Performance Evaluation of Machine Learning Techniques for Early Prediction of Brain Strokes

Md. Azizul Hakim
Student Id: 012182003

A Thesis

in

The Department

of

Computer Science and Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Computer Science and Engineering

United International University

Dhaka, Bangladesh

December 2019

Approval Certificate

This thesis titled " **Performance Evaluation of Machine Learning Techniques for Early Prediction of Brain Strokes**” submitted by **Md. Azizul Hakim**, Student ID: **012182003**, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 14th December 2019.

Board of Examiners

1.

______________________________ Supervisor

Dr. Mohammad Nurul Huda
Professor & Director- MSCSE Program
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

2.

______________________________ Head Examiner

Dr. Swakkhar Shatabda
Associate Professor & Undergraduate Program Coordinator
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

3.

______________________________ Examiner-I

Rubaiya Rahtin Khan
Assistant Professor
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

4.

______________________________ Examiner-II

Suman Ahmmed
Assistant Professor
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

5.

______________________________ Ex-Officio

Dr. Salekul Islam
Professor & Head
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

Declaration

This is to certify that the work entitled “**Performance Evaluation of Machine Learning Techniques for Early Prediction of Brain Strokes**" is the outcome of this thesis carried out by me under the supervision of Dr. Mohammad Nurul Huda, Professor & Director-MSCSE Program.

____

Md. Azizul Hakim,
Id: 012182003
Department of Computer Science and Engineering
United International University (UIU), Dhaka-1212.

In my capacity as supervisor of the candidate’s thesis, I certify that the above statements are true to the best of my knowledge.

____

Dr. Mohammad Nurul Huda
Professor & Director- MSCSE Program
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

Abstract

Brain Strokes are the prime cause of fatality in the world. Bangladesh probably has the highest rates of brain strokes among all South Asian countries and yet is the least studied. Predicting stroke effect from a set of predictive attributes may classify high-risk patients and guide cure approaches, leading to reduce relative incidence. In this work, we propose an intelligent system that can make an effectively prediction of a possible brain strokes using only eight (8) features. We also apply six (06) well known supervised machine learning algorithms on first Bangladeshi datasets collected from five different hospitals of Bangladesh to analyze the prediction accuracy. The overall process can be categorized into four phases. Phase 1: we have provided a comprehensive literature review where we summarize various related machine learning algorithms. Phase 2: we have collected brain stokes patients' data from five different hospitals of Bangladesh to create a dataset. Phase 3: we have selected the important features by using feature importance score. Finally, feed the data to appropriate machine learning algorithms to determine if the predictive model is accurate. It is observed that using our collected dataset for 8 features the classification accuracy of Bagging is almost 96% and it performs better than other classification algorithms such as Logistic Regression (94.82%), k-Nearest Neighbor (73.27%), support vector machines (93.96%), Naive Bayes (93.97%) and Decision Tree (89.66%). Whereas using the dataset with all 23 attributes the classification accuracy of Bagging is 93.43% and it also performs better than the other classification algorithms, such as Logistic Regression (92.24%), k-Nearest Neighbor (69.83%), support vector machines (90.52%), Naive Bayes (92.24%) and Decision Tree (87.07%).

Acknowledgement

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing that makes me possible to complete this thesis successfully. I would like to thank my supervisor Dr. Mohammad Nurul Huda, Professor & Director- MSCSE Program, Department of CSE United International University, Dhaka for his continuous guidance, inspiration, and leniency and for giving me the opportunity to do this work under his supervision. Deep Knowledge & keen interest of my supervisor in the field of "Machine Learning" to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

Finally, again I want to thank all the good wishers, friends, family, seniors for all the help and inspirations. This research is a result of hard work and all those inspirations and assistance.

Table of Contents

LIST OF TABLES..vii

LIST OF FIGURES viii

1. Introduction...1

1.1 Motivation……………………………………………………………………1

1.2 Objectives…………………………………………………………………….1

1.3 Contributions and Publication………..……………………………………2

1.3.1 Contribution……………………………………………………….……2

1.3.2 Publication…………………………………………….……………..2

1.4 Organization of Chapters…………………………………………………….3

2. Background Study and Literature Review...4

2.1 Brain Stroke……………………………………………………………….4

2.2 Machine Learning Overview………………………………………………4

2.2.1 Data Preprocessing………………………..………………………..........….5

2.2.2 Dimension Reduction Techniques………………………..………………..5

2.2.2.1 Feature Extraction……………………………………….…………6

2.2.2.2 Feature Selection………………………………………………..……..6

2.2.3 Classification Algorithms...…...6

2.2.3.1 Decision tree……………………………………………………..……7

2.2.3.2 Support Vector Machine……………………………………..……..7

2.2.3.3 Naïve Bayes………………………………………………..……….8

2.2.3.4 Logistic Regression………………………………………..…………..9

2.2.3.5 K nearest Neighbor………………………………………..………..9

2.2.3.6 Bagging……………………..……………………………………..10

2.4 Related Works…………………………………………………………………10

3. Methodology..14

3.1 Dataset and Features……………………………………………………...14

3.1.1 Data Collection Procedure……………………………………………..…14

3.1.2 Dataset Features…………………………………………………..……15

3.2 Feature Selection Task…………………………………………………...19

3.2.1 Handling Missing Data……………………………………………..……19

3.2.2 Feature Selection Task…………………………………………..……..19

3.3 Proposed Model……………………………………………………….20

4. Results 21

4.1 Experimental Results………………………………………………….21

5. Conclusion 27

5.1 Summary……………………………………………………………………27

5.2 Discussion…………………………………………………………………..27

5.3 Future Work………………………………………………………………...28

5.4 Conclusion………………………………………………………………….28

6. References..29

7. Appendix A..31

LIST OF TABLES

Table 2.1: Various supervised and unsupervised learning algorithms 6

Table 2.2: Summarized Related Works 12

Table 3.1: Dataset Features with Feature Description..15

Table 3.2: Missing Value Statistics of All Features……………………………………...19

Table 4.1: Confusion Matrix of Logistic Regression...23

Table 4.2: Confusion Matrix of k-Nearest Neighbor...23

Table 4.3: Confusion Matrix of Support Vector Machines..23

Table 4.4: Confusion Matrix of Naive Bayes..23

Table 4.5: Confusion Matrix of Bagging...23

Table 4.6: Confusion Matrix of Decision Tree...23

Table 4.7: Performance analysis of various supervised methods using relevant features…………………………………………………………………………………...24

Table 4.8: Performance analysis of various supervised methods using all features…...…..24

Table 4.9: Performance analysis of various supervised methods using irrelevant features…………………………………………………………………………………...24

LIST OF FIGURES

Figure 2.1: Basic Steps of Machine Learning ……………………………………..………......5

Figure 2.2: Decision Tree……………………..……………………...………………………...7

Figure 2.3: Support Vector Machine…………………………………...………….....……….8

Figure 3.1: Dataset Statistics…………………...………………………………………....15

Figure 3.2: Feature Importance Score using Extremely Randomize Tree……..………….20

Figure 3.3: Proposed flow model of research……………………....…………………...21

Figure 4.1: Comparison bar graph of classification accuracy using relevant features………...25

Figure 4.2: Comparison bar graph of classification accuracy using irrelevant features……....25

Figure 4.3: Comparison bar graph of classification accuracy using all features……………...26

Chapter 1

Introduction

According to most recent information published by World Health Organization (WHO), Strokes are the second major reason of death and third major reason of disability [1]. In Bangladesh strokes death achieved 16.27% of total deaths which makes Bangladesh in number 34 position in the world. Stroke is also the first prime cause of fatality in Bangladesh [2]. So an expert system is needed for early prediction and identification of stroke which will be helpful for detecting and medication of stroke.

Machine learning techniques can be used to find out the hidden pattern from patient's dataset that can be used for the development and improvement of an early prediction system that can help doctors is the treatment process. Data science through machine learning algorithm is becoming an essential aid for the diagnosis, treatment and prediction of different kind of diseases.

