



**RISK PREDICTION OF STROKE FROM THE  
RISK FACTORS USING DATA MINING  
TECHNIQUES**

Masum Mohammad Jubayel

ID: 151-115-105

Asima Akter Chowdhury

ID: 151-115-082

# Paper Content

- Introduction
- Related work
- Objectives & Contribution
- Dataset details table
- Questionnaire form of risk prediction of Stroke
- Flow chart of Methodology
- Mathematical Representation of Methodology
- System Architecture
- Decision Tree
- Mathematical representation of Decision Tree
- Result
- Comparison of evaluation metrics using 10-fold cross-validation and percentage split (80:20)
- Proposed Tool
- Limitation
- Conclusion & Future Scope
- References

# Introduction

## What is Stroke ?

A stroke occurs if the flow of oxygen-rich blood to a portion of the brain is blocked [\[1\]](#)

## Data mining for risk prediction model:

Data mining classification techniques like- NB, SVM, ANN, J48, RF, RT, LG are using for reliable risk prediction of disease due to having the capability of finding hidden of relevant data.

## Role of risk factor of stroke for prediction:

The more risk factor a patient has, the more chances to have a stroke [\[2\]](#)

Risk factor can be controlled [\[3\]](#)

# Related work

Paper Name	Author Name	Dataset information	Used data mining techniques	Obtained accuracy
STROKE RISK PREDICTION THROUGH NONLINEAR SUPPORT VECTOR CLASSIFICATIONS MODELS	Sabibullah Mohammed Hanifa and Kasmir Raja S.V	100 data with 8 attributes	Confusion matrix (RBF and polynomial)	98%, 92%
PREDICTION OF STROKE USING DATA MINING CLASSIFICATION OF STROKE	Ohoud Almadami and Riyadh Alshammari	969 data	C4.5, Jrip and multilayers perceptron	92.60%, 95.25%
ASSESSMENT OF STROKE RISK BASED ON MORPHOLOGICAL ULTRASOUND IMAGE ANALYSIS WITH CONFORMAL PREDICTION	Antonis Lambrou, Harris Papadopoulos	6 features	ANN, NB, SVM, and K-NN	98%
A MODEL FOR PREDICTING ISCHEMIC STROKE USING DATA MINING ALGORITHMS	Balar Khalid and Naji Abdelwahab	265 data	C4.5 DT algorithm and Logistic regression	77.58% , 83.03%

# Objectives & Contribution

- To prepare an appropriate dataset of risk factors
- To use the most popular risk prediction data mining algorithm for the dataset
- To obtain the best algorithm
- To design an end-user tool (website)  
for mass people with the  
\*checking feature developed with the best algorithm.

# Dataset details table

## 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

## Description of the training dataset

	Number of Attributes	Number of Instances
Report based dataset	15	435

# Questionnaire form of risk prediction of Stroke

## Questionnaire form of risk prediction of stroke

Name				
Age				
Gender	Male <input type="checkbox"/> Female <input type="checkbox"/>			
Blood pressure	SBP	a) 120> <input type="checkbox"/>	b) 120-139 <input type="checkbox"/>	c) 140-160 <input type="checkbox"/> d) 160< <input type="checkbox"/>
	DBP	a) 80> <input type="checkbox"/>	b) 80-94 <input type="checkbox"/> c) 95< <input type="checkbox"/>	
Diabetes	Diabetic <input type="checkbox"/> Non-diabetic <input type="checkbox"/>			
Ischaemic heart diseases	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Race	Asian <input type="checkbox"/> European <input type="checkbox"/> African <input type="checkbox"/> Tribal <input type="checkbox"/>			
Family history of stroke or TIA	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Alcohol or illegal drug use	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Lack of physical activity	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Smoking	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Stress and Depression	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Abnormal Cholesterol levels	TG ↑	Yes <input type="checkbox"/> No <input type="checkbox"/>		
	LDL ↑	Yes <input type="checkbox"/> No <input type="checkbox"/>		
	HDL ↓	Yes <input type="checkbox"/> No <input type="checkbox"/>		
Unhealthy diet	Saturated fat	Yes <input type="checkbox"/> No <input type="checkbox"/>		
	Fibber	Yes <input type="checkbox"/> No <input type="checkbox"/>		
CKD kidney	Yes <input type="checkbox"/> No <input type="checkbox"/>			
Class (Attribute)	Stroke <input type="checkbox"/> Non -Stroke <input type="checkbox"/>			

