**RISK PREDICTION OF STROKE**

**USING DATA MINING TECHNIQUES**

**Introduction:**

In healthcare industry, data mining plays an important role in predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of tests can be reduced. This reduced test plays an important role in time and performance [1]. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis [2].

Knowledge Discovery is the nontrivial process of extracting implicit, novel, and useful information from large volume of data. Major applications of Knowledge Discovery and Data Mining in healthcare fall into four categories: (a) Clinical Medicine: Modern hospitals and clinical centers surpassed their traditional role as a place for diseases’ diagnosis and treatment and now acting as a mass database and a source of complex clinical, laboratory, equipment use, and drug management data which can be analyzed for disease diagnosis and decision making; (b) Public Health: including early outbreak detection, healthcare and syndromic surveillance; (c) Healthcare Text mining: including mining medical literature, as well as mining clinical data such as patients’ clinical records; and (d) Healthcare Policy and Planning: including detecting expensive clinical profiles among patients diagnosed with a specific chronic illness which has a high disease’s burden such as diabetes[3].

Data mining plays an important role in predicting diseases. Through different types of data mining technique heart disease, Diabetes and Breast cancer and so on can be predicted easily [4].

A stroke is a sudden interruption in the blood supply of the brain. Most strokes are caused by an abrupt blockage of arteries leading to the brain ([ischemic stroke](http://www.strokecenter.org/patients/about-stroke/ischemic-stroke/)).  Other strokes are caused by bleeding into brain tissue when a blood vessel bursts ([hemorrhagic stroke](http://www.strokecenter.org/patients/about-stroke/intracerebral-hemorrhage/)). Because stroke occurs rapidly and requires immediate treatment, stroke is also called a brain attack. When the symptoms of a stroke last only a short time (less than an hour), this is called a transient ischemic attack (TIA) or mini-stroke. 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 you may be able to lower your risk factors and prevent or delay a stroke [6].

**Literature Review:**

1. **Multi Disease Prediction Using Data Mining Techniques.**[1]6

In this paper 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 the percentage split and cross-validation separately.

1. **Survey on Data Mining Algorithms in Disease Prediction.** [2]**7**

In this paper V. Kirubha and S. Manju Priya analyzed the application of the most popular data mining techniques in medical domain and they used some of the algorithms for disease prediction. In their study they showed that, by using various tools and techniques on different disease diagnosis a varieties number of results can be gained. However, they did not show the result which were obtained by using different data mining techniques.

1. **A Model for Predicting Ischemic Stroke Using Data Mining Algorithms.**[3]**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 have found the sensitivity rate was 77.58%, the specificity rate was 83.03% and the error rate was 19.7%.

1. **The development and implementation of stroke risk prediction model in National Health Insurance Service’s personal health record.**[4]**9**

In this paper the authors presented the prediction of risk of stroke within 10 years.

And their dataset was within 1500000 men and 1200000 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.

1. **Stroke Risk Prediction through Non-linear Support Vector Classification Models.**[5]**10**

In this paper authors 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 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 stoke-related risk factor data.

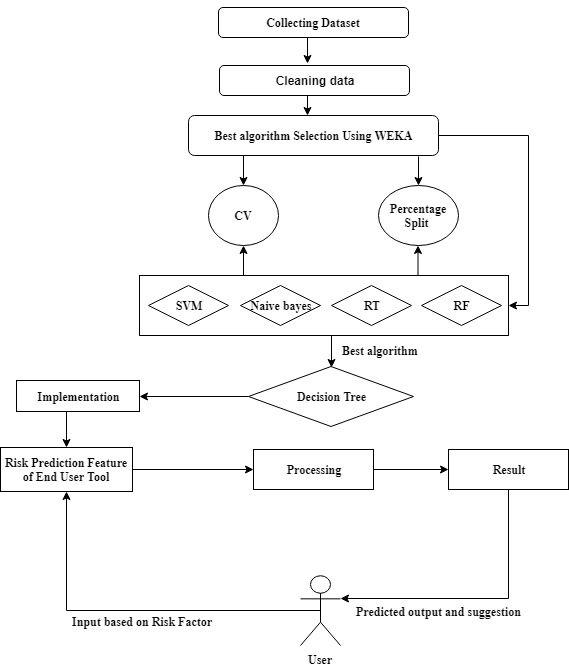
****

Figure 1:Proposed System Architecture

**System Architecture:** Our system architecture has been delineated in Figure 1. Initially, an original dataset including the risk factors of 435 people has been used for selecting the best prediction algorithm. Dataset was cleaned then as most of the people don’t have any idea of their Cholesterol level and all of them are in same races. 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.