1.1 Motivation

In Bangladesh 26% of the people aged 40 years or above are suffering from hypertension and of them 21.5% have suffered brain strokes (ref: The Daily Star - May 17, 2018), According to a study, Bangladeshis suffer the highest rate of strokes among 3 south Asian countries (Bangladesh, Sri Lanka and Pakistan). Also strokes death achieved 16.27% of total deaths which makes Bangladesh in number 34 position in the world. It is the high time that the health system of our country must be empowered for further research which may be required to find the factors leading to a higher rate of brain strokes in Bangladesh. So an expert decision support system using machine learning technique will be helpful for early prediction and detection of brain stroke which can reduce the severity.

1.2 Objectives

The motive of my research is to propose a model using machine learning approaches which can be used for early prediction of brain stroke. The objectives of my research are written as follows:

1. Collect real life valid raw data from different hospitals of Bangladesh.
2. Develop a standard model and algorithm that can help us to predict brain stroke in earlier stage.
3. Find out the best feature subset by using feature importance score based on Extra trees algorithm.
4. Analyze the brain strokes data using proposed methods and various machine learning methods and calculate the prediction accuracy.
5. Evaluate the prediction performance of proposed method and compared the performance with other method’s performance by using different performance evaluation matrices.

1.3 Contributions and Publication

1.3.1 Contribution

- We have collected both brain stroke and non-stroke dataset from different hospitals of Bangladesh and constructed a dataset of brain stroke dataset which is first in Bangladesh.
- We have provided a list of related works, where we summarized several works related to Machine Learning algorithms.
- We have reduced the feature by using feature importance score.
- We have proposed an algorithm and compared the performance of that algorithm with various Machine Learning algorithms.

1.3.2 Publication

While doing my M.Sc. thesis, a part of my thesis has been accepted in one international conferences. Publication information of this research work is included in Appendix A.

1.4 Organization of the Chapter

This thesis is organized into 5 chapters. Chapter 2 contains the background and literature review with various literatures findings. Chapter 3 describes the methodology of this research, proposed model, describes the dataset features and data preprocessing techniques. The experimental results are depicting in Chapter 4. Chapter 5 concludes the research with limitation and shows some future work.

Chapter 2

Background Study and Literature Review

2.1 Brain Stroke

Brain stroke is caused by the blockage of blood flow to the brain. This blockage of blood cause oxygen starvation which makes brain impairment and reduce the function of brain. Strokes are divided into three major categories [3]. Those are:

- Ischemic Stroke
- Hemorrhagic Stroke
- Transient ischemic attack

Among these three categories of strokes, Ischemic stroke occurs most often. An ischemic stroke is caused by lack of blood flow to brain tissue. This can happen a condition such as atherosclerosis or an embolism. Hemorrhagic stroke caused by high blood pressure which lead to brain hemorrhage. Transient ischemic attack is known as mini stroke. This types of stroke caused by short time blood flow interruption. Brain strokes can be the reason of permanent damage to the body. Partial damage, speech impairment and also memory loss can occur due to Brain stroke.

2.2 Machine Learning Overview

Machine Learning (ML) is the knowledge and practice of algorithms, mathematical optimization and statistical methods that is concerned with computer programs. In computer science ML is part of Artificial Intelligence (AI), which is used for data analysis, prediction, forecasting, knowledge discovering and improve their performance progressively. Machine Learning techniques construct a mathematical model by various sample data. According to sample data specification machine learning can make a decision with better performance. The sample data is divided into two sectors, namely training data and testing data. These are used in ML methods are delivered from mathematical optimization and is closely related to statistical method. In ML, training details used for learning knowledge discovering and the test data is used to evaluate accuracy and performance of a specific task. The Overview/Model of Machine Learning is shown in Figure 2.1.

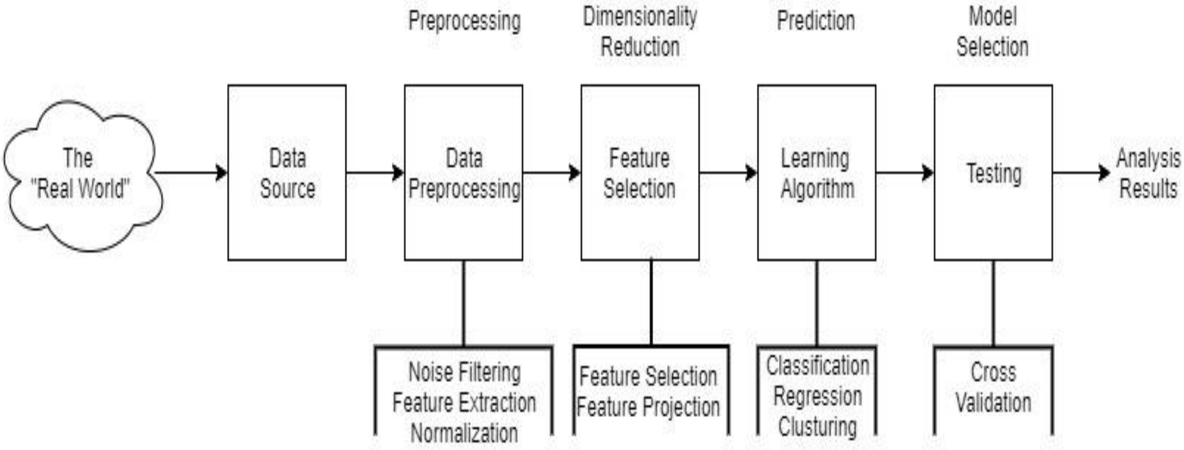

Figure 2.1: Basic Steps of Machine Learning

2.2.1 Data Preprocessing

Data preprocessing is the preliminary step in the machine learning process. Regarding data learning (i.e. data training and data testing) and data analysis it requires a suitable data format. Sometimes data preprocessing is also known as data wrangling (e.g. data in CSV format, noise reduction, normalization, etc.), where data wrangling refers to a set of steps which changes raw data into suitable environmental data format. In this research, we have applied CSV (Comma Separated Value) data format. To obtain better accuracy we have applied mean value imputation for handling missing data.

2.2.2 Dimension reduction techniques

A dimension reduction technique is the method or cognitive operation to eliminate a large number of variables or dimensions without losing information. A huge number of dimensions, duplicate variables, and duplicate or large dataset increase the inconsistency and complexity. To reduce inconsistency, time complexity, computational complexity with better performance including accurate analysis dimensionality retrenchment techniques is required [4]. Dimensionality reduction techniques have two types: Feature Selection and Feature Extraction.

2.2.2.1 Feature Extraction

Feature extraction is a dimensionality reduction process, where a primary set of raw variables is attenuated to more manageable features for processing, while still accurately and completely describing the original data set. The process of converting the input data into the set of features is called feature extraction. Feature extraction is the process of converting the input data into a set of features which can very well represent the input data. It is a process of extracting new features from the original features.

2.2.2.2 Feature Selection

In feature selection, i dimensions of feature are chosen out of j dimensions with important details and eliminate the $(j - i)$ features. The best feature or subset provides the least number of dimensions with low error function and better accuracy [4]. In this research we have used Extra Trees algorithm of finding the best feature of subsets using feature importance score.

2.2.3 Classification Algorithms

Machine Learning method or classification algorithm can produce reliable results and can learn from earlier computation where analyzing huge amount of data or creating predictive models manually could be impossible. There are basically of two types data used in machine learning. Those are: unlabeled data use for unsupervised learning and labeled data use for supervised learning. The various supervised learning and unsupervised learning methods of Machine learning algorithms are given in Table 2.1.

Table 2.1: Various supervised and unsupervised learning algorithms

Machine Learning Algorithms	
Supervised Learning	***Unsupervised Learning***
Decision Tree	K Mean Clustering
Naïve Bayes	Fuzzy C- Mean Clustering
Logistic Regression	Self-organizing Map
Support Vector Machines (SVM)	Hierarchical Clustering
Bagging	
K-Nearest Neighbor (KNN)	

In this research, we have applied well known most popular supervised learning algorithms. Some most popular supervised learning techniques are described below.

2.2.3.1 Decision Tree

Decision tree is a supervised learning classifier and most powerful tools which can be used in discrete/continuous data for prediction or classification. In machine learning decision tree represents tree structure form and classifies the data or instance by beginning at the root of the tree has no incoming edges & moving through it until a leaf node has exactly one incoming edge. Decision tree contains decision nodes, leaf nodes, edges and path. Using this contents decision tree can make a decision by input objects or a set of attributes. Structure diagram of decision tree shown in Figure 2.2.

