

Stroke Prediction Using Machine Learning Techniques

BY

Syed Washfi Ahmad
ID: 211-25-948

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

Supervised By

Md. Zahid Hasan
Associate Professor & Coordinator MIS
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Thesis titled “**Stroke Prediction Using Machine Learning Techniques**”, submitted by Syed Washfi Ahmad, ID No: 211-25-948 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 M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on January 22, 2022.

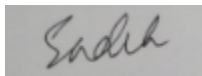
BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Md. Sadekur Rahman (SR)
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Moushumi Zaman Bonny
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



External Examiner

Dr. Shamim H Ripon

Professor

Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Md. Zahid Hasan, Associate Professor & Coordinator MIS, Department of CSE** Daffodil International University. I also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree or diploma.

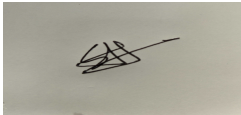
Supervised by:



Md. Zahid Hasan

Associate Professor & Coordinator MIS
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



Syed Washfi Ahmad

ID: 211-25-948
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to Almighty God for His divine blessing makes it possible to complete the final year thesis successfully.

I am really grateful and wish my profound indebtedness to **Md. Zahid Hasan, Associate Professor & Coordinator MIS**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of *Artificial Intelligence* to carry out this research. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this thesis.

I would like to express my heartiest gratitude to honorable Professor and Head, Department of CSE, **Professor Dr. Touhid Bhuiyan** for his kind help, to finish my research and also to other faculty members and the staff of the CSE department of Daffodil International University.

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

Finally, I must acknowledge with due respect, the constant support and patience of my parents.

ABSTRACT

Most of the strokes are due to an unanticipated blocking of courses by both the brain and the heart. Detection of different stroke warning signals can help to minimize the intensity of the stroke. This research suggests an early prediction of stroke illnesses by combining the incidence of hypertension, BMI, heart disease, average glucose level, smoking status, prior stroke, and age with various machine learning algorithms. For predicting strokes, seven different classifiers were trained using these high features. Logistics Regression, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, K-Nearest Neighbour Classifier, Random Forest Classifier, and XGBoost Classifier were used in the research. Furthermore, the proposed study produced a 94 percent accuracy rate, with the Random Forest classifier outperforming other classifiers. This model predicts strokes with the greatest accuracy. Random Forest has the lowest false positive and false negative rates when compared to other methods. As a consequence, Random Forest is nearly the ideal classifier for predicting stroke, which physicians and patients may use to prescribe and diagnose a probable stroke early.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	I- II
Declaration	III
Acknowledgements	IV
Abstract	V
CHAPTER	PAGE
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 Objectives	2
1.4 Expected Outcome	3
1.5 Report Layout	3
CHAPTER 2: LITERATURE REVIEW	4-7
2.1 Related Works	4-6
2.2 Scope of the problem	6
2.3 Challenges	7
CHAPTER 3: METHODOLOGY	8-17
3.1 Dataset and Features	8-9
3.2 Data Preprocessing	10-12
3.2.1 Handling Missing Data	10
3.2.2 Data Encoding	11

3.2.3 Feature Selection	11
3.2.4 Handling Imbalanced Dataset	12
3.2.5 Splitting the Data	12
3.3 Research Subject & Instrumentation	12-16
3.3.1 Classification Algorithms	13-14
3.3.1.1 Decision Tree	14-15
3.3.1.2 Random Forest	15
3.3.1.3 Naïve Bayes	15-16
3.3.1.4 K-Nearest Neighbor	16
3.4 Proposed Model	16-17
CHAPTER 4: RESULTS	18-23
4.1 Correlation Results	18
4.2 Experimental Results	19-23
CHAPTER 5: CONCLUSION	24
5.1 Summary	24
5.2 Future Work	24
REFERENCES	25-26

