A Decision Support System for Early Prediction of Brain Stroke Disease in Bangladesh

BY

Md. Mahabur Alam
ID: 152-15-5811

AND

Md. Mehadi Hasan

ID: 152-15-6016

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Azizul Hakim
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Nusrat Jahan
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

APRIL 2019

APPROVAL

This Project/internship titled **"A Decision Support System for Early Prediction of Brain Stroke Disease In Bangladesh"**, submitted by Md. Mahabur Alam, ID No: 152-15-5811 and Md. Mehadi Hasan, ID No: 152-15-6016 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 02-05-2019.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Ismail Jabiullah — **Internal Examiner**
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Saddam Hossain Mukta — **External Examiner**
Assistant Professor
Department of Computer Science and Engineering
United International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Azizul Hakim,** Lecturer and Co-supervision of **Nusrat Jahan,** lecturer, Department of CSE, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

6.5.19

Md. Azizul Hakim
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Nusrat Jahan
5.5.19

Nusrat Jahan
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Mahabur

Md. Mahabur Alam
ID: -152-15-5811
Department of CSE
Daffodil International University

mehadi

Md. Mehadi Hasan
ID: -152-15-6016
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At first, we are appreciative to Almighty Allah for his kindness and elegance without which we wouldn't be able to complete our project. We needed to strive to take care of business yet we are thankful to some other individuals, without the assistance of whom this venture couldn't be for what it's worth.

We would like to express our deep and sincere gratitude to our Supervisor **Md. Azizul Hakim**, **Lecturer and** Co-Supervisor **Nusrat Jahan**, **Lecturer,** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor and Co-supervisor in the field of "*Machine learning and Data mining*" helped us to carry out this project. This entire time they have upheld and enlivened us and demonstrated the correct way. Their endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stages have made it possible to complete this project.

They were so agreeable with us this entire time and that was the principle motivation for us. We are so fortunate to work under their watch and obviously, it has been a respect to work under their watch. We additionally need to our offer our most profound thanks to noteworthy Professor and Head of CSE department, **Prof. Dr. Syed Akhter Hossain.**

We would like to thank our entire course mate in Daffodil International University, who took part in discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Brain Stroke is a Neurological disease that occurs when the blood supply to brain is interrupted or reduce, depriving brain tissue of oxygen and nurturance. This can lead to brain damage to possibly death. Brain Stroke are a medical emergency and prompt treatment is essential because the sooner a person receives treatment stroke, the less damage is likely to happen. Brain Stroke is the second leading cause of death in worldwide and third in Bangladesh. Information gain from health data may lead to innovative solution of better treatment plan for patients. In order to gain knowledge intelligently from brain stroke data, some machine learning technique and utilized to process data and generated data model that can be used to predict brain stroke disease or extract valuable information about brain stroke disease. In this study a data mining model has been build using a few machine learning algorithm, which can find out the important features of brain stroke as well as efficiently predict brain stroke.

TABLE OF CONTENTS

LIST OF FIGURES

PAGE NO

FIGURES

LIST OF TABLES

PAGE NO

CHAPTER 1

INTRODUCTION

1.1 Introduction

A brain stroke occurs when the flow of blood to part of the brain is cut off or significantly reduced. Without the oxygen carried by the blood, brain cells die quickly which can cause permanent brain damage. Brain stroke can be major or minor and the consequences can range from complete recovery to fatality. There are two types of brain stroke:

- Ischemic
- Hemorrhagic

An ischemic stroke is caused by lack of blood flow to brain tissue. This can happen when the arteries in the brain narrow due to a condition such as atherosclerosis or an embolism. About 87% of strokes are ischemic. Strokes are caused by a rapture in a blood vessel in the brain. A hemorrhagic stroke is also called an intra-cerebral hemorrhage or an ICH. About 13% of strokes are hemorrhagic. Stroke is the 5th leading cause of death in the United States and a major cause of serious disability for adults.

In Bangladesh, Brain stroke has been ranked as the third leading cause of death after coronary heart disease and infectious diseases such as influenza and pneumonia. The mortality rate of brain stroke increases from 6.00% (in 2006) to 8.75%, (in 2011) with an age-adjusted mortality rate of 108.31 per 100000 people (in 2011). The World Health Organization (WHO) ranks mortality due to Brain stroke in Bangladesh as number 84 in the world [25]. The cured death rate per 1000 people in Bangladesh is reported at 5.8%; the female and male life expectancies are reported as 64.4 years old and 65.1 years old, respectively [26]. With regard to these findings and emphasis on prediction of stroke incidence to reduce complications, disabilities and healthcare costs, this study was aimed to investigate 23 risk factor for brain stroke.

After collecting data, we have utilized three machine learning algorithms in our model which are:

i. Support vector machine.
ii. Random forest.
iii. Logistic recreation.

Our model is trained in such a way in which we can determine that the brain stroke exists or not.

1.2 Motivation

In rural Bangladesh, 26% of the people aged 40 years or above are suffering from hypertension, and of them 21.5% have suffered brain stroke [24], says a new study on Bangladesh, Sri Lanka and Pakistan. According to the study, Bangladeshis suffer the highest rate of stocks among 3 south Asian countries. It is the high time that the health system of the country is empowered for further research which may be required to find the factors leading to a higher rate of brain stocks in Bangladesh. To find out the risk factors behind the increasing rate of brain stocks in Bangladesh, we came up with a plan to build a model using data science and machine learning algorithm which can tell us about high risk factors by analyzing the existing data.

Data science through machine learning algorithm is currently becoming an essential aid for the diagnosis, treatment and prediction of complications and patient outcomes in a number of diseases. The evaluation and treatment of acute ischemic stroke have experienced a significant advancement over the first few years. In this paper, we offer an insight into the recent developments and applications of machine learning in brain stroke data analysis and prediction.

1.3 The rational of the study

Machine learning (ML), considered a branch of Artificial Intelligence (AI), is a field of CSE that facilities extraction of database on pattern recognition. A computer learns from previous mistakes after repeated analysis of data and masters tasks that work previously considered too complex for machine to process. The development of this system to interpret data in brain stroke analysis has provided valuable information for research in matter of interaction, structure and mechanism of brain and behavior in certain neurological disorders.

Machine learning (ML) systems are now being implemented in clinical neurosciences to devise imaging based diagnostic and classification systems of certain neurological and psychotic disorders. In this paper, we discuss the present day role of machine learning focusing on brain stroke disease, analyzing the data using the algorithms and finding the key factors behind the growing rate of brain stroke of Bangladesh.

1.4 Research Questions

- Does it demonstrate the exact incentive to anticipate cerebrum stroke in early forecast?
- Can it classify the brain stroke disease using machine learning algorithm?
- Does every algorithm work perfectly (no)?

Effectively, a few destructive illnesses have been distinguished for an individual. Albeit each ailment has an answer for aversion it's unrealistic for everybody because of just for obviousness. Everybody needs to have a cheerful existence in where an infection is the main hindrance. Any sort of ailment anticipation is conceivable if that in stay essential stage. For that reason, we manufactured an expectation framework that recognizes the sickness arrange and gives us the outcome that the individual in question has cerebrum stroke or not. The majority of the illnesses, mind stroke infection is viewed as one of the main sicknesses. Numerous people groups are passed on due to this sickness. Mind stroke illness is the greatest enemy of the two people in the world. In our Bangladesh there are no outstanding framework for mind stroke. At last, we chosen it as our exploration theme in Bangladesh individuals for our pleasure and furthermore attempt to make a decent framework for forecast mind stroke ailments.