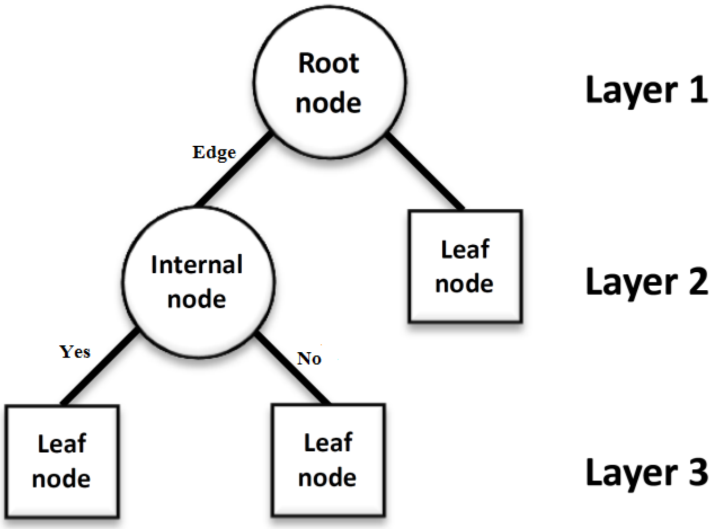

Figure 2.2: Decision Tree

2.2.3.2 Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning technique or discriminative classifier decided by a splitting hyper-plane. This hyper-plane is a line dividing a plane in two parts in two dimensional spaces where in each class retain in either sides and separate the different classes of data. Support vector machine constructed with the training data and it results the hyper-plane in the test data [5]. It tries to find the place in the matrix of data where different classes can be widely separated and draws a hyper-plane.

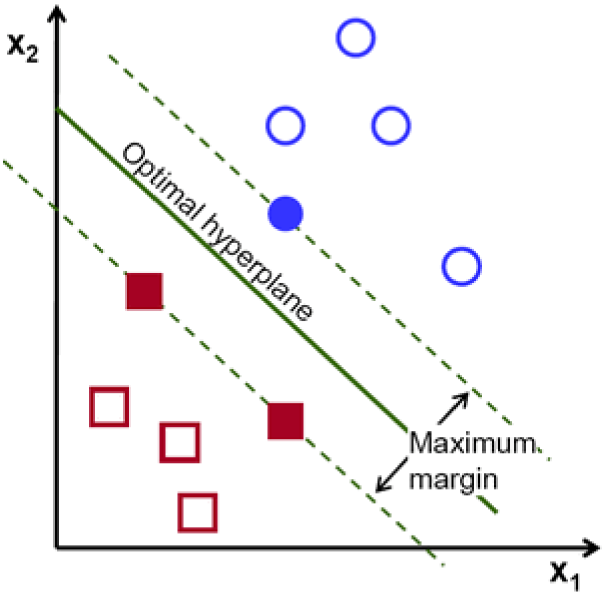

Figure 2.3: Support Vector Machine [6].

In Figure 2.3, the classes of training data points defined by red and blue color where data is labeled. This area has more than one point to draw a hyper-plane for classify the data linearly. A good hyper-plane is taken which increases the margin between the classes.

2.2.3.3 Naïve Bayes

Naïve Bayes is a well-known supervised machine learning algorithm or classifier. To classify the data Naïve Bayes applies the Bayes theorem to classify the data and it assumes that the probability of particular attribute A is fully independent of another attribute B [6]. Bayes theorem gives a theory to compute the probability with prior knowledge of the hypothesis.

$$Posterior = \frac{Likelihood * Prior}{Evidence}$$

where Posterior is the posterior probability of class (target) given predictor (attribute), Likelihood is the probability of predictor given class, Prior -is the prior probability of class, Evidence is the prior probability of predictor.

Naïve Bayes algorithm is divided into three types: Gaussian Naïve Bayes use for classification purposes, Multinomial Naïve Bayes is used in multinomial distributed data problems and Bernoulli Naïve Bayes is used in data with multivariate Bernoulli distribution problem. In this research we have applied Gaussian Naïve Bayes theorem.

2.2.3.4 Logistic Regression

Logistic regression is a statistical technique that analyze a dataset that contains one or more self-governing variables that decide the result. The result is calculated with a variable where there are two possible results. The objective of logistic regression is to identify the best suitable model that can narrate the connection between the variables [7]. Logistic regression creates the coefficients of a formula that can predict a logarithmic transformation of the probability of presence of the characteristic of interest.

$$\boldsymbol{logint}\ (\boldsymbol{p}) = a_0 + \ a_1 X_1 + \ a_2 X_2 + \ a_3 X_3 + \cdots + \ a_i X_i$$

where p is the probability of presence of the characteristic of interest. The $logint$ is defined as follows:

$$\boldsymbol{logint}(\boldsymbol{p}) = \boldsymbol{ln}\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right)$$

Where p is the probability of presence of the characteristic of interest and $(1-p)$ is the probability of absence of the characteristic of interest

2.2.3.5 K Nearest Neighbor

K nearest neighbors which works with distance function. By using a majority vote of its neighbor, a class is classified. Here mainly Euclidean, manhattan, minkowski and hamming distance formula are used for calculating the distance of nearest neighbor. We can implement the K Nearest Neighbor model according to following steps [8]:

1. Load the collected dataset
2. Initialize the value of k
3. For $n = 1$ to total number of training data do:
    - Calculate the distance between test data and each row of training data by using any distance measuring formula
    - Sort the calculated distances in ascending order based on distance values measured in previous step
    - Take the top k rows from the sorted array
    - Identify the most frequent class of those rows
    - Return the predicted class

2.2.3.6 Bagging

Bagging techniques mainly create to key elements. One is bootstrap and another is aggregation. At first some subsets are created from the original training set which is bootstrap technique [9]. Then bagging techniques aggregate the output of all base models into one output using voting method [10].

Given training dataset $Q = \{(x_1, y_1),\ (x_2, y_2) \dots (x_M, y_M)\}$, where $y_M = \{0, 1\}$ corresponds to the class types (Non-stroke and Stroke). BS is the number of base models or subsets and BC is the base classifier.

1. Generate N number of bootstrap datasets as $Q_n = \text{Bootstrap}(Q)$, where $n = 1, 2, \dots, BS$
2. Initialize the base model set $BS = 0$
3. Create base model $b_{mn} = (Q_n)$, where $n = 1, 2, \dots, BS$ by applying the base classifier BC on each bootstrap
4. For $n = 1, 2, \dots, BS$ do:

$$BS = BS \cup \{b_{mn}\}$$

5. Final result is given by bagging method is:

$F_{Bagging_{(x)}} =$ mostly voted class in $V_{m_{(x)}}$, where $V_m \in BS$ and x is the predicted test pattern.

2.3 Related Works

Machine learning approaches are widely being used for early prediction and identification of different types of disease. Several researchers are working hard to make co-relation between the medical data and machine learning techniques. The details of related works are described below.

Khosla et al. [11] compared the cox proportional hazards model with machine learning approaches for the prediction of stroke on the Cardiovascular Health Study (CHS) dataset. They propose an automatic feature selection algorithm named conservative mean (CM). They used support vector machine (SVM) and margin-based censored regression (MCR) learning algorithm for stoke prediction. By combining CM feature selection with MCR they got 0.777 average tests AUC to predict stroke which is the highest.

N. Kasabov et al. [12] a novel approach named PMeSNNr (Personalised modelling evolving spiking neural network reservoir system) for individualized modeling of spatio/spectro-temporal data (SSTD) and to predict events. This classification model is based on spiking neural networks (SNN) which is suitable for learning and classification of SSTD. They compared this model with traditional machine learning approaches multi linear regression (MLR), support vector machine (SVM) and multilayer perception (MLP). They achieved the best result with 94% accuracy by using PMeSNNr method.

P. Bentley et al. [13] used machine learning approach for predicting stroke thrombolysis outcome using imaging features based on computerized tomography (CT). They collected clinical domain records and CT brains of 116 acute ischemic stroke patients those who are treated with intravenous thrombolysis. They compared the performance of support vector machine (SVM) with other prognostication tools such as SEDAN and HAT scores and SVM achieved the best result with 0.744 AUC score.