# Flow chart of Methodology

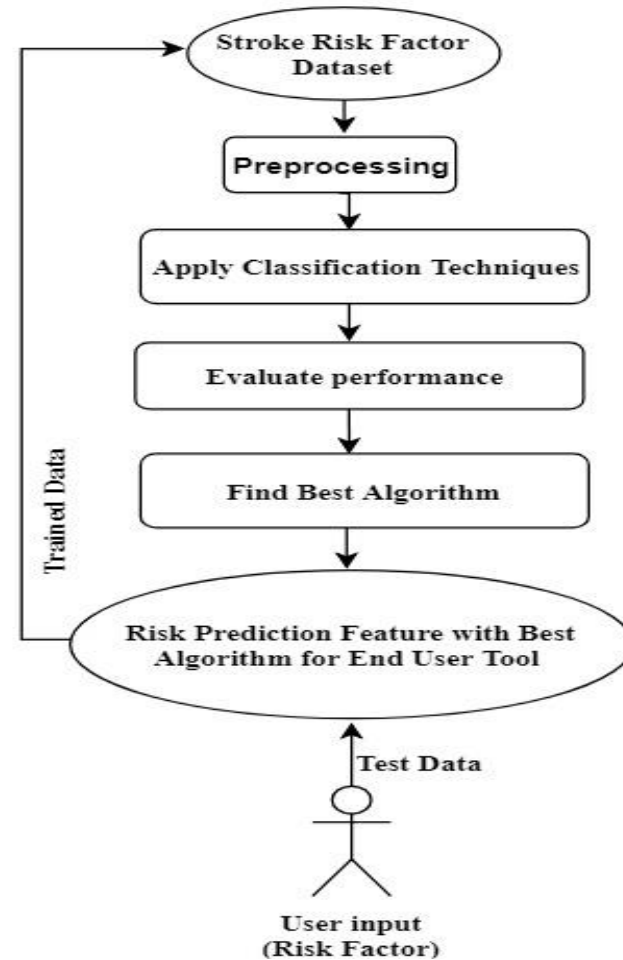


Fig. 1. Flow chart of Methodology



# Mathematical Representation of 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,

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

Where,

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