1.5 Expected Outcome

In our brain stroke system helps to generate an expected result based on given data according to our dataset. Here we used 70% training data and 30% of test data to get more accuracy result. Our result is how much accuracy it depends entirely on the training data set. Our machine learning system will be ready, after completing all the needed procedure of our system. To getting accurate output we applied various strategies. We got 95% accuracy from “Random Forest Algorithm and Logistic Regression algorithm” among all of our applied algorithm.

1.6 Layout of the Report

- Chapter 1 have demonstrated an Introduction to the Research with its Motivation, the rational of the study, research questions, expected outcome.
- Chapter 2 will have “Background” demonstrates introductions, related works, research summery and challenges.
- Chapter 3 will have Research Methodology.
- Chapter 4 will have Experimental Results and Discussions.
- Chapter 5 will have Summary and Conclusion.

CHAPTER 2

Background Study

2.1 Introduction

The incidence of brain stroke increasing in Bangladesh in comparison to developed country. It is very important to know the symptomatology of brain stroke in the vascular territory wise, but this is very complex. We have under taken this research project throughout which we will try to figure out the most important features that are responsible for brain stroke with probity of the brain stroke.

In this part, we will discuss our related jobs, project summary and challenges about this research. In related job section, we will discuss about other research papers which are related to our work. In summary part, we will discuss about our project summary and in challenge part, we will discuss about how we increase the accuracy level.

2.2 Related Work

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, depriving brain tissue of oxygen and nutrients. Within minutes, brain cells begin to die. A stroke is a medical emergency. Prompt treatment is crucial. Early action can minimize brain damage and potential complications. There is no doubt that brain is most complex and important part of every human body. To lead a better healthy life, we have to be very cautious. If we can identify this disease at its primary level it is possible to overcome it. Otherwise we will suffer for long time or face death.

2.3 Research Summary

In this table 2.3.1, we are show some short description of some research paper which are related our topic.

Table 2.3.1: Research paper summary

Serial No	Author Name	Methodology	Description	Outcome
1	Luan Gao, Feng-gang Li, Jian Wang, Yu Liang,Yu Li. [1]	Association rule method, Xin'an physician method.	probe into the medicine rules for stroke	Chenpi, Fuling, Gancao, Luxiancao,

			prevention treated by Xin'an physicians.	Xiqiancao, Baijili, and Shinanye this seven used for Xin'an physicians' stroke prevention.
2	Sheng-feng sung, cheng-yang hsieh, yea-huei kao yang, Hue-juan lin, chih-hung chen, Yu-wei chen, Ya-han hu. [2]	Stroke severity index (SSI) and National Institute of health stroke scale (NIHSS) method.	Using administrative data it develop stroke sequrity index.	Identify seven predictive feature and three models.
3	Ahmet K. Arslan, Cemil Colak, Ediz Sarihan. [3]	Support Vector Machine (SVM), Stochastic gradient boosting (SGB), Panalized logistic regration (PLG).	Intend to access different medical data mining approaches to predict ischemic stroke.	Accuracy value is about 95%.
4	Hyo-Ki Lee, Joo-Han Kim, Hyoun-Seok Myoung, Jung-Hun Lee and Kyoung-Joung Lee. [4]	Morphological Operators, Bland-Altman Plot.	evaluate the repeatability of the accelerometric-method to detect step events for hemiparetic stroke patients.	A good result achieved.
5	efthyvoulos c. kyriacou, constantinos s. pattichis, minas a. karaolis, christos p. loizou, christodoulosi. christodoulou, marios s. pattichis, stavros kakkos, and andrew nicolaides. [5]	probabilistic neural network (PNN) and support vector machines (SVM) classifiers.	an integrated system for the assessment of risk of stroke based on two modules: 1) clinical risk factors and noninvasive investigations, and 2) carotid plaque texture analysis.	Demonstrate a better accuracy.
6	Ye Liu, Hao Zhang, Min Chen and Liqing Zhang. [6]	Resample Heuristic Algorithm, Common Spatial-Spectral	Imagery Electroencephalogram (EEG) based Brain-Computer Interface (BCI) system	Average outcome.

		Boosting Pattern (CSSBP) Algorithm.	can be used as a rehabilitation tool for stroke patients.	
7	Jose Rafael Romero, Jane Morris, Aleksandra pikula. [7]	Framingham Stroke Risk profile (FSRP)	The impact of modifiable traditional vascular risk factor on ischemic stroke, Interventions for stroke prevention and evidence for early treatment.	Identify the risk factor of stroke.
8	Seung Nam Min, Kyung-Sun Lee, Se Jin Park, Murali Subramaniyam, and Dong Joon Kim. [8]	Logistic Regression (LR), Deep Learning.	a model equation for developing a stroke pre-diagnosis algorithm with the potentially modifiable risk factors.	Sensitivity 64.7%.
9	Safa Aouinti, Hela Mallek, Dhafer Malouche, Olfa Saidi ,Olfa Lassouedi, Faycel Hentati, Habiba Ben Romdhane. [9]	EM-algorithm, BIC model, graphical interaction models.	Evaluate the medical cost of managing this disease and to identify risk factors that influence its variation in Tunisia. And identify the factors that make these costs high.	This is deduced from a multimodal probability distribution taking the appearance of a mixture of Gaussian distribution.
10	M. Sheetal Singh, Prakash Choudhary. [10]	Decision tree, back propagation neural network classification algorithm.	compare different methods for stroke prediction on the Cardiovascular Health Study (CHS) dataset.	97.7% accuracy gained from implemented models.
11	P. Chantamit-o-pas, M. Goyal. [11]	Case-based reasoning (CBR) method.	Conceptual CBR framework for stroke disease prediction that uses previous case-based knowledge.	quite significant decision-making for patients.

Cerebrum stroke is an illness that assaults the mind. The cerebrum is a very imperative piece of each individual. In this way, in the event that we need to have a solid existence, we must be careful. In the event that we discover this sickness from the get-go in the underlying stage, we can without much of a stretch beat it. Else we should languish over this ailment for our future life. Subsequent to making the choice dependent on the ebb and flow circumstance, we needed to build up a

framework that provides better execution because of malady and comprehend the circumstance of influenced patients.

At long last, we contact our normal objective for God's favoring, which we have thought to execute.

2.4 Challenges

Data collection is the biggest challenges to get our predicting accuracy. In Bangladesh perspective data collection is not a simple process. There have no collection data in any hospital or medical center in our country which is helpful for our project. As a result we went many more hospital and collected data manually and then we store it in CSV file in our pc. For this reason we contacted every hospitals register office for many time that is so much time consuming. Firstly, they didn't agree to permit for collecting data. Then we informed our super visor sir, then he contacted with department head sir and then he provide a reference as the side of our research work. Then hospitals resister office accept our application but they took too much time (minimum 1 week) to provide us final permission. We didn't get perfect data. There had also many missing data. So, we imputed those missing data. Therefore, different algorithm has been applied to the proposed architecture. Finally, the implementation process has been established to get accurate predicted value. There were several challenges rising according to the working procedure.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this section the methodology explained including on how the data set is obtained, attributes, machine learning algorithm and evaluation criteria.