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2.3: Feature Importance Score using Random Forest	11
Figure 3.3: Basic Steps of Machine Learning	13
Figure 3.3.1: Decision Tree	15
Figure 3.4: Proposed flow model	17
Figure 4.1: Correlation matrices among different features	18
Figure 4.2: Confusion matrix for XGBClassifier	20
Figure 4.3: Confusion matrix for for AdaBoostClassifier	20
Figure 4.4: Confusion matrix for LogisticRegression	21
Figure 4.5: Confusion matrix for Naive Bayes	21
Figure 4.6: Confusion matrix for Random Forest	22
Figure 4.7: Comparison of classification accuracy	23

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Summary of related works	6
Table 3.1: Features description of the Dataset	8
Table 3.2.2: Missing Values for All Features Statistics	10
Table 3.3.1: Algorithms for supervised and unsupervised learning	14
Table 4.2: Analysis of Performance for various classifiers	22

Chapter 1

Introduction

1.1 Introduction

Everyone's health is seen as vital, and there is a need for a system that keeps track of illnesses and their connections. Patient case summaries, clinic medical records, and other manually maintained data include the bulk of disease-related information. To interpret the texts in them, text mining and machine learning (ML) techniques might be applied. Machine learning is a process for spreading material as part of information retrieval, with an emphasis on the content's semantic and syntactic features. ML and text mining algorithms are introduced and applied for feature extraction and classification. The majority of healthcare professionals use the term "stroke" to describe damage to the brain and spinal cord caused by blood flow abnormalities. Stroke might have a variety of connotations depending on who you ask, but it always provokes a clear bodily reaction. Each person's respiration and movement are supported by brain activity. For more than five decades (since 1970), the number of individuals who die from stroke has been ten times higher in developing nations, and it is expected to quadruple globally by 2030. In general, there are three forms of stroke: Hemorrhagic stroke (HE), Ischemic stroke (IS), and Transient Ischemic Attack (TIA). The most prevalent form of stroke is ischemic stroke. The American Heart Association (AHA) has predicted that 87% of strokes are ischemic strokes, which occur if a clot or an obstacle persists in a blood vessel of the brain. Ischemic stroke has two categories: embolic stroke and thrombotic stroke [1].

Brain function is lost when brain cells die. It is possible that one will be unable to do tasks that are controlled by that area of the brain. A stroke, for example, can damage your ability to move, speak, eat, drink, and swallow, as well as your capacity to see properly, think and recall, and solve issues. Ischemic strokes account for the majority of strokes (87 percent). Blockages caused by blood clots frequently result in ischemic strokes. [2]

1.2 Motivation

According to the American Heart Association (AHA), ischemic stroke accounts for 87 percent of all strokes. [3] In Bangladesh, 26% of persons aged 40 or older have hypertension, and 21.5% have had a stroke. Bangladeshis had the highest risk of strokes among three South Asian countries, according to research (Bangladesh, Sri Lanka, and Pakistan). Furthermore, stroke fatalities accounted for 16.27% of all deaths, ranking Bangladesh 34th in the world. It is long past time for our country's health system to be able to do extensive research that is needed to determine the components that are behind the soaring number of stroke patients in Bangladesh. As a result, an intelligent decision support system based on Machine Learning(ML) techniques would be beneficial in the early diagnosis and prediction of stroke, reducing the severity of the condition.

1.3 Objectives

The objective of this research is to build a model that may be used to predict brain stroke early utilizing machine learning techniques.

1. Create a common model and algorithm to help in the early diagnosis of a stroke in the brain.
2. Using the Extra trees technique and the feature significance score, choose the optimum feature subset.
3. Calculate the prediction accuracy by analyzing the brain stroke data using the provided approaches and other machine learning algorithms.
4. Using various performance evaluation matrices, evaluate the proposed method's prediction performance and compare it to the performance of other methods.

1.4 Expected Outcome

This system assists in the production of an expected outcome in my stroke prediction system depending on the dataset provided. In this strategy, I used 80% of the training to generate more accurate predictions. The training of the dataset determines how accurate the system is. After completing all of the proposed system's necessary procedures, the system was ready to forecast on a real-world dataset. The expected outcome is to predict stroke from real-world datasets with the achieved accuracy of 94% with precision

1.5 Report Layout

Chapter 1 provided an overview of the research, including its motivation, objectives, and expected outcome. Chapter 2 will show the related works, the scope of the problem, and its challenges. The research methodology will be covered in Chapter 3. Results will be found in Chapter 4. The conclusion will be found in Chapter 5.

