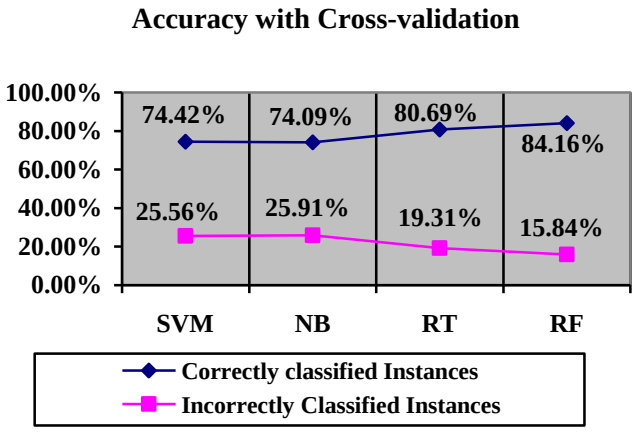

Fig. 6 Performance of Classification Algorithms Using 10 Fold Cross-Validation Technique

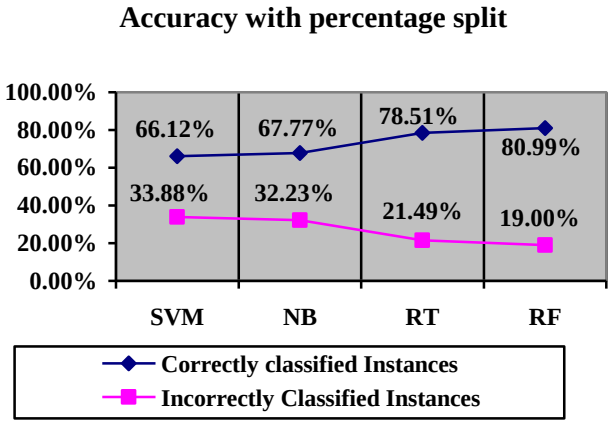

Fig. 7 Performance of Classification Algorithms Using Percentage Split Technique

From Table 3 to Table 18 we can see that the decision tree gives the best accuracy performance according to our dataset. After seeing this, we have analyzed our dataset by implementing the code of the decision tree in python. We implemented the code in python in two ways. Firstly, we implemented the code for 10-fold cross-validation. For this, we splitted the dataset into 10-fold and calculate the percentage of the performance accuracy. Then we calculated the Gini Index of the attributes and attribute values for the dataset and create a terminal node value. Secondly, we implemented the code for Percentage Split. For this, we splitted the dataset, 80% as trained data and 20% as test data. We found 76.17% the mean accuracy for 10 Fold cross-validation and 80.33% accuracy for Percentage Split and generated a decision tree.

VII. PROPOSED TOOL FOR THE END USERS

We propose a web application with a user-friendly interface, where a knowledge-based website asking questions in both Bengali and English could be beneficial to check the risk of developing stroke from user's common risk factors as input. However, any other region adopting this idea can change the language according to them. This concept was made due to reach mass people of every stage with the contribution of this research work. This website would be able to provide

some useful suggestions and tips to the end-users to avoid developing stroke and seek medical attention timely. In Fig. 8 and Fig. 9 a demo input page has been depicted. In Fig. 10 and Fig. 11 a demo output page is depicted, which shows the generated prediction of having a stroke resulting in whether an individual is at risk or not.