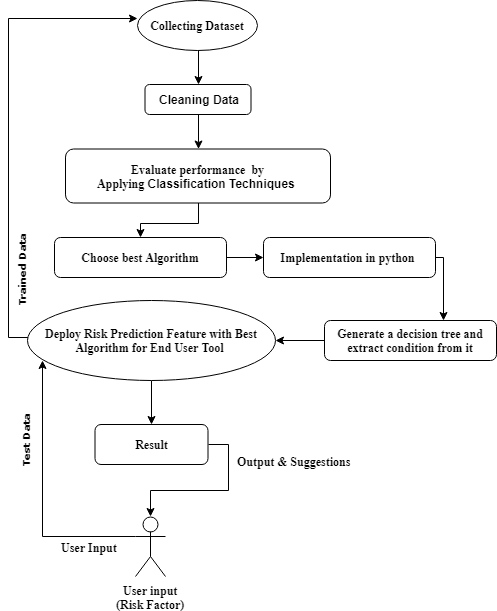
****

Figure 2:Flow Chart of Methodology:

**Flow Chart:** A flow chart of our methodology has been portrayed in Figure 2. Our methodology works accord zing to this flow chart. Here in the first step dataset will be collected and cleaned. After that Dataset will be sent 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.

**Experimental Analysis**

Dataset details and the result analysis is represented in this section.

**Dataset Details**

This dataset contains the information of 606 persons. It includes data about peoples including risk factors of developing stroke that may cause stroke. 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 questionnaire from diﬀerent hospital in Sylhet, Bangladesh. We have collected the information from Sylhet Woman Medical College & Hospital and Jalalabad Ragib Rabeya Medical College & hospital. The description of dataset is given below.

|  |  |  |
| --- | --- | --- |
|  | Number of Attributes | Number of Instances |
| Risk Factors Dataset | 15 | 606 |

# Table 1:Description of Dataset

|  |  |
| --- | --- |
| 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 |

# Table 2:Description of Attribute

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 been remained in total. Among them, 451 are positive values and 155 are negative values. The detail description of the attributes is shown in Table 2. Two class variables are used to ﬁnd whether the patient is having a risk of developing of stroke (positive) or not (negative).

Figure 3:Class Attributes Distribution

**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 data prediction. In case of our dataset, the accuracy was the lowest for both of the Cross-validation method and also for the Percentage Split. However, the best result was achieved by using RandomForest decision tree. Where using 10-fold cross validation, 84.16% instances were classiﬁed correctly and using percentage split technique, it could classify 80.99% of the instances correctly. In table 3 to table 18 we have depicted the detail analysis of the result. We have found the correctly classiﬁed instances and incorrectly classiﬁed instances for each algorithm.

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

# Table 3: Performance Results from RandomForest decision tree using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 4: Detailed Accuracy by class from RandomForest decision tree using 10-fold Cross Validation Technique

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

# Table 5: Performance Results from RandomForest decision tree using Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 6: Detailed Accuracy by class from RandomForest decision tree using - Percentage Split (80:20)

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

# Table 7: Performance Results from Naïve Bayes Algorithm using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 8: Detailed Accuracy from class Naïve Bayes with 10-fold Cross Validation technique

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

# Table 9: Performance Results from Naïve Bayes - Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 10: Detailed by class Accuracy from Naïve Bayes– Percentage Split (80:20)

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

# Table 11: Performance Results from RandomTree, a Decision Tree Algorithm using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 12: Detailed Accuracy from RandomTree using 10-fold Cross Validation technique