Chapter 2

Literature Review

2.1 Related Works

Machine learning techniques are frequently employed in the early detection and prediction of a variety of diseases. A number of academics are attempting to establish a link between medical data and machine learning techniques. The following sections go into the specifics of linked works.

N. Kasabov et al. [4] proposed PMeSNNr (Personalized modeling evolving spiking neural network reservoir system) as a novel approach for customized modeling of spectrotemporal data (SSTD) and event prediction. The classification technique is built on spiking neural networks (SNN), this may be used to train and classify SSTD. They compared this model against support vector machine (SVM), multilinear regression (MLR), and multilayer perceptron, which are all classic machine learning methodologies (MLP). Using the PMeSNNr technique, they got the best result with 94 percent accuracy.

On the Cardiovascular Health Study (CHS) dataset, Khosla et al. [5] compared the cox proportional hazards model with machine learning algorithms for the prediction of stroke. They offer a conservative mean algorithm for autonomous feature selection (CM). For stroke prediction, they employed a margin-based censored regression (MCR) and a support vector machine (SVM) learning technique. They acquired 0.777 average tests AUC to predict stroke by combining CM feature selection with MCR, which is by far the highest.

P. Bentley et al. [6] employed a machine learning technique to predict stroke thrombolysis results using computed tomography imaging data (CT). They gathered CT brain scans and clinical domain records from 116 patients who had an acute ischemic stroke and were given intravenous thrombolysis. Their way of evaluation included the support vector machine's (SVM) performance to that of other prognostication tools including HAT and SEDAN scores and found that SVM had the highest AUC score of 0.744.

To predict ischemic stroke, Ahmet K. Arslan et al. [7] employed stochastic gradient boosting (SGB), penalized logistic regression (PLR), and support vector machine (SVM)

approaches. They used their strategies using a medical dataset that only included 80 patients' medical records and 112 healthy people's medical records, each with 17 characteristics. Accuracy, AUC, sensitivity, specificity, positive predictive value, and negative predictive value were their performance evaluation measures, and they used a 10-fold cross-validation procedure. The grid search approach was used to tune the model's parameters. The greatest predictor in terms of accuracy was the SVM model with a score of 96 percent.

B. Letham et al. [8] introduced Bayesian Rule Lists (BRL), which is a rule-based model for predicting stroke that can yield a posterior grouping over permutations of if-then rules. This model was compared to others such as CART, CHADS2-VASc, CHADS2, C5.0, 11 logistic regression, Random Forests, and SVM. With an AUC score of 0.775, BRL is the best performing approach among all. They also stated that the BRL method was comparable to the random forests method in terms of performance. They also assessed performance by running the models on two different groups of data, one for the female patients and the other for the male patients. BRL models outperformed the other models in this case, and the results were consistent with the random forests method's findings.

With the use of administrative data from patients who have an acute ischemic stroke, Sung S-F et al. [9] suggested a process to establish a stroke severity index (SSI). The National Institute of Health Stroke Scale was used to determine the severity of the stroke, according to the researchers (NIHSS). With a correlation value of 0.743, they discovered that the k-nearest neighbor outperforms the others.

Artificial Neural Networks (ANN) were used by D. Shanthi et al. [10] to predict thromboembolic stroke pathology. They employed a clinical dataset of just 50 individuals who were experiencing stroke symptoms. They employed the backward stepwise technique to pick features. Using an ANN-based prediction model, they were able to achieve a prediction accuracy of 89 percent. They also claimed that ANN-based stroke disease prediction improved therapy accuracy and consistency. In table 2.1 a synopsis of the related studies can be found.