3.2 Data Collection Procedure

3.2.1 Data Collection:

Data used in this project 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)", "Shaheed Sheikh Abu Naser Specialized Hospital Khulna", "Gazi Medical College Khulna", "Islami Bank Hospital Khulna", " Jhenaidah Sadar Hospital, Jhenaidah" and "Islami Bank Community Hospital Jhenaidah".

This data set consists of 385 instances, Where male 209 (54%) and female 176 (46%). Our data set also contains brain stroke are 234 (61%) and non-brain stroke are 151(39%).

That is show on bar chart:

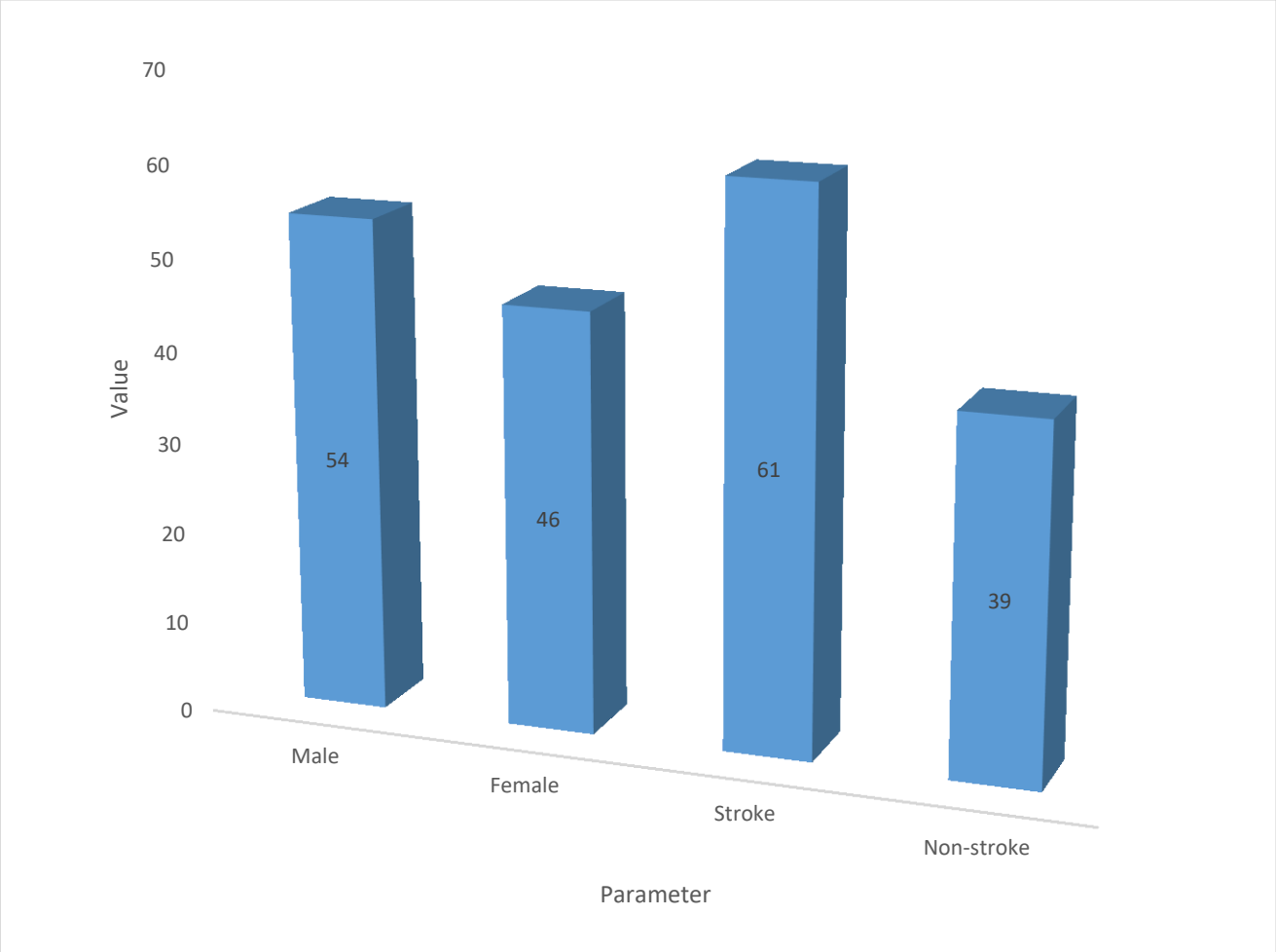

Figure 3.2.1.1: display results.

In figure 3.2.1.1 we can see the percentage of gender and final result of our dataset. From bar chat we can say that male patients are more face brain stroke disease.

3.2.2 Attribute:

The obtained data set contains 24 attributes. The main attributes are (Hyper tension, LDL, HDL, RBS, Triglyceride, Heart disease, RBC, Age, Average Glucose Level, BMI, Serum Cholesterol, HbA1c, Hb, WBC, Work type, Smoking, Serum Creatine, ESR). Moreover, data divided to two separates data sets “Training data set to build the model and test data set to evaluate the model”. We find out this figure to apply the Random Forest Algorithm.