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

# Table 13: Performance Results from RandomTree using - Percentage Split

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 14: Detailed Accuracy from RandomTree using - Percentage Split (80:20)

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

# Table 15: Performance Results from Support Vector Machine Algorithm using (Cross Validation)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0 | 0 | ? | 0 | ? |
|  | 1 | 1 | 0.786 | 1 | 0.853 |
| Weighted Average | 0.744 | 0.744 | ? | 0.744 | ? |

# Table 16: Detailed Accuracy from Support Vector Machine using - Cross Validation

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

# Table 17: Performance Results from Support Vector Machine Algorithm – Percentage Split

[

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TP Rate | FP Rate | Precision | Recall | F-measure |
|  | 0 | 0 | ? | 0 | ? |
|  | 1 | 1 | 0.661 | 1 | 0.796 |
| Weighted Average | 0.661 | 0.661 | ? | 0.661 | ? |

# Table 18: Detailed Accuracy from Support Vector Machine using 10 Percentage Split (80:20)

For the more semantic view of the performance of used algorithms using both evaluation techniques are depicted in graphs. In Fig. 4, the performance of the algorithms using Cross-validation evaluation is depicted and in Fig. 5, the results from percentage split have been shown to represent the comparative accuracy of the used algorithms.

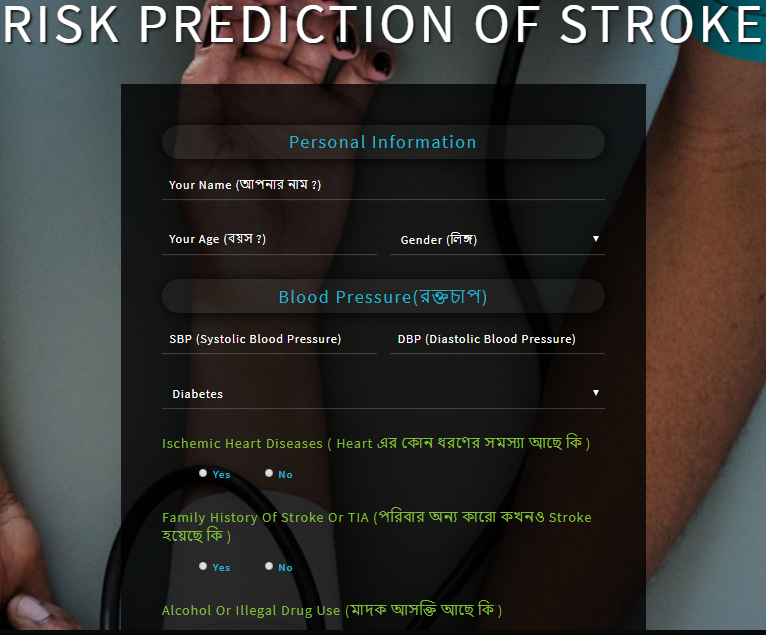
# Figure 4:Performance of Classiﬁcation Algorithms Using Cross-Validation Technique