O. Almadani et al. [14] designed a framework based on different data mining techniques to predict stroke. They have divided their dataset into two classes. One class includes the data those who have stroke and another class includes the stroke mimic patients. Their dataset consists of 969 instances. Among them 899 instances are stoke patients and 69 instances are stroke mimic patients. They used principal component analysis (PCA) as attribute selection technique. J48, JRip and multilayer perception (MLP) were applied by using 10-fold cross validation method on the data set for building a model. They got their best accuracy of 95.25% by using C4.5 along with PCA.

B. Letham et al. [15] proposed a model called Bayesian Rule Lists (BRL), a rule based model that produces a posterior distribution over permutations of if….then rules for predicting stroke. They compared this model with other models such as $CHADS_{2,}$ $CHADS_2$-VASc, CART, C5.0, l_1 logistic regression, SVM, Random forests. Among them BRL is the best performing method with AUC score of 0. 775.They also said that BRL method matched random forests method for the best performing method. They also measured the performance by applying the models in to two subsets of the data one is female patient's subset and another is male patient's subset. Here BRL models again outperformed the other models and it matched with the result of random forests method.

D. Shanthi et al. [16] applied Artificial Neural Networks (ANN) for predicting Thrombo-embolic stroke disease. They used the clinical dataset of only 50 patients those who have the symptoms of stroke. For feature selection they used backward stepwise method. They got 89% prediction accuracy by using ANN based prediction model. They also said that ANN based prediction of stroke disease improves the treatment accuracy with higher consistency.

Ahmet K. Arslan et al. [17] used support vector machine (SVM), stochastic gradient boosting (SGB) and penalized logistic regression (PLR) techniques to predict ischemic stroke. They applied these techniques on medical dataset which includes the medical records of only 80 patients and 112 healthy persons with 17 attributes. They utilized 10-fold cross validation method and their performance evaluation metrics includes accuracy, AUC, sensitivity, specificity, positive predictive value and negative predictive value. For tuning the parameters of the model grid search method was applied. In case of accuracy the SVM model was the best predictor with 96% of accuracy.

Sung S-F et al. [18] proposed a method for developing a stroke severity index (SSI) by using administrative data of patients with acute ischemic stroke. They said stroke severity was measured by using National Institute of Health Stroke Scale (NIHSS). They developed three models by using *k*-nearest neighbor, multiple linear regression and regression tree and measure the model performance according to the pearson correlation coefficient between SSI and the NIHSS. They found that the *k*-nearest neighbor performs well than the others with correlation coefficient of 0.743.

In table 2.2 we have summarized the related works

Table 2.2: Summarized Related Works

Author's Name	**Working Principle**	**Used ML Algorithm**	**Performance**	**Year**
Ahmet K. Arslan et al.	Predicting ischemic stroke using medical dataset includes 192 records	Support Vector Machine (SVM), Stochastic Gradient Boosting (SGB) and Penalized Logistic Regression (PLR) .	96%	2016
Sung SF et al.	Proposed a method for creating a Stroke	*k*-Nearest Neighbor, Multiple Linear	0.743 (AUC Score)	2015

	Severity Index (SSI) and measure the performance of model according to the pearson correlation coefficient between SSI and the NIHSS.	Regression and Regression Tree		
P. Bentley et al.	Predicting stroke thrombolysis outcome using imaging features based upon Computerized Tomography (CT).	Support Vector Machine (SVM)	0.744 (AUC Score)	2014
A. Khosla et al.	Comparing between the cox proportional hazards model and machine learning approaches for the prediction of stroke and proposed an automatic feature selection algorithm named Conservative Mean	Support Vector Machine (SVM) and Margin-based Censored Regression (MCR)	0.777 (AUC Score)	2010
D. Shanthi et al.	Predicting Thrombo-embolic stroke disease	Artificial Neural Networks (ANN)	89%	2009

According to the review of earlier research, it observes that different researchers proposed different approaches to predict strokes in earlier stage but no one used ensemble based machine learning techniques before. In Bangladesh no one works before with stoke dataset. That’s why are highly motivated to develop a suitable model for early prediction of stroke with higher accuracy.

Chapter 3

Methodology

In our research, we have tried to collect data from different medical hospital of Bangladesh. We have also tried to make our research unique and make accurate prediction in our research. After collecting data, we got some missing value which we resolved. To get the proper prediction, we've completed the feature scaling process. Datasets are used for training and testing purposes and here are some of them included algorithm Logistic Regression (LR), k-Nearest Neighbor (KNN), support vector machines (SVM), Naive Bayes and Bagging. The appropriate algorithm scenario has been given based on the working procedure.

3.1 Dataset and Features

3.1.1 Data Collection Procedure

Data used in this research are collected from various medical institutions in our country (Bangladesh). Most of the data are collected from Bangabandhu Sheikh Mujib Medical University (BSMMU) and National Institute of Neurosciences & Hospital (NINS) of Dhaka Bangladesh. Rest of the data are collected from Dhaka Medical College and Hospital (DMC), Khulna Medical College and Hospital (KMC) and Jhenaidah Sadar Hospital, Jhenaidah. We have collected the data of a brain stroke patient after he/she get attacked with brain stroke and get admitted into the mentioned hospitals. All these data are collected manually from medical register book. This data set consists of 385 instances, where number of males are 209 (54%) and the number of females are 176 (46%). Our data set also contains 234 (61%) brain stroke patient's data and 151(39%) non-brain stroke people's data. Figure 3.1 shows the percentage of gender and final result of our dataset. From bar chart we can say that males face brain stroke disease more than the females in Bangladesh.

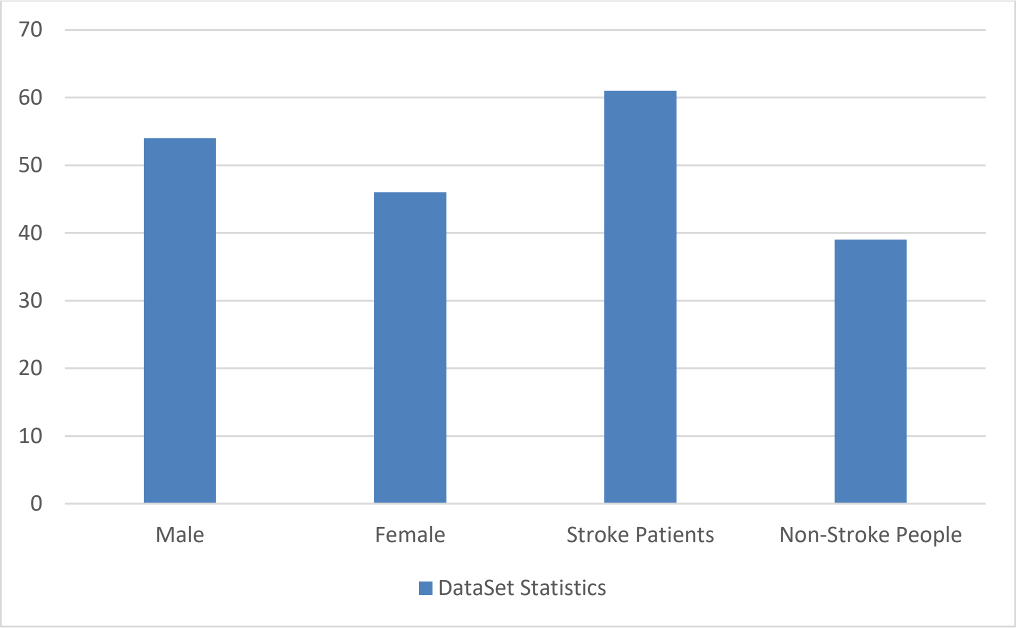

Figure 3.1: Dataset Statistics

3.1.2 Dataset Features

The obtained data set contains 23 features and a class variable. The main features are (Gender, Age, Work Type, Residence Area, Marital Status, Hypertension, Heart Disease, Smoking Habit, Average Glucose Level, Weight, Height, BMI, RBS, Serum Creatinine, Serum Cholesterol, LDL, HDL, Triglyceried, HbA1c, Hb, RBC, WBC, ESR). Table 3.1 shows the features and features descriptions of our research.