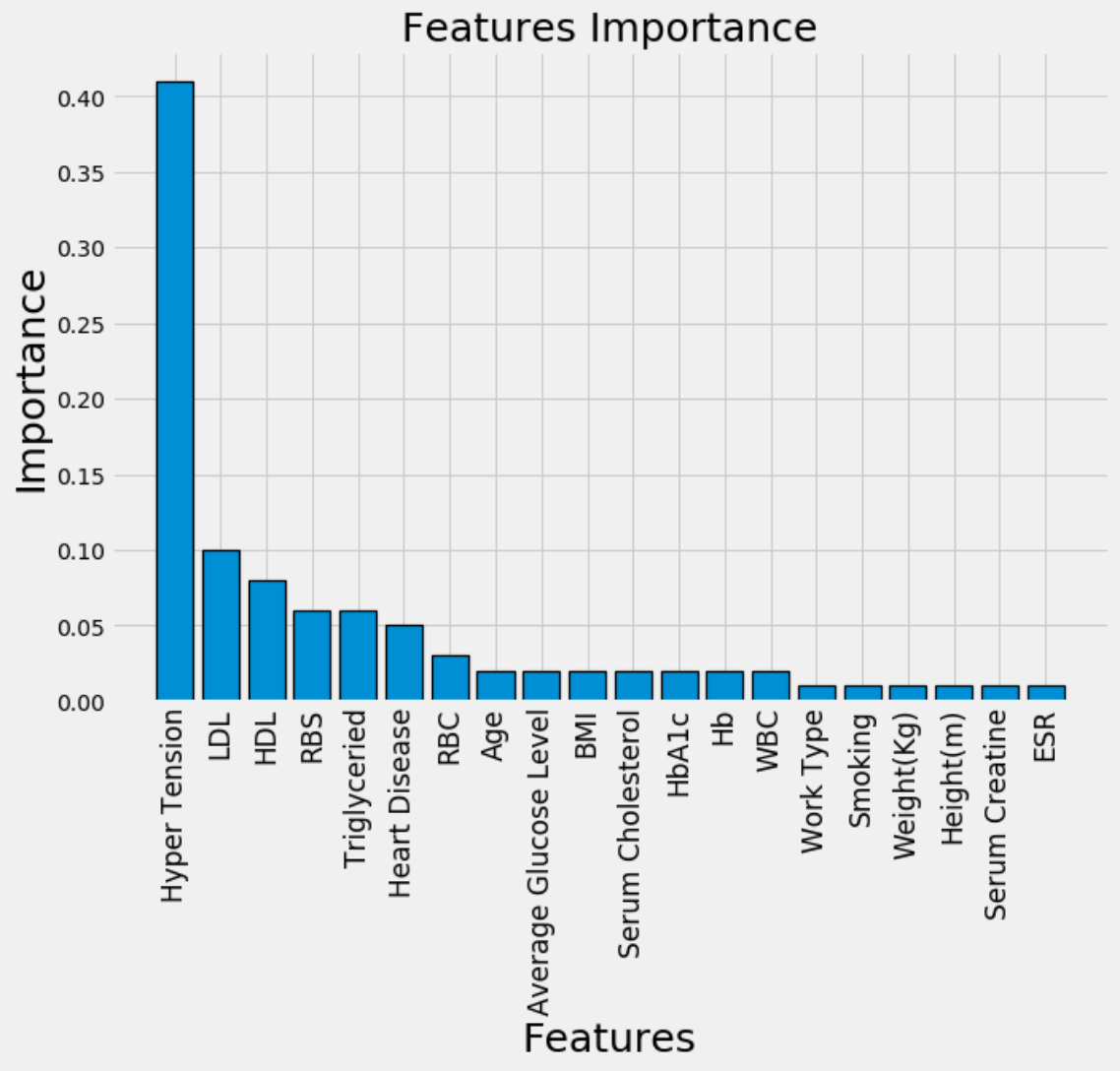

Figure 3.2.2.1: Important Features.

We need to demonstrate our dataset data where highlights are hyper tension, LDL (Low Density Lipoproteins). HDL (High Density Lipoproteins), RBS (Random Blood Sugar), Triglyceride, Hearth Disease, RBC (Red Blood Cells), Age, Average Glucose Level, BMI/(Weight and Height), Serum Cholesterol, HbA1c, Hb (Hemoglobin), WBC (White blood Cells), Work Type, Smoking, Serum Creatine, ESR (erythrocyte sedimentation rate).

3.2.3 Missing Data Imputation:

Data sets contains 385 records in which some values missing. We have handle the missing values by using Mean imputation. It is a method in which the missing value on a certain attribute is replaced by the min of the available cases. This method maintain the sample size and easy to use.

3.2.4 Machine Learning Algorithms:

For their known accuracy rate Random Forest Algorithm, Logistic Regression, Support Vector Machine algorithm we use to applied on the training data set to build a model. All the applied algorithm on python and libraries are NumPy, Pandas, Scikit-learn and Matpolib.

3.3 Statistical Analysis

In our dataset contains 385 records where 234 data for brain stroke and 151 data for non-brain stroke patients data. Here, in our model we selected 70% train data and 30% test data. To applying machine learning algorithm such as (Random forest algorithm, Support Vector Machine, Logistic Regression) we try to find out better accuracy in our model. In this figure 3.3.1 we are shown that dataset flowchart and how we use the dataset for proposed model.

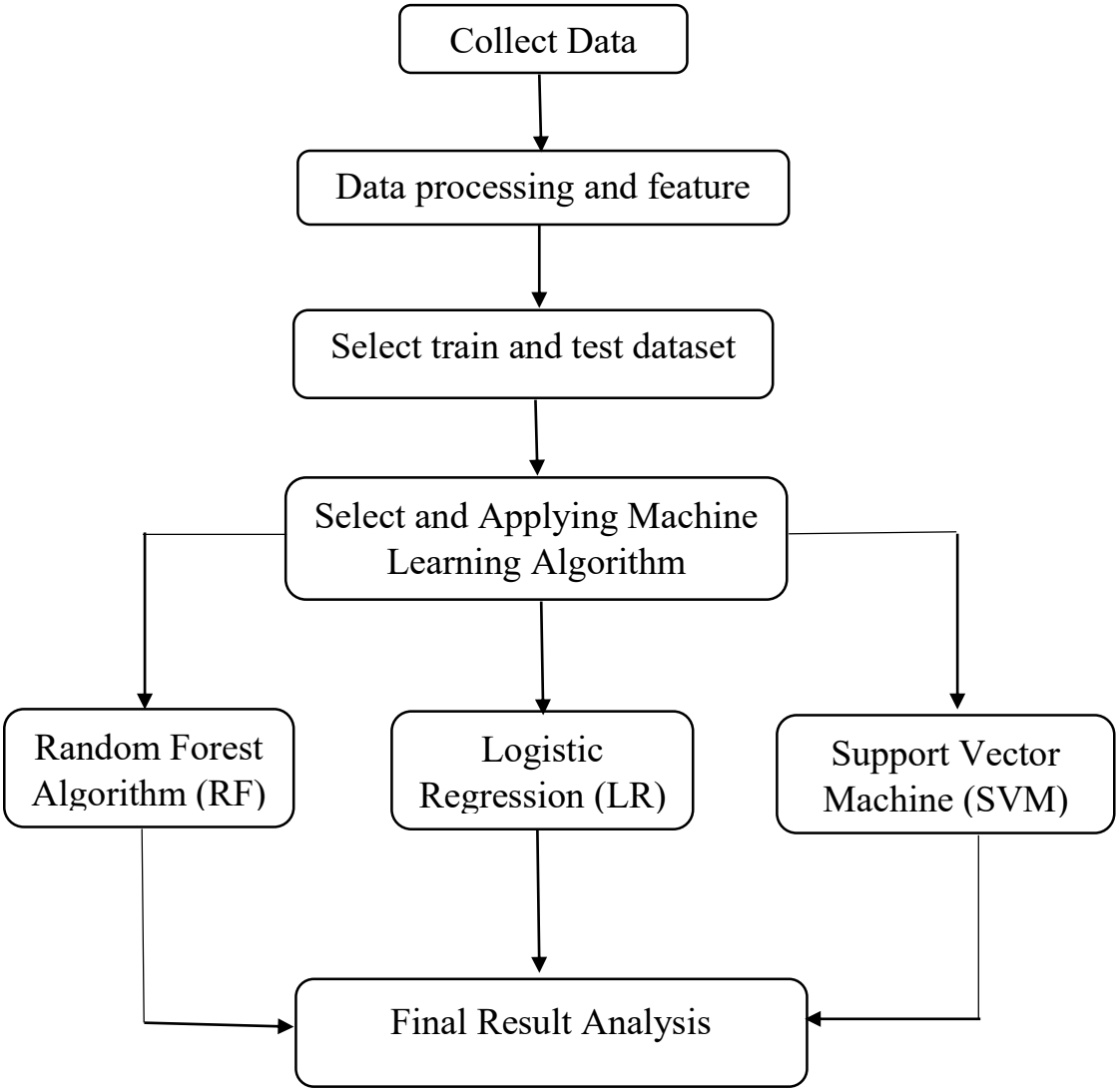

Figure 3.3.1: proposed model Structure

In this figure, we have shown that how we do our research shortly details, In this figure (3.3.1) we can know how to go ahead for our target step by step.