Table 2.1: Synopsis of related works

Author's Name	Working Principle	Used ML Algorithm	Performance	Year
P. Bentley et al.	Prediction of stroke thrombolysis result using imaging features based on Computerized Tomography (CT).	Support Vector Machine (SVM)	0.744 (AUC Score)	2014
Ahmet K. Arslan et al.	The medical dataset used to predict ischemic stroke containing 192 records.	Penalized Logistic Regression (PLR), Support Vector Machine (SVM),	96%	2016
D. Shanthi et al.	Thromboembolic Stroke Disease Prediction	Artificial Neural Networks (ANN)	89%	2009

2.2 Scope of the problem

Stroke has become one of the most serious diseases. A brain stroke has claimed the lives of people from all walks of life. Because it is linked to the heart, it is thought to be the most appealing sickness. After doing some research and breaking it down, I've decided on cerebral stroke as the focus of my investigation. The research topic was chosen in retrospect as a high number of people who died as a result of a stroke. Finally, the study has been working on this to develop a better technique that encourages us to lessen the number of people who die in their advanced age groups.

2.3 Challenges

One of the most difficult aspects of forecasting accuracy is gathering data. It is impossible to forecast without data, and it is impossible to predict without data. Then there's preprocessing, which is another problem. The data set has no null values after preprocessing, which aids us in making a decent forecast. Following that, feature scaling aids in putting all feature values on the same value scale. As a result, the suggested design has been subjected to a variety of algorithms. Finally, a technique has been created for obtaining reliable anticipated values. Several obstacles arose as a result of the working method. So, using various machine learning methods, I attempted to raise and improve the model's performance.

Chapter 3 Methodology

3.1 Dataset and Features

There are 11 characteristics and a class variable in the data set that was collected. The research's features and feature descriptions are listed in Table 3.1.

Table 3.1: Features description of the Dataset

No.	Feature Name	Feature Description
1	Age	Age in years
2	Gender	Female: 0 Male: 1
3	Hypertension	No: 0 Yes: 1
4	Heart disease	No: 0 Yes: 1
5	Ever married	True: 1 False: 0
6	Work type	Private: 1 Self-employed: 2 Govt. Job: 3 Children: 4 Never worked: 5
7	Residence type	Urban: 1 Rural: 0
8	Average glucose level	55.1 - 272 mg/dl
9	BMI	10.3 - 97.6
10	Smoking status	Formerly Smoked: 1 Never Smoked: 2 Smokes: 3 Unknown: 4
11	Stroke	No: 0 Yes: 1

Gender: Men are more likely than women to suffer from a stroke. After a certain age, though, women's risk increases. Males account for 41% of the total 5110 cases, while females account for 59%.

Age: Strokes in the brain become more likely as people age. Over the age of 65, approximately 75% of all strokes occur, and the chance of stroke gets doubled after the age of 55.

Heart Disease: Those who have heart disease or who have had a cardiac attack owing to atherosclerosis which is the tightening of the arteries are more likely to have a stroke.

Smoking Status: Smoking raises the risk of cerebrovascular disease, which can lead to an increased risk of stroke. Smoking has been related to a higher risk of ischemic strokes and arachnoid hemorrhage, both of which can result in a brain stroke.

Hypertension: High blood pressure damages the blood arteries in the brain, which can result in a stroke. It is also a significant contributor to the development of a stroke in the brain.

Average Glucose Level: A typical person's blood glucose level is less than 100 mg/dL. Hyperglycemia can occur if the glucose level is more than 126 mg/dL (>6.0 mmol/L) on average. Hyperglycemia can also result in an ischemic stroke.