Table 3.1: Dataset Features with Feature Description

No.	Feature Name	Feature Description
1	Gender	Male: 1 Female: 0
2	Age	Age in years
3	Work Type	Govt. job: 1 Private job: 2 Farmer: 3 Business: 4

		Politics: 5 Teacher: 6 None: 7 Self Employed: 8
4	Residence Area	Rural: 1 Urban: 0
5	Marital Status	Yes: 1 No: 0
6	Hypertension	Yes: 1 No: 0
7	Heart Disease	Yes: 1 No: 0
8	Smoking Habit	Yes: 1 No: 0
9	Average Glucose Level	>6.0 mmol/L
10	Weight	Weight in Kilograms
11	Height	Height in Meters
12	BMI	Kg/m^2
13	RBS	<7.8mmol/L
14	Serum Creatinine	<1.2mg/dL
15	Serum Cholesterol	150-220m
16	LDL	<130 mg/dL
17	HDL	>40 mg/dL
18	Triglyceried	<150mg/dl
19	HbA1c	4-5.6%
20	Hb	12.5 – 17.5 g/dL
21	RBC	Male:4.7- 6.1 mcL, Female:4.2 – 5.4 mcL
22	WBC	4000 – 11000 $cell/mm^3$
23	ESR	0 to 29 mm/hr
24	Class: Stroke	Yes: 1 No: 0

Gender: Men are generally at greater risk of brain stroke. However, women's risk increases after a certain age. In our total collected 385 instances total male are 209 (54%) and females are 176 (46%).

Age: Aging increases the risk of brain strokes. Around 75% of all strokes occur in the people over the age of 65 and the strokes risk are double after the age of 55. [19]

Hypertension: High blood pressure can lead to brain stroke by damaging the brain's blood vessels. It is also an important factor for causing brain stroke.

Heart Disease: People who are suffering with heart disease, angina or those who have had a heart attack due to hardening of the arteries which called atherosclerosis have more risk of stroke.

Smoking Habit: Smoking increase the occurrences of cerebrovascular disease which can causes higher risk of stroke. Smoking is linked with the risk of different ischemic stroke and arachnoid hemorrhage which can cause brain stroke.

Average Glucose Level: The average glucose level of a normal people is less than 00 mg/dL. But when the average glucose level is greater than 108 mg/dL (>6.0 mmol/L) then it can cause hyperglycemia. And hyperglycemia can lead to ischemic stroke.

Weight: Stroke patients those who are underweight 67% more likely to die than patients those who are normal weight.

RBS: Random blood sugar (RBS) is a measures of blood glucose. The ideal value is less than 7.8mmol/L. RBS value, greater than 7.8mmol/L can cause higher risk of brain strokes.

Serum Creatinine: Normal creatinine level in men varies from 0.9 to 1.3 mg/dL and for women it varies from 0.6 to 1.1 mg/dL. People with higher serum creatinine level in the blood indicates that the kidneys are not working properly which is a potential risk factor for stroke. People who are suffering with chronic kidney diseases are at higher risk of brain strokes.

Serum Cholesterol: Total serum cholesterol level of an adults is less than 200 mg/dL. Cholesterol level greater than 239 mg/dL is considered as a high level. Cholesterol is a waxy substance that builds up in artery walls and contributes to the formation of plaque, which refers to narrowing the arteries. This makes it difficult for blood to be pumped effectively from the heart to the rest of the body, setting the stage for a possible future stroke.

LDL: LDL cholesterol (bad cholesterol) alone is a relatively poor predictor of risk. In biomedical report or blood test report we have found an optimal level of LDL cholesterol is <130 milligrams per deciliter (mg/dL), or 3.37 millimoles per liter (mmol/L).

HDL: High-density lipoprotein (HDL) good cholesterol, which helps remove cholesterol from the arteries and prevent fatty buildup, and various non-HDL lipoproteins, which in excess are linked to artery damage, heart disease and stroke. In biomedical report or blood test report we have found an optimal level of HDL is >40 mg/dL (1.3 mmol/L).

Triglyceride: Triglycerides are one kind of fat which is found in blood. Higher triglyceride level can block the proper blood flow. The normal value of triglyceride of a human body is less than 150 mg/dL or less than 1.7 mmol/L.

HbA1c: HbA1c is the glycated hemoglobin level. Higher HbA1c indicated the higher risk for ischemic stroke. HbA1c value greater than 5.6% can cause the future risk of stroke. Normal value of HbA1c is less than 5.6%

Hb: High hemoglobin level makes too many red blood cells which makes the blood thicker than usual. This can lead to stokes. Normal hemoglobin level for men varies from 13.5 to 17.5 g/dL and for women it varies from 12.5 to 15.5 g/dL.

WBC: WBC is the white blood cells distribution in blood. Abnormal WBC is lined with high mortality rate after ischemic stroke. The normal range of WBC is between 4,000 and 11,000 per microliter of blood.

ESR: High Erythocyte sedimentation rate is considered to be a cause of ischemic stroke. The normal range is 0 to 22 mm/hr for men and 0 to 29 mm/hr for women.

RBC: Red blood count is the red blood cell count in blood. High red blood cells makes the blood thicker. This can cause ischemic stokes. The normal range of RBC for men is 4.7 to 6.1 mcL (million cells per microliter) and for women it is 4.2 to 5.4 mcL (million cells per microliter)

3.2 Data Preprocessing

3.2.1 Handling Missing Data

Clinical domain data has several delusions due to the process of collecting data. Missing entries can lead to an improper predictive model. Data imputation can be helpful to amend missing entries. There are several different techniques of missing data imputation such as Mean Imputation, Regression Imputation, Hot-deck Imputation, Expectation Maximization (EM) [20] and so on. We used Mean Imputation techniques because this is a widely used imputation techniques and it is fast, simple and ease to implement. Table 3.2 shows the missing value statistics of 22 features.

Table 3.2: Missing Value Statistics of All Features

Feature Name	Number of Missing Values (Percentage %)
Gender	0%
Age	0%
Work Type	0%
Residence Area	0%
Marital Status	0%
Hypertension	0%
Heart Disease	0%
Smoking Habit	0%
Average Glucose Level	17%
Weight	0%
Height	0%

BMI	0%
RBS	25%
Serum Creatinine	19%
Serum Cholesterol	25%
LDL	14%
HDL	24%
Triglyceried	22%
HbA1c	23%
Hb	2%
RBC	13%
WBC	1%
ESR	0%

3.2.2 Feature Selection Task

Due the irrelevant features in the dataset under fitting and overfitting problems on the classification model can occurs. It can also decrease the training period and performance on the test set. Extremely Randomize Trees (Extra Trees) is executed for choosing the subset of significant features and separate unimportant features by using feature significance result. Figure 3.2 shows the feature importance score. Extra tress contains lots of decision trees which randomly select the observations from the training dataset and features. Not every tree has all the features or all the observations which guarantees that the trees are de-correlated which makes it less prone to over-fitting. We choose 70% data as a training dataset. In our work, we choose the first eight features according to their feature importance score. Feature importance score of the eight selected features are Hypertension, Heart disease, HDL, LDL, Triglyceride, RBS, HbA1c and Age with 0.53, 0.061, 0.421, 0.03, 0.031, 0.022, 0.024 and 0.023 respectively. We also showed the comparison between the accuracy with selected features, irrelevant features and all features.