# Figure 5:Performance of Classiﬁcation Algorithms Using Percentage Split Technique

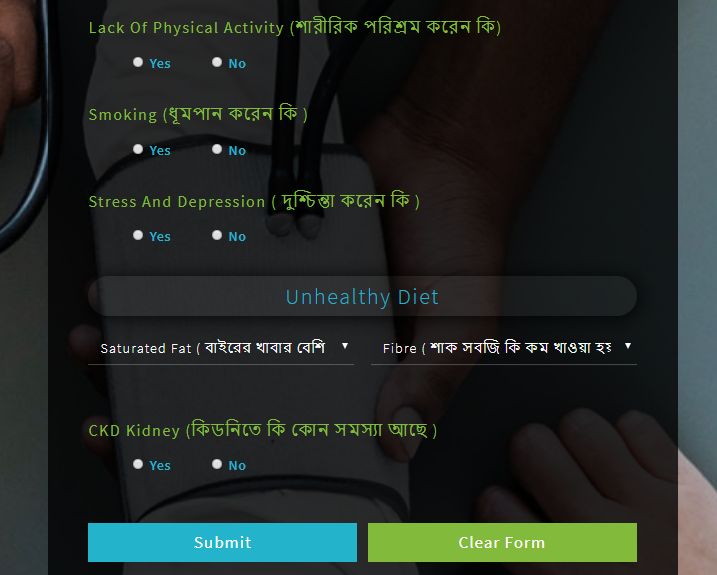
From table 3 to 18 we can see that, the decision tree gives the best accuracy performance according to our dataset. After seeing this we have analyzed our dataset by implementing the code of the decision tree in python. We implemented the code in python in two ways. First, we implemented the code with k-fold cross validation. For this we split the dataset into k-fold and calculate the percentage of the performance accuracy. Then we calculate the Gini Index of the attributes and attributes values for the dataset and create a terminal node value. Then we generate a decision tree by following the CART (Classification and Regression Tree) algorithm and find 80.33% the mean accuracy of the algorithm.

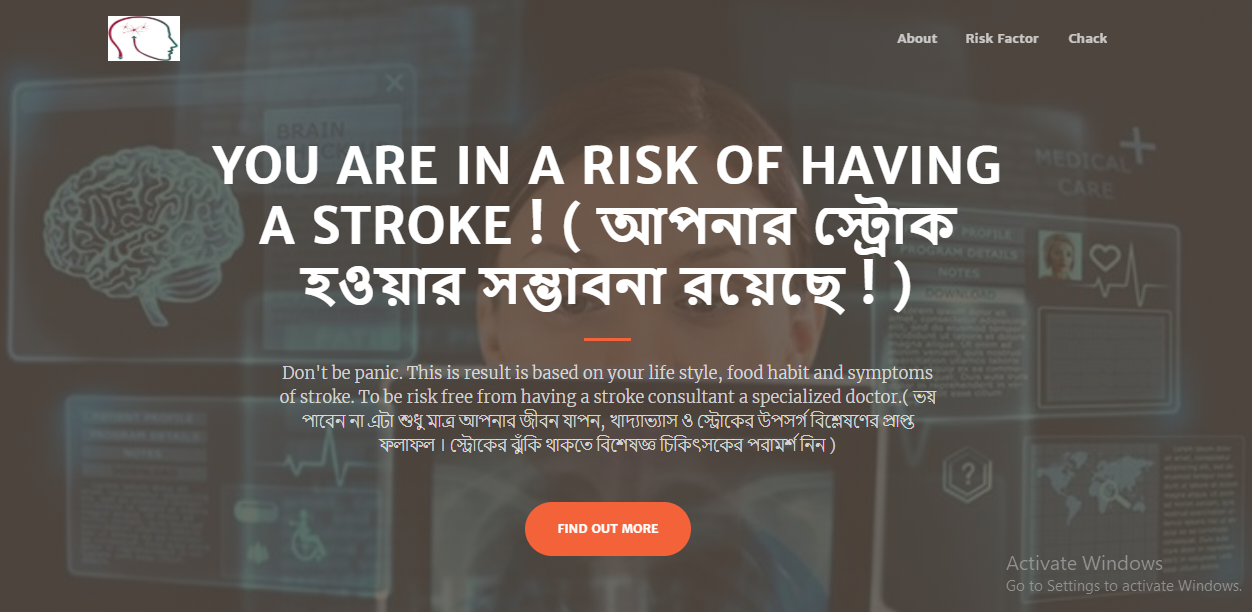
**Proposed Tool for the End Users**

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. However, any other regions adopting this idea can change the language according to theirs. This concept was made due to reach mass people of every stage with the contribution of this research work. 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 figure 6 a demo input page is given below:

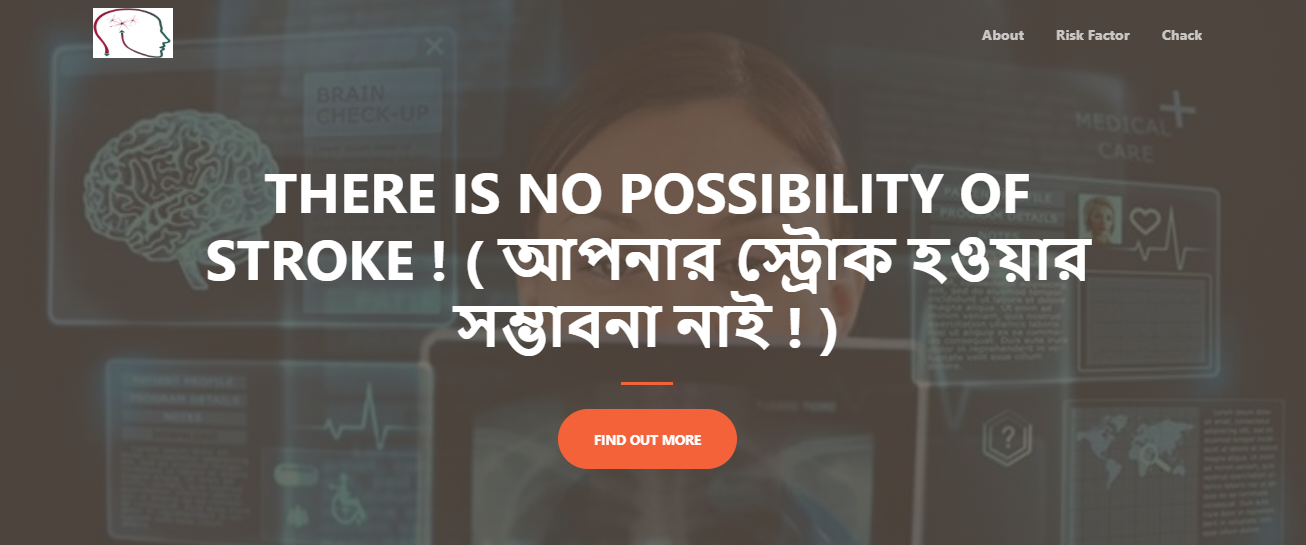
* Can predict the risk of developing stroke in the early stage.
* Can find some healthy tips and helpful suggestions.
* Can be aware before developing stroke.

# Figure 6:Homepage of the Proposed Website/System



Figure 7:Homepage of the Proposed Website/System

# Figure 8:Risk Checking Page of Stroke



# Figure 9:Risk Checking Page of Stroke

# **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 , RandomForest, RandomTree and so on.[11] We observed in this work that the J48 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 606 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.

## **Limitations**

* As it is a newly created dataset within the risk factors only. So, it is quite small.
* Since, we collected our dataset from different hospitals in Sylhet, it was hard to find the stroke patients. So, if anyone can collect data from different states then the effectiveness of the dataset can be increased

# References

1.*AICN*. (2014, April 30). Retrieved from AICN: http://www.ieeeottawa.ca/aicn/data-mining-and-knowledge-discovery-in-healthcare-and-medicine/

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

National Heart, L. a. (n.d.). *patient, patient and family*. Retrieved from National Heart, Lung and Blood Inistitution: https://www.nhlbi.nih.gov/health-topics/stroke

*Patient*. (n.d.). Retrieved from The internet stroke Center: http://www.strokecenter.org/patients/about-stroke/what-is-a-stroke/

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

# Bibliography

|  |  |
| --- | --- |
| [1] | D. D. S. P. K.Gomathi Kamaraj, "ResearchGate," 17 September 2017. [Online]. Available: https://www.researchgate.net/publication/319851535\_Multi\_Disease\_Prediction\_using\_Data\_Mining\_Techniques. [Accessed 03 July 2019]. |
| [2] | 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. |
| [3] | 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. |
| [4] | H.-s. L. Jae-woo-Lee, "The development and implementation of stroke risk prediciton model in National Health insurance service's personal health record.". |
| [5] | 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. |
| [6] |  |
| [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] | H.-s. L. Jae-woo-Lee, "The development and implementation of stroke risk prediction model in National Health Insurance Service's personal health record," *Elsevier,* pp. 253-257, 2017. |