3.4 Research subject and Instrumentation

Starting late, the acclaim of machine learning figuring's is rising exponentially. Machine learning Algorithms enable PCs to pick up from data with the help of quantifiable strategies. A machine can find the inside data precedent and convey a decision or farsighted learning accordingly without the help of express coding is considered as the most premium part. Along these lines, a comparative estimation can be associated with dataset of different regions without having a difference in its

inside structures. There are particular sorts of machine learning counts, anyway we have used some of them to our system.

3.4.1 Random Forest (RF)

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

Let's suppose you have decided to ask your friends, and talked with them about their past travel experience to various places. You will get some recommendations from every friend. Now you have to make a list of those recommended places. Then, you ask them to vote (or select one best place for the trip) from the list of recommended places you made. The place with the highest number of votes will be your final choice for the trip.

In the above decision process, there are two parts. First, asking your friends about their individual travel experience and getting one recommendation out of multiple places they have visited. This part is like using the decision tree algorithm. Here, each friend makes a selection of the places he or she has visited so far.

The second part, after collecting all the recommendations, is the voting procedure for selecting the best place in the list of recommendations. This whole process of getting recommendations from friends and voting on them to find the best place is known as the random forests algorithm.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class

is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms. [20]

The default values for the parameters controlling the size of the trees (e.g. max_depth, min_samples_leaf, etc.) lead to fully grown and unpruned trees which can potentially be very large on some data sets. To reduce memory consumption, the complexity and size of the trees should be controlled by setting those parameter values.

The features are always randomly permuted at each split. Therefore, the best found split may vary, even with the same training data, max_features=n_features and bootstrap=False, if the improvement of the criterion is identical for several splits enumerated during the search of the best split. To obtain a deterministic behaviour during fitting, random_state has to be fixed.

The default value max_features="auto" uses n_features rather than n_features / 3. The latter was originally suggested in [22], whereas the former was more recently justified empirically in [23]

3.4.2 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally characterized by an isolating hyperplane. At the end of the day, given marked preparing information (managed taking in), the calculation yields an ideal hyperplane which arranges new precedents. In two-dimensional space this hyperplane is a line partitioning a plane in two sections where in each class lay in either side the numeric info factors (x) in your information (the sections) shape a n-dimensional space. For instance, on the off chance that you had two information factors, this would frame a two-dimensional space [16].

A hyperplane is a line that parts the information variable space. In SVM, a hyperplane is chosen to best separate the focuses in the info variable space by their class, either class 0 or class 1. In two-measurements you can envision this as a line and we should expect that the majority of our info focuses can be totally isolated by this line. For instance:

$$B0 + (B1*X1) + (B2*X2) = 0 \ldots\ldots\ldots\ldots\ldots (3)$$

From equation 3, here the coefficients (B1 and B2) that decide the incline of the line and the capture (B0) are found by the learning calculation, and X1 and X2 are the two info factors [16].

SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points.

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

3.4.3 Logistic Regression (LR)

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

The implementation of logistic regression in scikit-learn can be accessed from class Logistic Regression.

This implementation can fit binary, One-vs- Rest, or multinomial logistic regression with optional L2 or L1 regularization.

As an optimization problem, binary class L2 penalized logistic regression minimizes the following cost function:

$$\min_{w,\,c} \frac{1}{2} w^T w + C \sum \log\left(\exp(-y_i(X_i^T w+c))+1\right).$$

Similarly, L1 regularized logistic regression solves the following optimization problem.

$$\min_{w,\,c} \|w\|_1 + C \sum \log\left(\exp(-y_i(X_i^T w+c))+1\right).$$

Note that, in this notation, it's assumed that the observation y_i takes values in the set -1, 1 at trial i The solvers implemented in the class Logistic Regression are "liblinear", "newton-cg", "lbfgs", "sag" and "saga":

The solver "liblinear" uses a coordinate descent (CD) algorithm, and relies on the excellent C++ LIBLINEAR library, which is shipped with scikit-learn. However, the CD algorithm implemented

in liblinear cannot learn a true multinomial (multiclass) model; instead, the optimization problem is decomposed in a “one-vs-rest” fashion so separate binary classifiers are trained for all classes. This happens under the hood, so Logistic Regression instances using this solver behave as multiclass classifiers. For L1 penalization sklearn.svm.l1_min_c allows to calculate the lower bound for C in order to get a non “null” (all feature weights to zero) model.

The “lbfgs”, “sag” and “newton-cg” solvers only support L2 penalization and are found to converge faster for some high dimensional data. Setting multi_class to “multinomial” with these solvers learns a true multinomial logistic regression model [12], which means that its probability estimates should be better calibrated than the default “one-vs-rest” setting.

The “sag” solver uses a Stochastic Average Gradient descent [15]. It is faster than other solvers for large datasets, when both the number of samples and the number of features are large.

The “saga” solver [17] is a variant of “sag” that also supports the non-smooth penalty=“l1” option. This is therefore the solver of choice for sparse multinomial logistic regression.
The “lbfgs” is an optimization algorithm that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm [18], which belongs to quasi-Newton methods. The “lbfgs” solver is recommended for use for small data-sets but for larger datasets its performance suffers. [19]

3.5 Selected Algorithm

We utilize different algorithm to get highest accuracy to exactness from our dataset. In this figure, we are show which algorithm are given best precision among another algorithm.