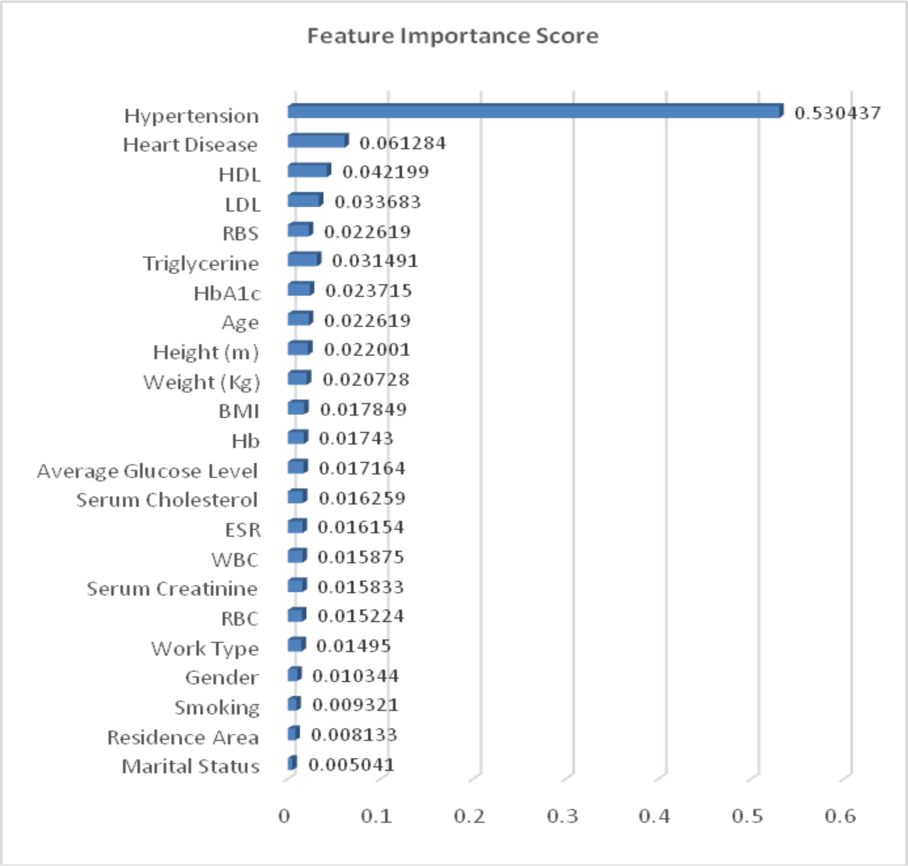

Figure 3.2: Feature Importance Score using Extremely Randomize Tree.

3.3 Proposed Model

The proposed model (see the Figure 3.3) of this work described below:

- Collect actual valid raw survey data from different hospitals of Bangladesh.
- Select Important features using feature importance score.
- Analyze the stroke data using various machine learning methods/data mining techniques for better prediction and accuracy.
- Analyze performance accuracy using different performance evaluation matrix.
- Validate the prediction result by medical professionals.

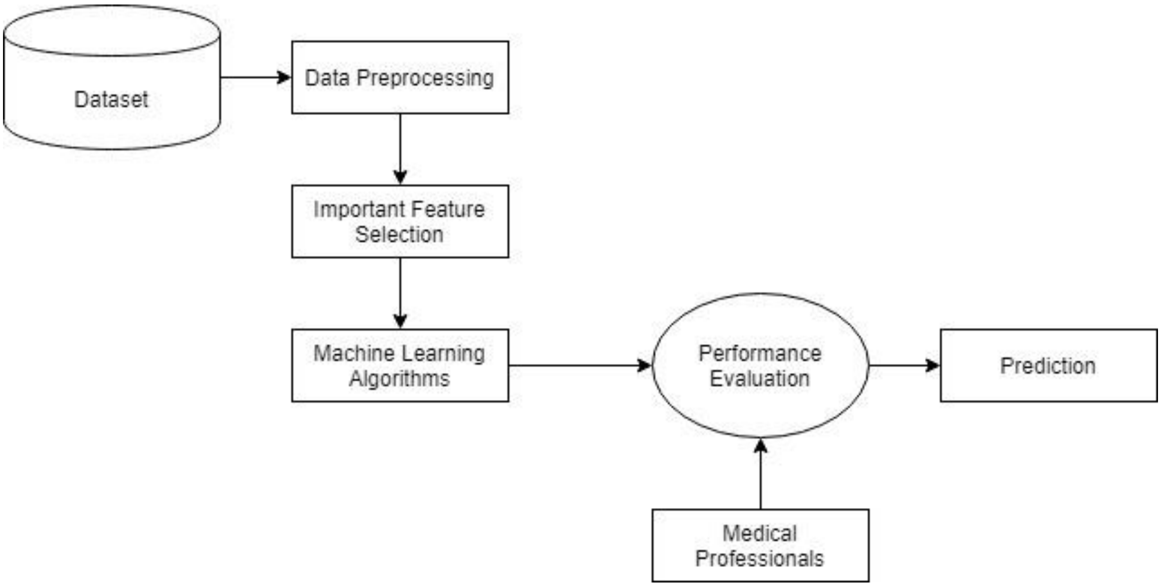

Figure 3.3: Proposed flow model of research

Chapter 4

Results

4.1 Experimental Results

We have evaluated this classification and prediction model by using four evaluation metrics. Stroke patients and non-stroke persons are labeled as the presence and the absence class respectively. Here, Performance evaluation metrics are demonstrated as follows:

- $\text{Accuracy} = \frac{(T_P+T_N)}{(T_P+F_N+F_P+T_N)}$
- $\text{Recall} = \frac{T_P}{(T_P+F_N)}$
- $\text{Precision} = \frac{T_P}{(T_P+F_P)}$
- $\text{F1} - \text{Score} = \frac{2\times(\text{Precision} \times \text{Sensitivity})}{(\text{Precision}+ \text{Sensitivity})}$

Where, T_P, F_P, F_N and T_N represents True Positive, False Positive, False Negative and True Negative respectively.

We have trained six machine learning classifiers namely Logistic Regression (LR), K-Nearest Neighbor (KNN), Support vector machines (SVM), Decision Tree, Naive Bayes and Bagging with eight selective features discussed in feature selection task section. All the classification tasks were done by executing the python. We have used 70% of data for training purpose and 30% of data for testing purpose. We have also used 10-fold cross validation to estimate accuracy.

It is observed that using 385 data with 8 selective features, the classification accuracy of bagging is almost 96% and it performs better than other classification algorithm such as Logistic Regression (94.82%), k-Nearest Neighbor (73.27%), support vector machines (93.96%), Decision Tree (89.66%), and Naive Bayes (93.97%). The confusion matrix of Logistic Regression (LR), k-Nearest Neighbor (KNN), support vector machines (SVM), Naive Bayes, Decision Tree and Bagging are showed in Table 4.1 to Table 4.6 respectively. The experimental results are displayed in Table 4.7 to Table 4.9 respectively. Figure 4.1 shows the comparison bar graph of classification accuracy using eight relevant features.

Figure 4.2 shows the comparison bar graph of classification accuracy using irrelevant features and Figure 4.3 shows the comparison bar graph of classification accuracy using all 23 features.

Table 4.1: Confusion Matrix of Logistic Regression

Logistic Regression			
Stroke Disease		Predicted Class	
		Presence	Absence
Actual Class	Presence	44	3
	Absence	3	66

Table 4.2: Confusion Matrix of k-Nearest Neighbor

k-Nearest Neighbor			
Stroke Disease		Predicted Class	
		Presence	Absence
Actual Class	Presence	28	14
	Absence	17	57

Table 4.3: Confusion Matrix of Support Vector Machines

Support Vector Machines			
Stroke Disease		Predicted Class	
		Presence	Absence
Actual Class	Presence	50	5
	Absence	2	59

Table 4.4: Confusion Matrix of Naive Bayes

Naive Bayes			
Stroke Disease		Predicted Class	
		Presence	Absence
Actual Class	Presence	37	3
	Absence	4	72

Table 4.5: Confusion Matrix of Bagging

Bagging			
Stroke Disease		Predicted Class	
		Presence	Absence
Actual Class	Presence	68	3
	Absence	2	43

Table 4.6: Confusion Matrix of Decision Tree

Decision Tree			
Stroke Disease		Predicted Class	
		Presence	Absence
Actual Class	Presence	36	5
	Absence	7	68

Table 4.7: Performance analysis of various supervised methods using relevant features

Techniques/ Methods	Accuracy	Recall	Precision	F1-Score
Logistic Regression	94.82%	95.65%	95.65%	95.65%
k-Nearest Neighbor	73.27%	77.02%	80.28%	78.62%
Support Vector Machines	93.96%	96.72%	92.18%	94.40%
Naive Bayes	93.97%	94.74%	96.00%	95.36%
Bagging	95.69%	97.14%	95.77%	96.45%
Decision Tree	89.66%	90.67%	93.15%	91.89%