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

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

$$\text{best\_accuracy} = \max(\text{Acc}_i) \quad (3)$$

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

# System Architecture

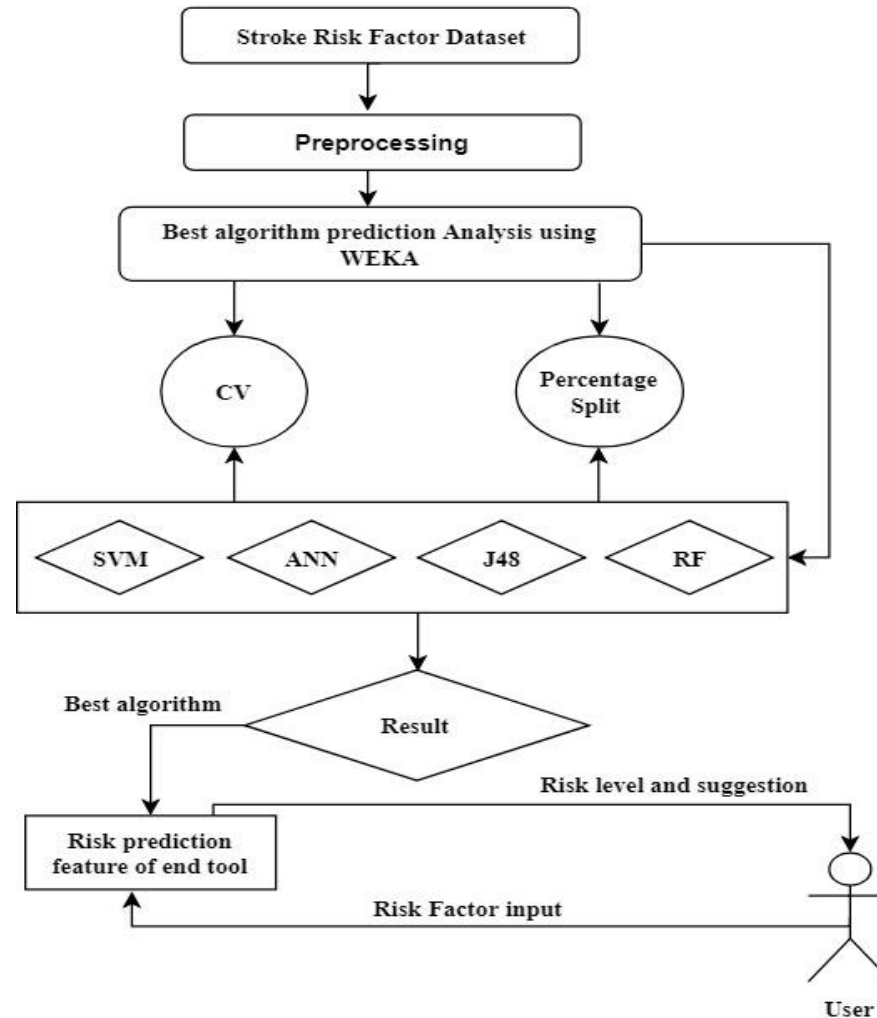
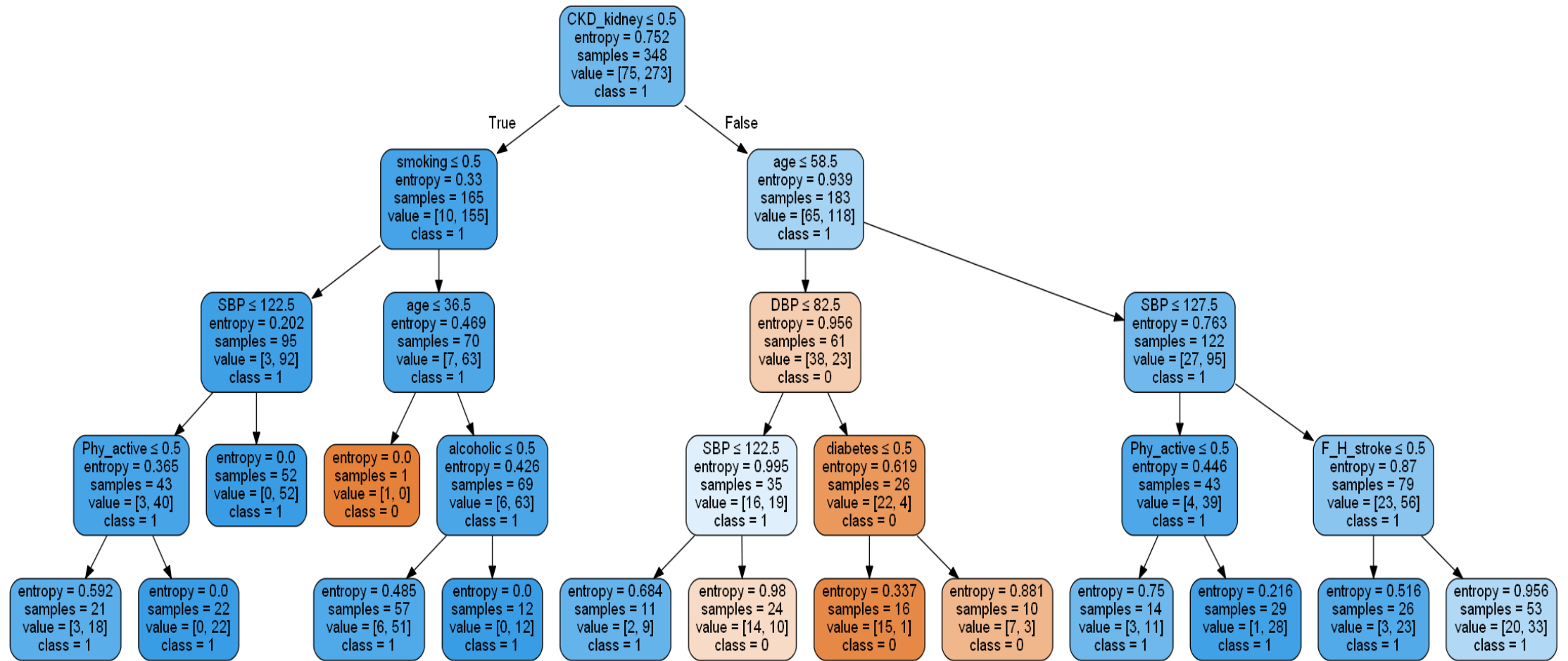