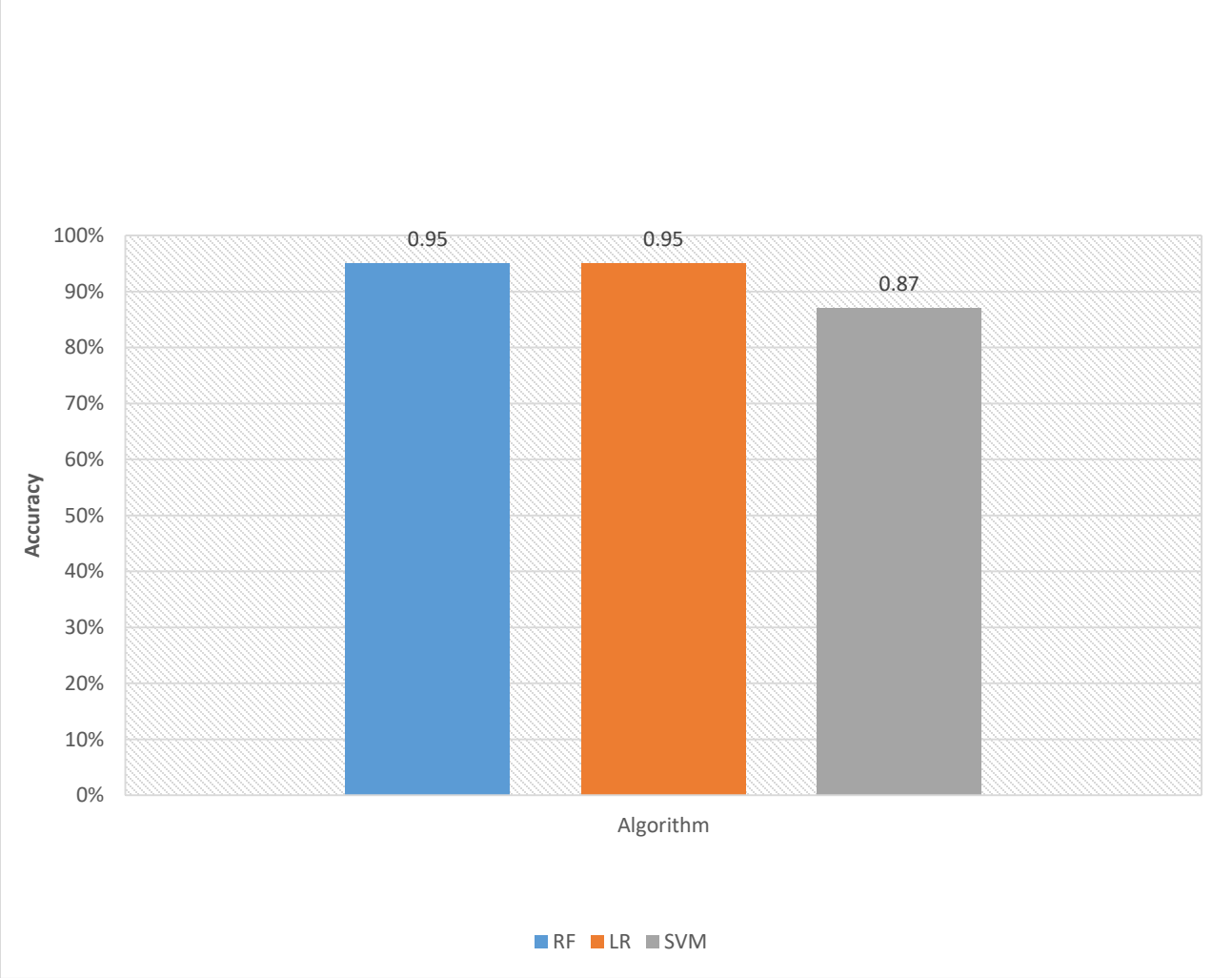

Figure 3.5.1: Accuracy Level

In figure 3.5.1, we show 3 models Random forest, Logistic Regression and Support vector machine and precision estimations for each. We need to balance the models with each other and select the most accuracy Random forest and Logistic regression.

We choose the model of Random forest, because it is very face and time consuming. That's why for our model we choose random forest algorithm.

3.6 Proposed Algorithm

In our proposed strategy, we use Random Forest algorithm. In our algorithm, we endeavor to construct a model which predicts brain stroke disease. Our proposed system intends to improve the execution of Random forest classifier for disease gauge. Anaconda is an environment that consists of Python, R and all deep learning packages. Python variant 3.6.3 has been utilized and considered as a most recent form of python. Different kinds of library Functions have been utilized for usage. In our method –

Step 1.Firstly we select our datasets which contain brain stroke and normal dataset.

Step 2.Classification of dataset into patient with brain stroke and non-brain stroke.

Step 3: Input the dataset in Jupiter notebook.

Step 4: Apply machine learning algorithm in python.

Step 5: Find out highest accuracy from dataset of different machine algorithm.

Step 6: Get most astounding exactness utilizing Random forest algorithm.

Step 7: Apply Random forest algorithm we can find out the important features in brain stroke disease of our dataset.

Step 8: Measure the performance of the model.

Random forest algorithm takes the brain stroke dataset and classify whether a person is having brain stroke or not, it can also classify the important features of brain stroke. The Random forest algorithm is applied on pre-processed dataset and performance is measure.

We need to understand that the model we made is any extraordinary. A short time later, we will use real procedures to assess the precision of the models that we make on covered data. We furthermore need an increasingly strong measure of the precision of the best model on subtle data by surveying it on real covered data.

That is, we will hold down a couple of data that the computations won't get the chance to see also, we will use this data to get a second and free idea of how exact test. Best model may truly be. We will part the stacked dataset into two, 70% of which we will use to set up our models and 30% that we will hold down as an endorsement dataset. We will utilize 10-crease cross approval to gauge

exactness. This will part our dataset into 10 sections, train on 9 and test on 1 and rehash for all blends of train-test parts.

Cross-endorsement is a resampling strategy used to survey machine learning models on a limited data test. The technique has a singular parameter considered that insinuates the amount of social events that a given data test is to be part into.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Results

To measuring the performance of our system we used in dataset 23 features to test the accuracy. Our dataset contains 385 data which is cause for brain stroke and get the accuracy 95%. We use real dataset that's why we find the highest accuracy from the dataset. We also use confusion matrix to calculate precision, recall, F-measure, Support, True Positive Rate, True Negative Rate and accuracy of the model.

The confusion matrix is a table to describe the performance of a classification model on a set of test data. Confusion matrix can define four terms:

True Positive (TP): Accurately detected i.e. Stroke disease patient is classified as presence class of stroke disease.

True Negative (TN): Accurately rejected i.e. Non-stroke disease person is classified as absence class of stroke disease.

False Positive (FP): Inaccurately detected i.e. Non-Stroke disease person is classified as presence class of stroke disease.

False Negative (FN): Inaccurately rejected i.e. Stroke disease patient is classified as absence class of stroke disease.

Precision: Precision is the piece of related instances among the retrieved instances. High precision means that an algorithm returned substantially more relevant results than irrelevant ones.

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}}$$

Recall: Recall is the piece of relevant instances that have been retrieved over the total amount of relevant instances. High recall means that an algorithm returned most of the relevant result.

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fp}}$$

F-measure: F-score is a measure of test's accuracy by considering both precision and recall. It is a harmonic average of precision and recall.

$$F - score = \frac{2 * precision * recall}{\text{precision+recall}}$$

Accuracy: Accuracy refers to the familiarity of the measured value to a known value.

$$accuracy = \frac{\text{tp+tn}}{\text{tp+tn+fp+fn}}$$

True Positive Rate: True positive rate are refers that our proposed method predict the brain stroke is no brain stroke when it's actually brain stroke. Calculate the true positive rate by the given equation:

$$True\ positiverate = \frac{\text{TP}}{\text{TN+Fp}}$$

Specificity: Specificity refers that our proposed method predicts the brain stroke when it's actually occur brain stroke. Calculate the specificity of the given equation:

$$specificity = \frac{\text{TN}}{\text{TN+Fp}}$$

We know about confusion matrix which can help in ascertaining further developed arrangement measurements, for example, precision, recall, specificity and sensitivity of our classifier.

4.1.1 Random Forest (RF):

Confusion Matrix:

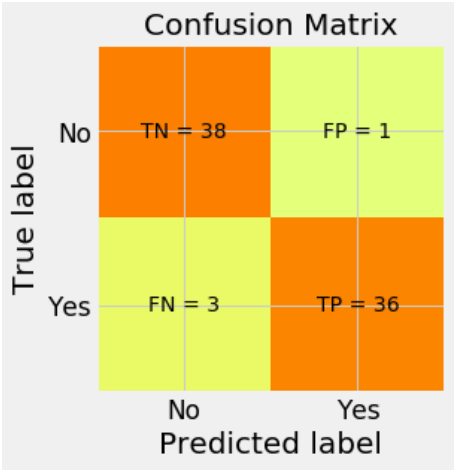

Figure 4.1.1.1: Confusion Matrix of RF

From figure 4.1.1.1 we can see in first row True Negative (TN) is 38 and False Positive (TP) is 1, so total no is 39. The second row is same result but here False Negative (FN) is 3 and True Positive (TP) is 36.