Table 4.8: Performance analysis of various supervised methods using all features

Techniques/ Methods	Accuracy	Recall	Precision	F1-Score
Logistic Regression	92.24%	92.75%	94.12%	93.43%
k-Nearest Neighbor	69.83%	70.77%	74.19%	72.44%
Support Vector Machines	90.52%	89.71%	93.85%	91.72%
Naive Bayes	92.24%	92.10%	95.89%	93.95%
Bagging	93.43%	92.89%	95.97%	94.02%
Decision Tree	87.07%	86.30%	92.65%	89.36%

Table 4.9: Performance analysis of various supervised methods using irrelevant features

Techniques/ Methods	Accuracy	Recall	Precision	F1-Score
Logistic Regression	58.62%	84.05%	61.05%	70.73%
k-Nearest Neighbor	52.58%	61.64%	62.50%	62.06%
Support Vector Machines	64.65%	01.00%	64.65%	78.53%
Naive Bayes	62.93%	94.73%	64.86%	77.05%
Bagging	65.23%	67.10%	65.19%	66.36%
Decision Tree	52.58%	57.14%	61.53%	59.25%

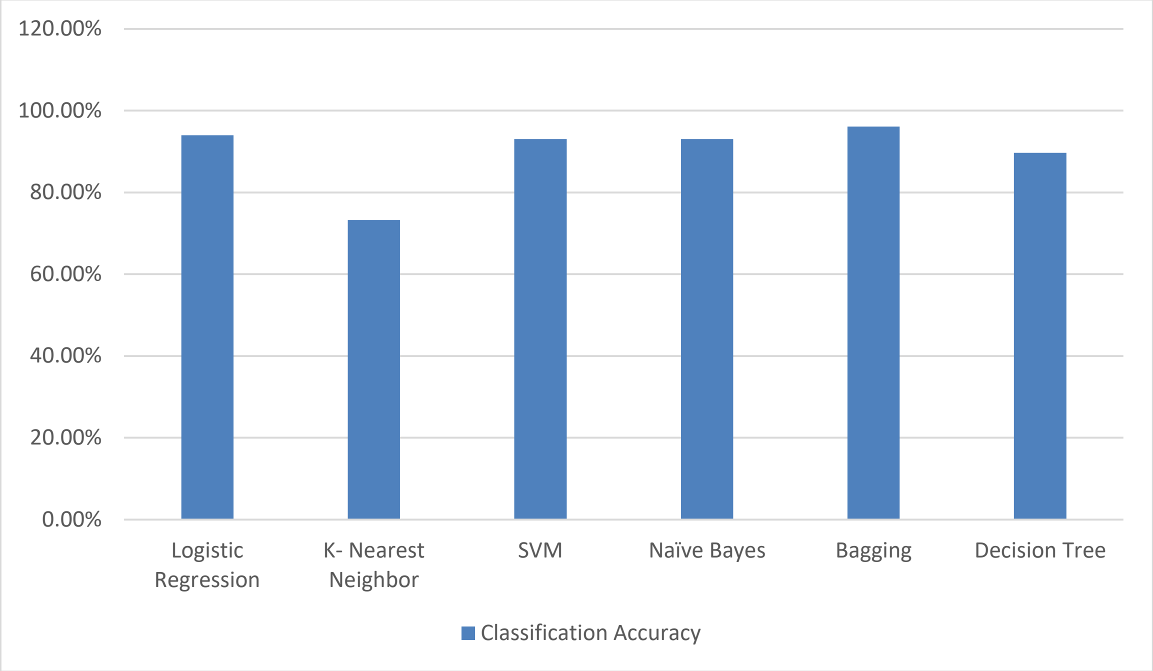

Figure 4.1: Comparison bar graph of classification accuracy using relevant features

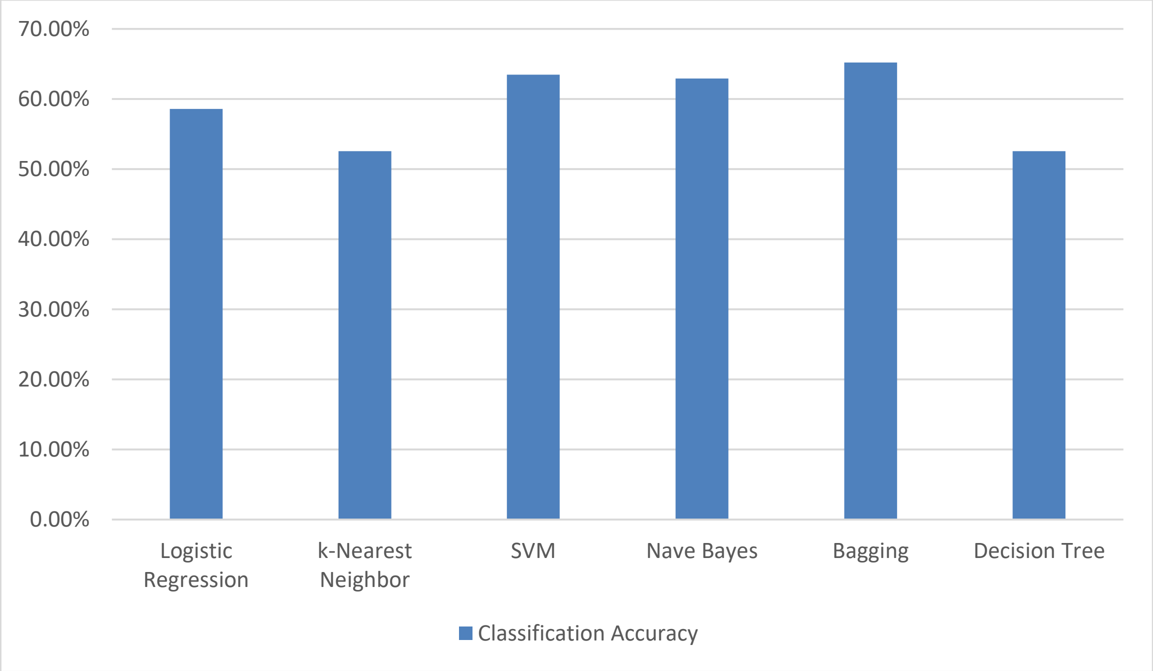

Figure 4.2: Comparison bar graph of classification accuracy using irrelevant features

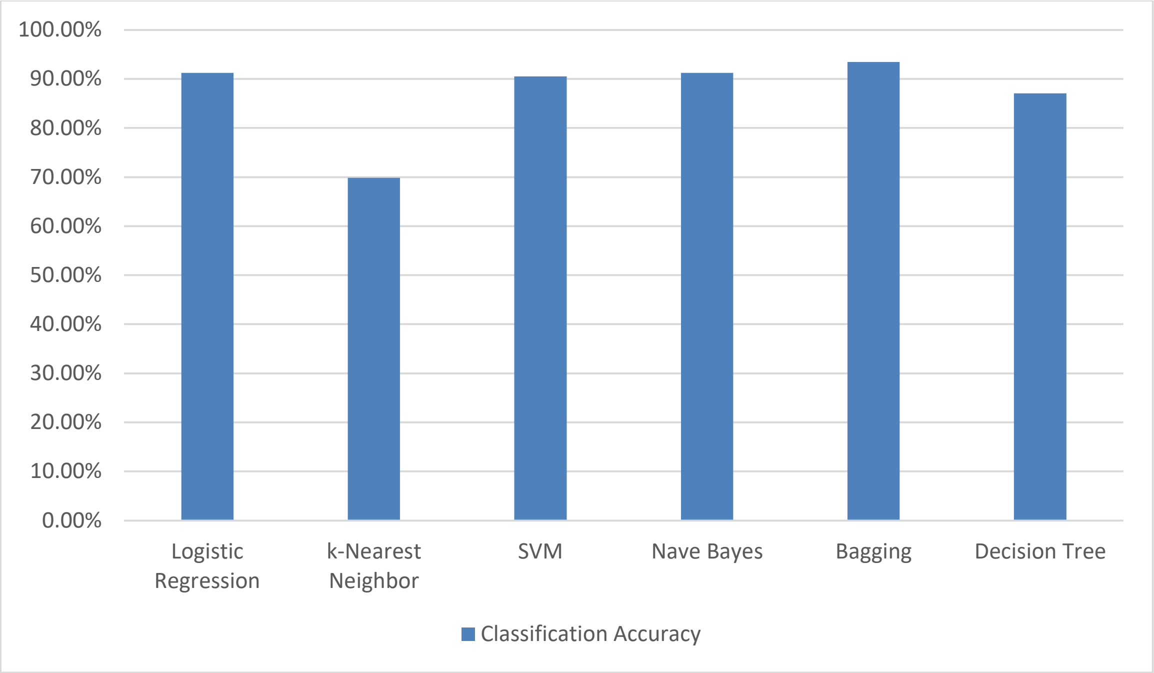

Figure 4.3: Comparison bar graph of classification accuracy using all features

Chapter 5

Conclusion

5.1 Summary

In this work we have collected data from five different hospitals of Bangladesh and developed a brain stroke prediction model to assists doctors in predicting brain stroke based on a patients’ clinical data. We have collected total 385 instances with 23 features and developed a prediction model using six supervised machine learning algorithms. It is observed that Bagging algorithm is the best predictor with almost 96% accuracy.