Fig. 2. Proposed System architecture

# Decision Tree



# Used Equations for Decision Tree

## Information Gain

$$\text{Info}(s) = - \sum_{i=0}^k P_i \log_2 P_i$$

Where,  $P_i$  is the probability that an attribute tuple  $S$  belongs to class  $C_i$ .

$$\text{Info}_A(S) = \sum_{j=1}^V \left( \frac{|S_j|}{|S|} * \text{Info}(S_j) \right)$$

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

Where,  $\text{Info}(S)$  is the required information for identifying the class label within  $S$  attribute tuple,  $\text{Info}_A(S)$  is expected information,  $\frac{|S_j|}{|S|}$  is the weight of the  $j^{\text{th}}$  partition and  $i$  is an incremental number. Here, attribute  $A$  will be selected within the highest range of value of  $\text{Gain}(A)$  as the splitting attribute.

# Used Equations for Decision Tree

## Gini Index

$$\text{Gini}(S) = 1 - \sum_{i=1}^m P_i^2$$

Where, S is the sample,  $P_i$  is the probability that a tuple in S belongs to class  $C_i$ .

If a binary split on attribute A partitions the sample or data S into S1 and S2. The Gini Index of S is:

$$\text{Gini}_A(S) = \frac{|S1|}{|S|} \text{Gini}(S1) + \frac{|S2|}{|S|} \text{Gini}(S2)$$

Finding the attribute with minimum Gini index

$$\Delta \text{Gini}(A) = \text{Gini}(S) - \text{Gini}_A(S)$$

Where, A is the attribute with the minimum value of Gini Index is chosen for splitting attribute

# Used Equations for Decision Tree

Gain Ratio:

$$\text{SplitInfo}_A(S) = - \sum_{j=1}^v \frac{|S_j|}{|S|} * \log_2 \frac{|S_j|}{|S|}$$

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}_A(S)}$$

Where,  $\frac{|S_j|}{|S|}$  is weight of the jth partition and the best attribute will be selected with the highest gain ratio.

10 Data

Jupyter screenshot



- The results are retrieved using the data mining tool Weka(3.8) and evaluated with 10-fold Cross validation and Percentage Split method.
- Accuracy table

Algorithms	Accuracy Found with 10-fold cross validation	Accuracy found with Percentage Split Method
Naïve Bayes	79.31%	79.31%
Random forest	82.29%	78.16%
J48 decision tree	83.91%	80.46%
Logistic Regression	80.92%	77.01%
Artificial Neural Network	81.15%	79.31%

# Comparison of evaluation metrics using 10-fold 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	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