3.2 Data Preprocessing

3.2.1 Handling Missing Data

Missing values and duplicate values are examined first while processing the data. An incorrect prediction model might be caused by missing values. Missing data can be corrected via data imputation. Missing data imputation approaches include Regression Imputation, Mean Imputation, Expectation-Maximization, and Hot-deck Imputation among others. In the BMI property, some missing values were discovered. The mean imputation approach is employed in this study because it is a frequently used imputation technique that is quick, simple, and straightforward to apply. There are no duplicate values in the dataset. The missing value statistics of all 12 features are shown in Table 3.2.2

Table 3.2.2: Missing Values for All Features Statistics

Feature Name	Number of Missing Values (Percentage %)
Id	0%
Gender	0%
Age	0%
Hypertension	0%
Heart disease	0%
Ever married	0%
Work type	0%
Residence type	0%
Average glucose level	0%
BMI	4%
Smoking status	0%

3.2.2 Data Encoding

In the dataset, there are different types of attributes including both categorical and numerical. For better prediction, the categorical data has been label-encoded to numerical data. Then it will find all the data set as a numerical value. For example, the Work type attribute has values like Private, Self-employed which have been label-encoded to 1 and 2.

3.2.3 Feature Selection

Feature importance is one of the major steps of the Machine Learning model development process. The result of the feature importance score is a set of features along with the statistics of their importance. When the importance of the features is determined, the features can be chosen properly. Using the Random forest, feature importance can be calculated since the average impurity decrease is computed from all decision trees in the forest. Also, this result is independent of the fact whether the data is linear or non-linear. The feature importance score is shown in Figure 3.2.3.

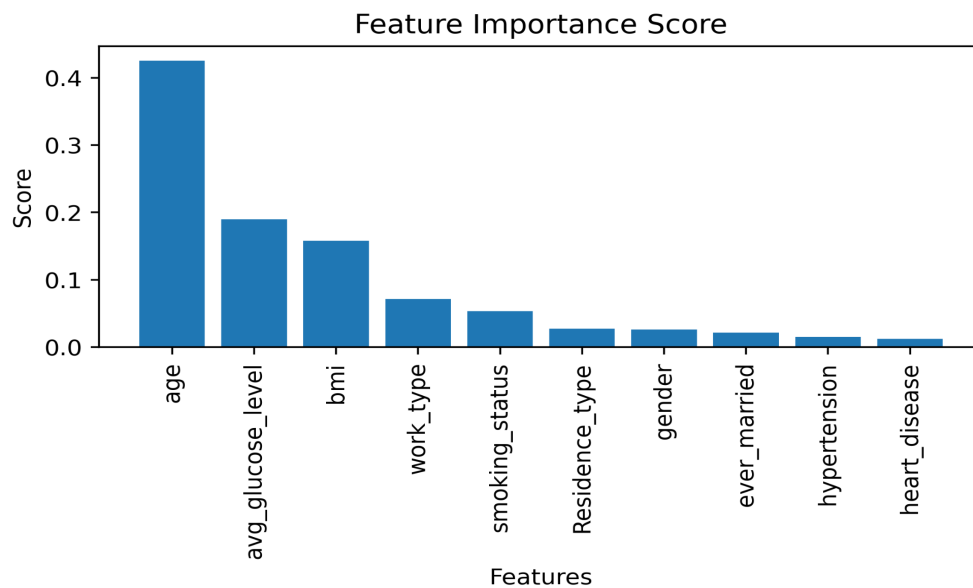


Figure 3.2.3: Random Forest Feature Importance Score

3.2.4 Handling Imbalanced Dataset

While dealing with unbalanced datasets, the problem is that most machine learning approaches will overlook the minority class, resulting in poor performance, despite the fact that performance on the minority class is often the most significant. There is relatively little data in the dataset for qualities with Stroke. As a result, there is an imbalance that might lead to poor model performance in the future. The Synthetic Minority Oversampling Technique(SMOTE), is a kind of data augmentation for the minority population. New instances are synthesized from existing data in SMOTE. To put it another way, SMOTE looks at minority class instances and uses k closest neighbor to find a random nearest neighbor, after which a synthetic instance is constructed in feature space at random. SMOTE is utilized to balance the dataset in this research.

3.2.5 Splitting the Data

Splitting the dataset entails dividing it into two categories: training and testing. Dataset splitting becomes necessary in Machine Learning algorithms to reduce bias in training data. Modifying parameters of an algorithm to perfectly suit the training data usually leads to an overfit algorithm that performs badly on actual test data. For this reason, we partition the dataset into distinct, discrete subsets on which we train various parameters. The split approach is utilized to separate the Training and Test data in this study.