Accuracy:

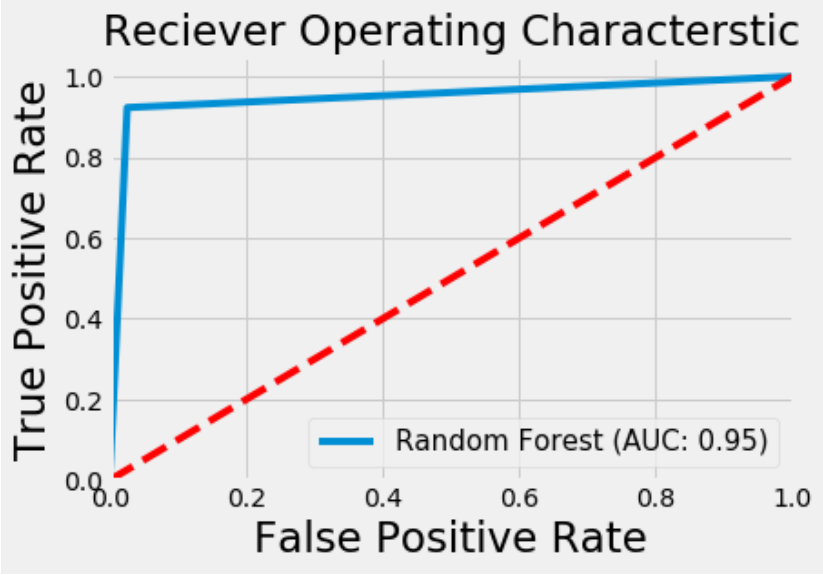

Figure 4.1.1.2: Accuracy Level of RF

We also get high Accuracy 95% for our research using “Random Forest” algorithm. In this figure 4.1.1.2 we get accuracy curve for our dataset and we get 95% accuracy for our dataset.

4.1.2 Logistic Regression (LR):

Confusion Matrix:

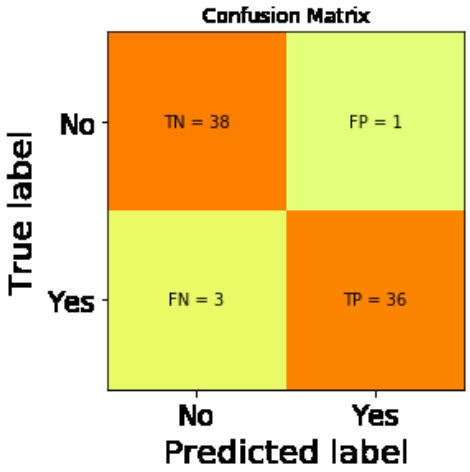

Figure 4.1.2.1: Confusion Matrix of LR

From figure 4.1.2.1 we can see in first row True Negative (TN) is 38 and False Positive (TP) is 1, so total no is 39. The second row is same result but here False Negative (FN) is 3 and True Positive (TP) is 36.

Accuracy:

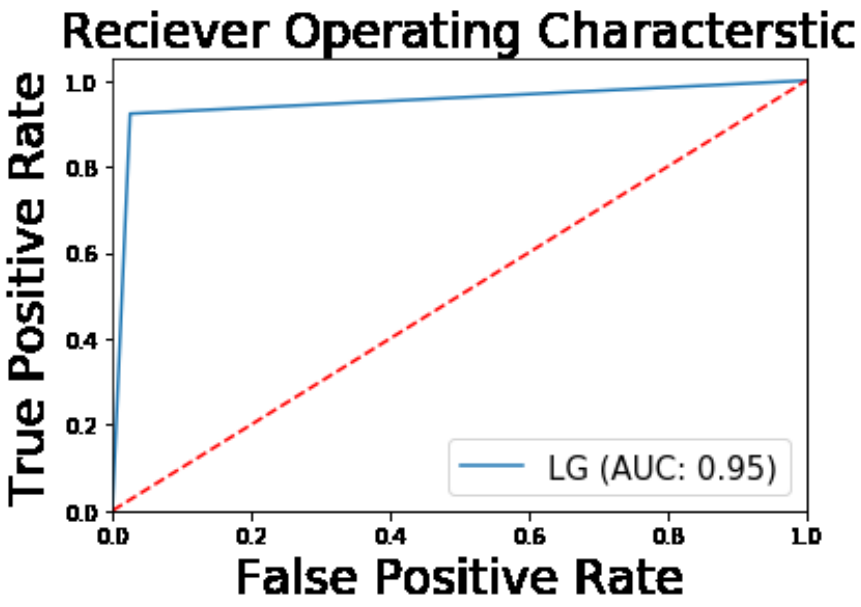

Figure 4.1.2.2: Accuracy Level LR

We also get high Accuracy 95% for our research using “Logistic Regression” algorithm. In this figure 4.1.2.2 we get accuracy curve for our dataset and we get 95% accuracy for our dataset.

4.1.3 Support Vector Machine (SVM):

Confusion Matrix:

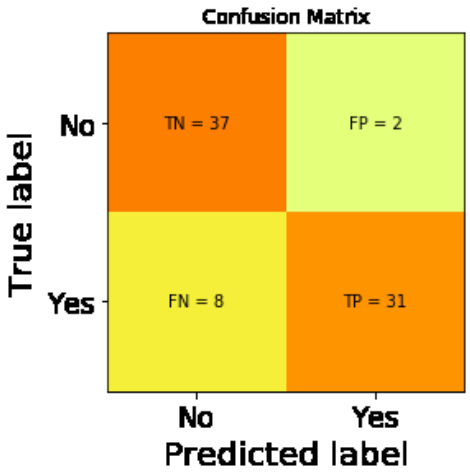

Figure 4.1.3.1: Confusion Matrix of SVM

From figure 4.1.3.1 we can see in first row True Negative (TN) is 37 and False Positive (TP) is 2, so total no is 39. The second row is same result but here False Negative (FN) is 8 and True Positive (TP) is 31.

Accuracy:

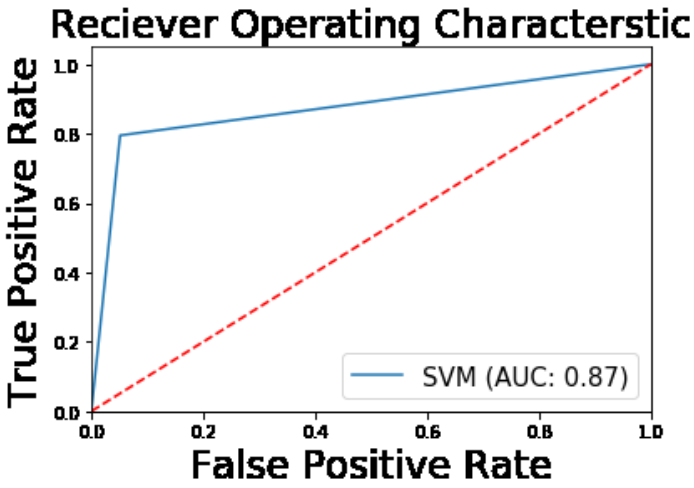

Figure 4.1.3.2: Accuracy Level of SVM

We also get high Accuracy 87% for our research using "Support Vector Machine" algorithm. In this figure 4.1.3.2 we get accuracy curve for our dataset and we get 87% accuracy for our dataset.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

High pressure of blood flow is the cause of brain stroke. The current investigation has gone for utilizing existing information and to evaluate the weight of stroke. The estimations depend on a few suppositions where the estimation of every part is questionable [21].