# Proposed Tool

## 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 ধরনের সমস্যা আছে কি )

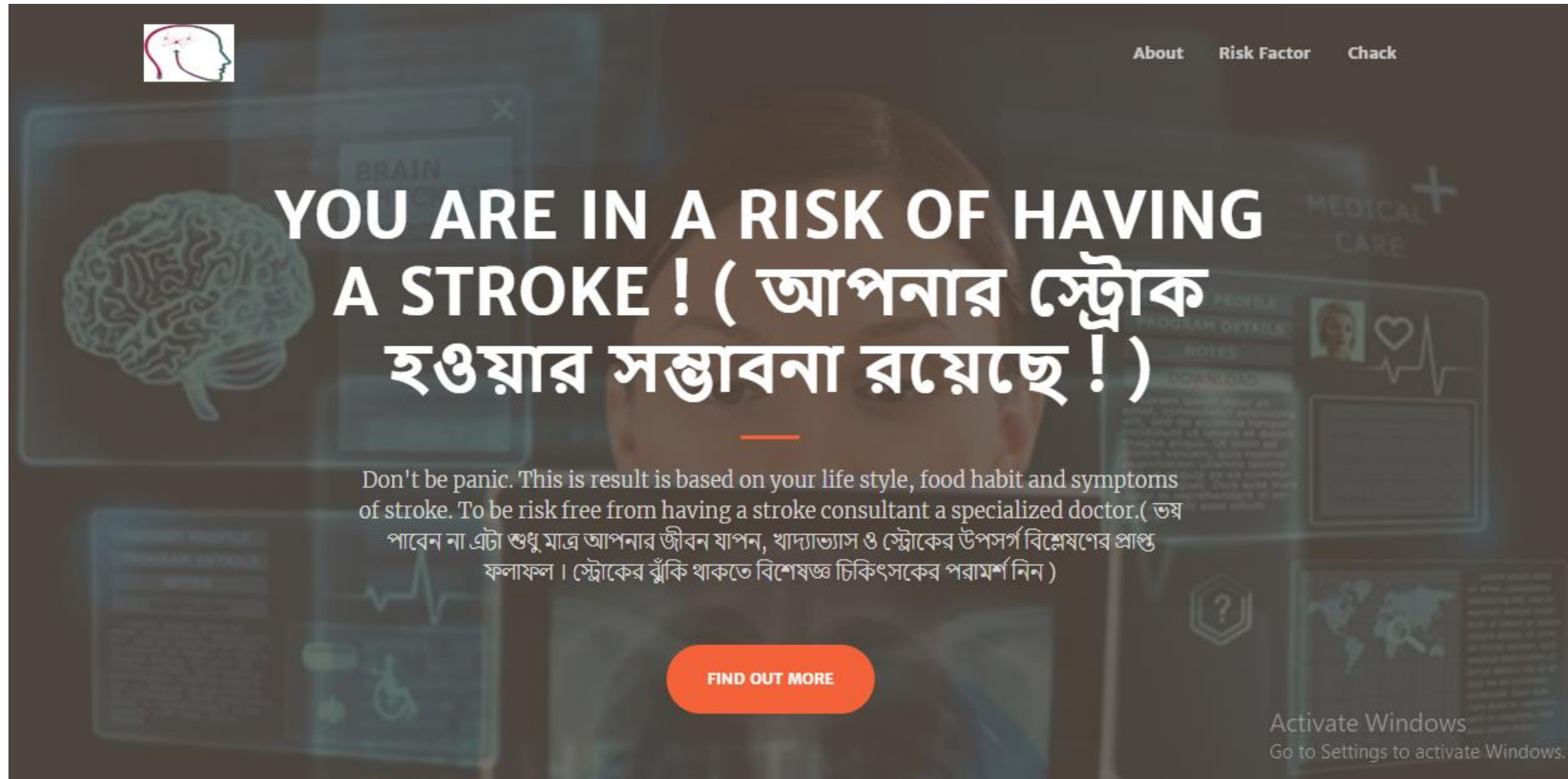
☐ Yes    ☐ No


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

☐ Yes    ☐ No

Alcohol Or Illegal Drug Use (মাদক আসক্তি আছে কি )