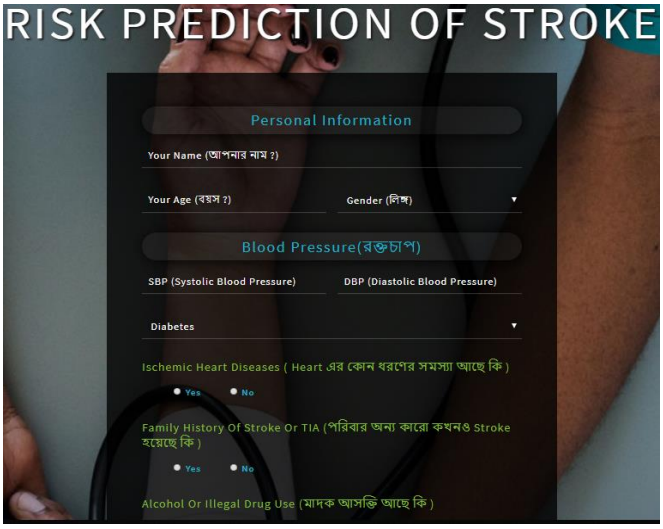

Fig. 8 Homepage of the Proposed Website/System

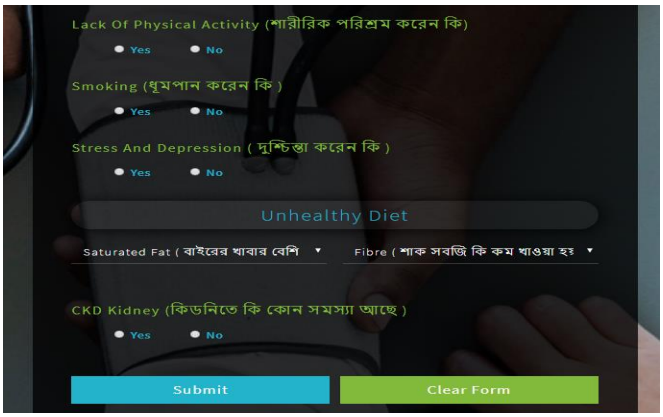

Fig. 9 Homepage of the Proposed Website/System

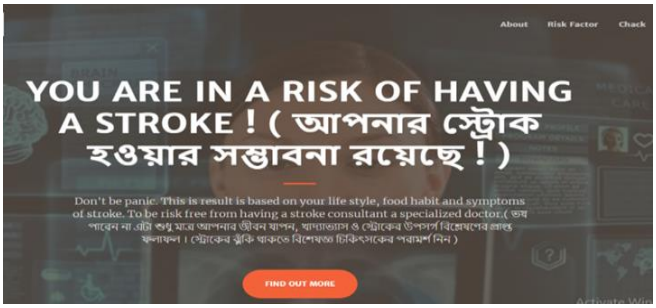

Fig. 10 Risk Checking Page of Stroke

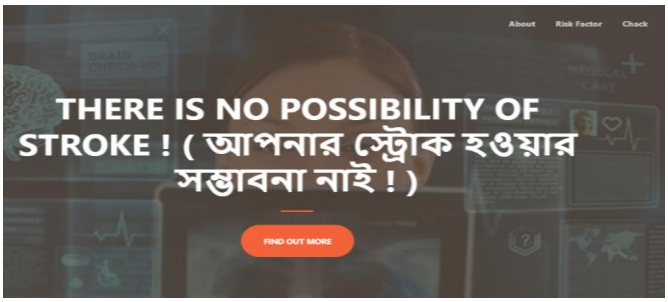

Fig. 11 Risk Checking Page of Stroke

VIII. Conclusion

Most of the statistics are showing that the global prevalence of stroke is rising where people are still unaware of the risk factors of developing a stroke. Therefore, knowing the chance of developing risk by any means could make them alert to re-duce the incidence of stroke and its aftermaths effectively. In this work, we have conducted a detailed analysis using multiple Data Mining Techniques and found Random forest Decision Tree as the best one providing accuracy at 84.16%. We also proposed a prototype for the end-user application. However, analysis with more size-able dataset and a widely deployed mobile apps could be an integrated research scope of this work.

References

[1] *AICN*. (2014, April 30). Retrieved from AICN: http://www.ieeeottawa.ca/aicn/data-mining-and-knowledge-discovery-in-healthcare-and-medicine/Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).

[2] Kamaraj, K. (2016). Multi Disease Prediction using Data Mining Techniques. *Research Gatee*, 1-2.

[3] Chandratheva A1, L. D. (2010, April 15). Population-based study of behavior immediately after transient ischemic attack and minor stroke in 1000 consecutive patients: lessons for public education. Retrieved from US National Library of Medicine National Institutes of Health: https://www.ncbi.nlm.nih.gov/pubmed/20395614

[4] Jiawei Han, M. K. (2001). *Data Mining Concepts and Techniques.* Academic Press, Morgan Kaufmann Publishers.

[5] Siri Krishan Wasan, V. B. (2006). The Impact of Data Mining Techniques On Medical Diagnostics. *Data Science Journal*, 119-126.

[6] D. D. S. P. K.GomathiKamaraj, "ResearchGate," 17 September 2017. [Online]. Available: https://www.researchgate.net/publication/319851535_Multi_Disease_Prediction_using_Data_Mining_Techniques. [Accessed 03 July 2019].

[7] S. M. P. V. Kirubha, "Survey on Data Mining Algorithms in Disease Prediction," International Journal of Computer Trends and Technology (IJCTT), vol. 38, pp. 124-128, 2016.

[8] N. A. Balar Khalid, "A Model for Predicting Ishcemic Stroke Using Data Mining Algorithms," IJISET- International Journal of Innovative Science, Engineering & Technology, vol. 2, no. 11, pp. 18-23, 2015.

[9] H.-s. L. Jae-woo-Lee, "The development and implementation of stroke risk prediciton model in National Health insurance service's personal health record.".

[10] K. R. S. Sabibullah Mohamed Hanifa, "Stroke Risk Prediction Through Non-linear Support Vector Classification Models," Internation Journal of Advanced Research in Computer Science, vol. 1, pp. 47-53, 2010.

[11] W. S. Organization, "World Stroke Organization," [Online]. Available:https://www.worldstroke.org/component/content/article/16-forpatients/84-facts-and-figures-about-stroke.