5.2 Discussion

In this thesis we have applied different machine learning algorithms for all 23 features, 8 selective relevant features and 8 irrelevant features. We have three investigations for machine learning classifiers. One is 385 patient’s data with all 23 features, the classification accuracy of Logistic Regression (92.24%), k-Nearest Neighbor (69.83%), support vector machines (90.52%), Naive Bayes (92.24%), Bagging (93.43%) and Decision Tree (87.07%). Here Bagging algorithm gives better result than other.

The second investigation is 385 patient’s data with 8 irrelevant features selected by features important score discussed in feature selection task section. The classification accuracy of this investigations are Logistic Regression (58.62%), k-Nearest Neighbor (52.58%), support vector machines (64.65%), Naive Bayes (62.93%), Bagging (65.23%) and Decision Tree (52.58%). Here the classification accuracies are very low compare than the first investigation. But in this case also the bagging algorithm gives the better accuracy than others.

Third investigation is 385 data with 8 relevant features selected by features important score discussed in feature selection task section. The classification accuracy of Bagging algorithm is 95.69% and it performs better than other classification algorithms such as Logistic Regression (94.82%), k-Nearest Neighbor (73.27%), support vector machines (93.96%), Naive Bayes (93.97%) and Decision Tree (89.66%).

According to these three investigations we can say that among those 23 features, 8 features have better importance scores than others. While predicting brain strokes we can use only these eight selected features instead of using all 23 features. If we use all features, then a person needs to do 12 medical tests and on the other hand if we use eight relevant features then a person needs to do only 5 medical tests. It will also help to minimize the cost of medical tests [21].

5.3 Future Work

In future we would like to concentrate on two things:

- Prepare a repository database based on stoke patients of Bangladesh
- Developed a mobile based stroke prediction and suggestions application through internet.

5.4 Conclusion

On the whole the motive of our work is to predict brain stroke in earlier stage and build a dataset of our own which will be the first brain stroke data repository in Bangladesh. Therefore, we have collected brain stroke patient's data from different hospitals of Bangladesh. We also finalize minimum number of features dimensionality by using feature importance score for our system to reduce hassles. Six machine learning classification techniques such as of Logistic Regression, K-Nearest Neighbor, Support Vector Machines, Naive Bayes, Bagging and Decision Tree have been applied. It has been seen that Bagging technique performs better than other classifiers with almost 96% of accuracy.

References

[1] Walter Johnson, Oyere Onuma, Mayowa Owolabi, and Sonal Sachdev, “Stroke: A global response is needed,” Bulletin of the World Health Organization, Volume 94, Pages 634–635A, 2016.

[2] World Life Expectancy (2019). Stroke in Bangladesh. [online] Available at: https://www. Worldlifeexpectancy.com/ bangladesh-stroke [Accessed 15 Feb. 2019].

[3] "Types of Strokes: Causes, Symptoms, and Treatments", Healthline, 2019. [Online]. Available: https://www.healthline.com/health/stroke-types. [Accessed: 15 Feb. 2019].

[4] Shai Shalev-Shwartz, Shai Ben-David, ―Understanding Machine Learning, ‖ Cambridge University Press, New York,NY 10013-2473, 2014.

[5] Introduction to Support Vector Machines [online]. Available: http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html. [Accessed: 20 Feb. 2019].

[6] Ethan Alapydin, ―Introduction to Machine Learning, ‖ 2nd ed. Cambridge Massachusetts, MIT Press, London, 2010.

[7] F.Schoonjans, "Logistic regression", *MedCalc,* 2019. [Online]. Available: https:// www.Medcalc.org/ manual/logistic_regression.php. [Accessed: 21- Sep- 2019]

[8] B. data and I. R, "Introduction to KNN, K-Nearest Neighbors: Simplified", *Analytics Vidhya*, 2019. [Online]. Available: https://www.analyticsvidhya.com/blog/ 2018/03/ introduction k neighbours algorithm clustering [Accessed: 21- Sep- 2019]

[9] Bradley Efron and R. J. Tibshirani, “An Introduction to the Bootstrap”, Chapman and Hall/CRC, New York, NY, USA, 1994.

[10] Zhi-Hua Zhou, “Ensemble Methods: Foundations and Algorithms”, Chapman & Hall/CRC, Boca Raton, Fla, USA, 2012.

[11] "Types of Strokes: Causes, Symptoms, and Treatments", Healthline, 2019. [Online]. Available: https://www.healthline.com/health/stroke-types. [Accessed: 15 Feb. 2019].

[12] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, Honglak Lee, “An Integrated Machine Learning Approach to Stroke Prediction Categories and Subject Descriptors”, KDD’10 Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and data Mining, Pages 183–192, 2010.

[13] Nikola Kasabov, Valery Feigin, Zeng-Guang Hou, Yixiong Chen, Linda Liang, Rita Krishnamurthi, Muhaini Othman, Priya Parmar, “Evolving spiking neural networks for personalised modelling, classification and prediction of spatio-temporal patterns with a case study on stroke”, Neurocomputing, Volume 134, Pages 269-279, 2014.

[14] Paul Bentley, Jeban Ganesalingam, Anoma Lalani Carlton Jones, Kate Mahady, Sarah Epton, Paul Rinne Pankaj Sharma, Omid Halse, Amrish Mehta, Daniel Rueckert, “Prediction of stroke thrombolysis outcome using CT brain machine learning”, NeuroImage: Clinical, Volume 4, Pages 635-640, 2014.

[15] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, David Madugan, “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model”, The Annals of Applied Statistics, Volume 9, Issue 3, Pages 1350-1371, 2015.

[16] Shanthi Dhanushkodi, G. Sahoo, Saravanan Nallaperumal, “Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke” Imternational Journal of Biometrics and Bioinformatics, Volume 3, Pages 10-18, 2009.

[17] Arslan AK, Colak C, Sarihan ME, “Different medical data mining approaches based prediction of ischemic stroke”, Computer Methods and Programs in Biomedicine, 2016.

[18] Sung SF, Hsieh CY, Kao Yang YH, Lin HJ, Chen CH, Chen YW, Hu YH, “Developing a stroke severity index based on administrative data was feasible using data mining techniques”, Journal of Clinical Epidemiology, Volume 68, Issue 11, Pages 1292-1300, 2015.

[19] "Stroke Statistics | Internet Stroke Center", *Strokecenter.org*, 2019. [Online]. Available: http://www. strokecenter. org/patients/about-stroke/stroke-statistics/. [Accessed: 21- Sep- 2019]

[20] Qinbao Song and Marin John Shepperd, "Missing Data Imputation Techniques", International Journal of Business Intelligence and Data Mining, Volume 2(3), Pages 261-291, 2007.

[21] "Hospital Test Rate | National Heart Foundation of Bangladesh ", *Nhf.org.bd*, 2019. [Online]. Available: http://www.nhf.org.bd/hospital_charge.php?id=4. [Accessed: 05-Nov- 2019]

Appendix A: Publication from this Research Work

<u>**Conference Paper:**</u>

1. **Md. Azizul Hakim**, Md. Zahid Hasan, Md. Mahabur Alam, Md. Mehadi Hasan and Mohammad Nurul Huda, “An Efficient Modified Bagging Method for Early Prediction of Brain Stroke”, *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 11-12 July, 2019.* (Accepted)