# Demo of predicted result Stroke





About Risk Factor Chack

## YOU ARE IN A RISK OF HAVING A STROKE ! ( আপনার স্ট্রোক হওয়ার সম্ভাবনা রয়েছে ! )

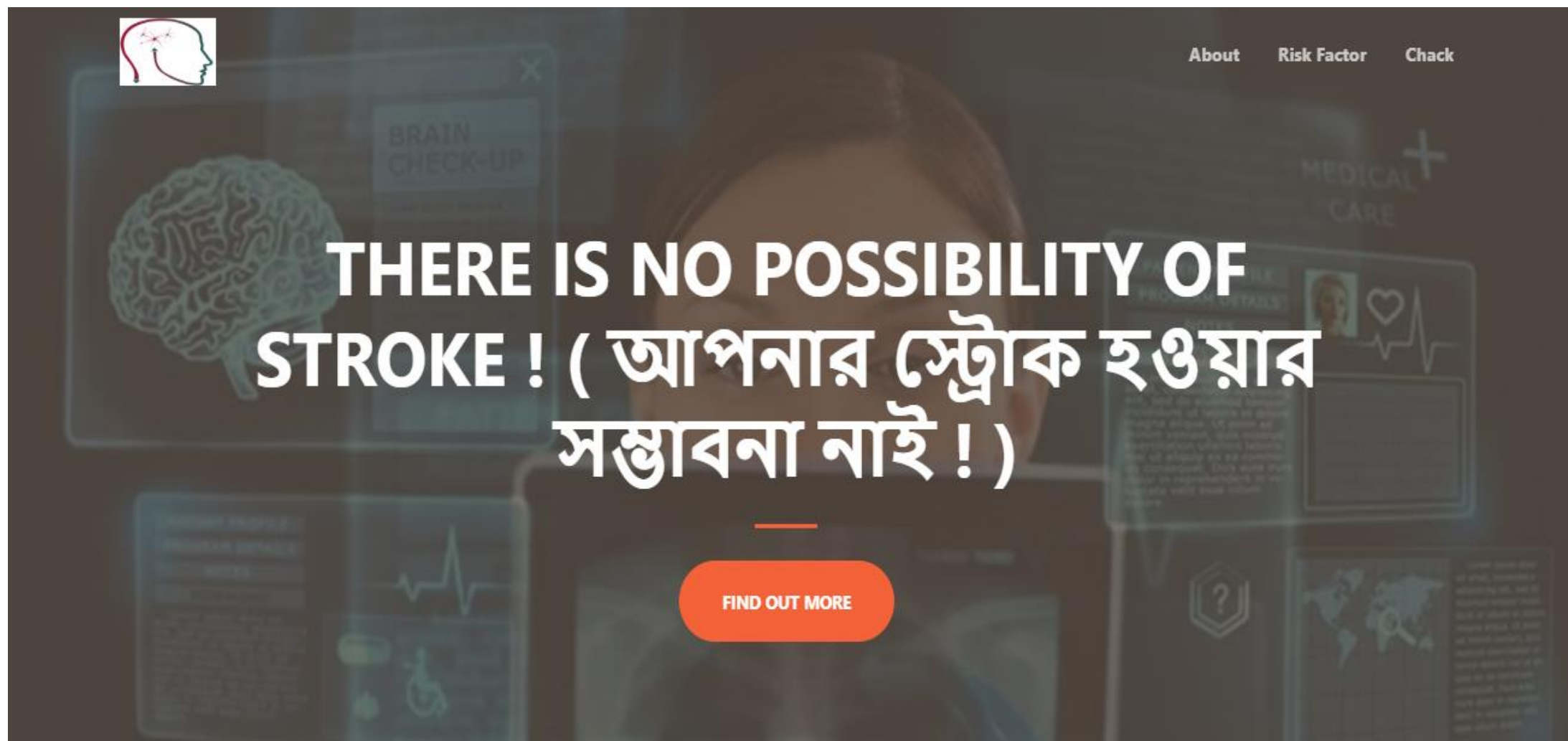
Don't be panic. This is result is based on your life style, food habit and symptoms of stroke. To be risk free from having a stroke consultant a specialized doctor.( ভয় পাবেন না এটা শুধু মাত্র আপনার জীবন যাপন, খাদ্যাভ্যাস ও স্ট্রোকের উপসর্গ বিশ্লেষণের প্রাপ্ত ফলাফল । স্ট্রোকের ঝুঁকি থাকতে বিশেষজ্ঞ চিকিৎসকের পরামর্শ নিন )


[FIND OUT MORE](#)

Activate Windows  
Go to Settings to activate Windows.

# Demo of predicted result Non-stroke

Page: 2





[About](#) [Risk Factor](#) [Check](#)

BRAIN  
CHECK-UP

MEDICAL  
CARE

**THERE IS NO POSSIBILITY OF  
STROKE ! ( আপনার স্ট্রোক হওয়ার  
সম্ভাবনা নাই ! )**

**FIND OUT MORE**

# Challenge

# Limitation

# Conclusion & Future Scope

## ➤ Conclusion:

1. A proper analysis of the relevant dataset to obtain the prediction method of best accuracy a reliable prediction tool is required.

## ➤ Future scope:

1. Deploying a website or mobile apps developed using Decision Tree method for mass people can be considered as the future scope of this work.
2. Regularly updated data from the user in the system will enrich the dataset.
3. Analysis of real system performance can also be conducted.



# References

- [1] National Heart, L. a. B. I., n.d. National Heart, Lung, and Blood Institute. [Online]  
Available at: <http://www.nhibi.nih.gov/health-topics/stroke>  
[Accessed 3 January 2019].
- [2] National Heart, L. a. B. I., n.d. National Heart, Lung, and Blood Institute. [Online]  
Available at: <http://www.nhibi.nih.gov/health-topics/stroke>  
[Accessed 3 January 2019].
- [3] National Heart, L. a. B. I., n.d. National Heart, Lung, and Blood Institute. [Online]  
Available at: <http://www.nhibi.nih.gov/health-topics/stroke>  
[Accessed 3 January 2019].

# Thank You !



*Any Query ?*