5.2 Future Work

All of the underlying methodological and computational complexities aside, our long term goal is to design an easy to use online system, allowing for relative prediction of the clinical outcome based on the demographics and clinical findings. Such a system has the potential for fine adjustment from the continuous training provided via handling large scale national or international multi-institutional users, with the advantage of easily incorporating newly available data to improve prediction performance. Another side to focusses our work perinatal Stroke, childhood Stroke.

REFERENCES

[1]L. Gao, F. Li, J. Wang, Y. Liang and Y. Li, "Analysis on Medicine Compounding for Stroke Prevention Treated by Xin'an Physicians Based on Association Rules", *2012 Fifth International Conference on Intelligent Computation Technology and Automation*, 2012. Available: 10.1109/icicta.2012.118 [Accessed 1 April 2019].

[2]S. Sung et al., "Developing a stroke severity index based on administrative data was feasible using data mining techniques", *Journal of Clinical Epidemiology*, vol. 68, no. 11, pp. 1292-1300, 2015. Available: 10.1016/j.jclinepi.2015.01.009 [Accessed 1 April 2019].

[3]A. Arslan, C. Colak and M. Sarihan, "Different medical data mining approaches based prediction of ischemic stroke", 2019.

[4]Hyo-Ki Lee, Joo-Han Kim, Hyoun-Seok Myoung, Jung-Hun Lee and Kyoung-Joung Lee, "Repeatability of the accelerometric-based method to detect step events for hemiparetic stroke patients", *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011. Available: 10.1109/iembs.2011.6091285 [Accessed 1 April 2019].

[5]E. Kyriacou et al., "An Integrated System for Assessing Stroke Risk", *IEEE Engineering in Medicine and Biology Magazine*, vol. 26, no. 5, pp. 43-50, 2007. Available: 10.1109/emb.2007.901794 [Accessed 1 April 2019].

[6]Y. Liu, H. Zhang, M. Chen and L. Zhang, "A Boosting-Based Spatial-Spectral Model for Stroke Patients' EEG Analysis in Rehabilitation Training", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 1, pp. 169-179, 2016. Available: 10.1109/tnsre.2015.2466079 [Accessed 1 April 2019].

[7]J. Romero, J. Morris and A. Pikula, "Review: Stroke prevention: modifying risk factors", *Therapeutic Advances in Cardiovascular Disease*, vol. 2, no. 4, pp. 287-303, 2008. Available: 10.1177/1753944708093847 [Accessed 1 April 2019].

[8]S. Min, K. Lee, S. Park, M. Subramaniyam and D. Kim, "Development of Stroke Diagnosis Algorithm Through Logistic Regression Analysis with National Health Insurance Database", *Advances in Intelligent Systems and Computing*, pp. 364-366, 2017. Available: 10.1007/978-3-319-60483-1_37 [Accessed 1 April 2019].

[9]S. Aouinti et al., "Graphical interaction models to extract predictive risk factors of the cost of managing stroke in Tunisia", *2013 International Conference on Computer Medical Applications (ICCMA)*, 2013. Available: 10.1109/iccma.2013.6506162 [Accessed 1 April 2019].

[10]M. Singh and P. Choudhary, "Stroke prediction using artificial intelligence", *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, 2017. Available: 10.1109/iemecon.2017.8079581 [Accessed 1 April 2019].

[11]P. Chantamit-o-pas and M. Goyal, "A Case-Based Reasoning Framework for Prediction of Stroke", *Information and Communication Technology*, pp. 219-227, 2017. Available: 10.1007/978-981-10-5508-9_21 [Accessed 1 April 2019].

[12]Christopher M. Bishop: Pattern Recognition and Machine Learning, Chapter 4.3.4 [Accessed 1 April 2019].

[13]L. Zeng, C. Meng, Z. Liang, X. Huang and Z. Li, "Stroke unit of integrative medicine for post stroke comorbid anxiety and depression: A systematic review and meta-analysis of 25 randomized controlled trials", *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014. Available: 10.1109/bibm.2014.6999360 [Accessed 1 April 2019].

[14]T. Feng and Q. Zhang, "Model of Environmental Reason Analysis and Prevention of Stroke on Linear Regression", *2015 11th International Conference on Computational Intelligence and Security (CIS)*, 2015. Available: 10.1109/cis.2015.26 [Accessed 1 April 2019].

[15] Mark Schmidt, Nicolas Le Roux, and Francis Bach: Minimizing Finite Sums with the Stochastic Average Gradient.

[16] Machine learning 101, https://medium.com/machine-learning-101/chapter-2-svm-support-vectormachine-theory-f0812effc72 (accessed on April 10, 2016).

[17] Aaron Defazio, Francis Bach, Simon Lacoste-Julien: SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives.

[18]"Broyden–Fletcher–Goldfarb–Shanno algorithm", *En.wikipedia.org*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Broyden%E2%80%93Fletcher%E2%80%93Goldfarb%E2%80%93Shanno_algorithm. [Accessed: 01- Apr- 2019].

[19]*Scikit-learn.org*, 2019. [Online]. Available: https://scikit-learn.org/0.20/_sources/modules/linear_model.rst.txt. [Accessed: 01- Apr- 2019].

[20]"Random Forests Classifiers in Python", *DataCamp Community*, 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/random-forests-classifier-python. [Accessed: 01- Apr- 2019].

[21] Kidwell CS, Warach S (December 2003). "Acute ischemic cerebrovascular syndrome: diagnostic criteria". Stroke. 34 (12): 2995–8.

[22]*Stat.berkeley.edu*, 2019. [Online]. Available: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. [Accessed: 01- Apr- 2019].

[23]P. Geurts, D. Ernst and L. Wehenkel, "Extremely randomized trees", 2019.

[24]B. Page and P. Palma, "Rate of strokes very high in Bangladesh", *The Daily Star*, 2019. [Online]. Available: https://www.thedailystar.net/backpage/rate-strokes-very-high-bangladesh-1577461. [Accessed: 14- Apr- 2019].

[25]"WHO Kobe", *WHO Kobe*, 2019. [Online]. Available: http://www.who.or.jp/uhcprofiles/Bangladesh.pdf. [Accessed: 14- Apr- 2019].

[26]2019. [Online]. Available: http://www.Worldlifeexpectancy. Com/bangladesh-stroke. [Accessed: 14- Apr- 2019].

Stroke Prediction

ORIGINALITY REPORT

20%	%	%	20%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Bahcesehir University Student Paper	4%
2	Submitted to Universidad Carlos III de Madrid Student Paper	2%
3	Submitted to National Institute Of Technology, Tiruchirappalli Student Paper	1%
4	Submitted to CSU, San Jose State University Student Paper	1%
5	Submitted to University College London Student Paper	1%
6	Submitted to Institute of Technology, Nirma University Student Paper	1%
7	Submitted to University of Ulster Student Paper	1%
8	Submitted to University of Sheffield Student Paper	1%