3.3 Research Subject & Instrumentation

Machine learning is capable of making a better choice based on the sample data parameters. The training and testing data are separated from the sample data. These are closely related to statistical approaches and are utilized in machine learning methodologies. They're the result of mathematical optimization. Test data is used in machine learning to evaluate a model's accuracy and performance. Figure 3.3 shows the trivial flow diagram training a model.

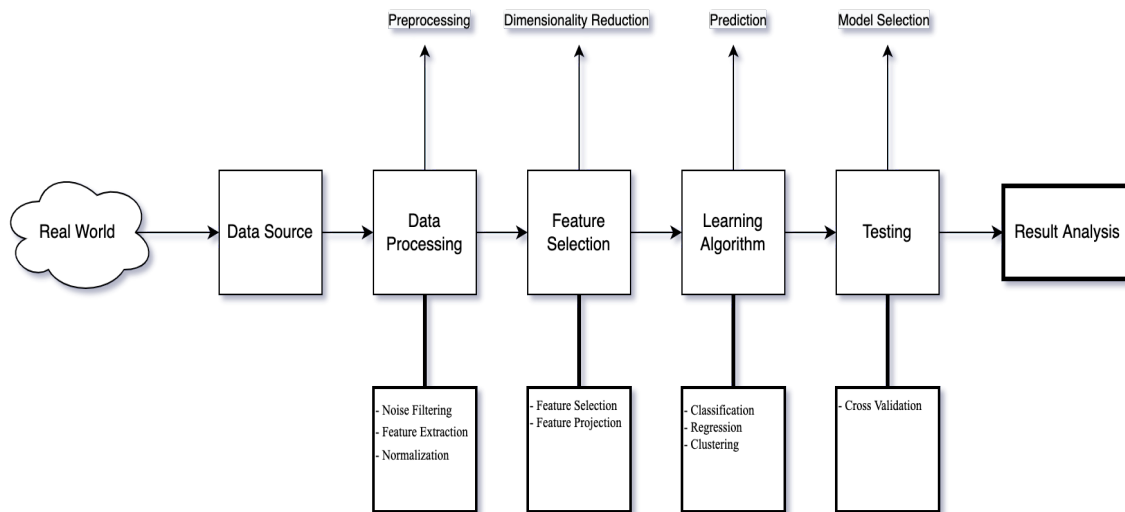


Figure 3.3: Basic Steps of Machine Learning

3.3.1 Classification Algorithms

By evaluating a large quantity of data or constructing prediction models, machine learning methods or classification algorithms may deliver trustworthy findings and learn from previous computations. There are two types of data that are mostly used for machine learning. The two possibilities are unlabeled data which is used for unsupervised learning and labeled data which is used for supervised learning. Machine learning algorithms employ a variety of supervised and unsupervised learning approaches, as shown in Table 3.3.1

Table 3.3.1: Algorithms for supervised and unsupervised learning

Supervised Learning	Unsupervised Learning
Naïve Bayes	Fuzzy C- Mean Clustering
Logistic Regression	K-Mean Clustering

Decision Tree	Self-organizing Map
K-Nearest Neighbor (KNN)	Hierarchical Clustering
Support Vector Machines (SVM)	
Random Forest	

The following are some of the most common supervised learning approaches.

3.3.1.1 Decision Tree

Among other supervised learning approaches Decision Tree is an approach that may be used to solve classification and regression problems, however, it is most often used to solve classification problems. In this tree-structured classifier, internal nodes hold dataset properties, branches provide decision rules, and each leaf node offers the conclusion. There are two nodes in a Decision Tree, the Decision Node and the Leaf Node. Leaf nodes are the result of those decisions and they do not include any more branches. On the other hand, Decision nodes are used for making a decision and have several branches. A decision tree's general structure is seen in the diagram below.

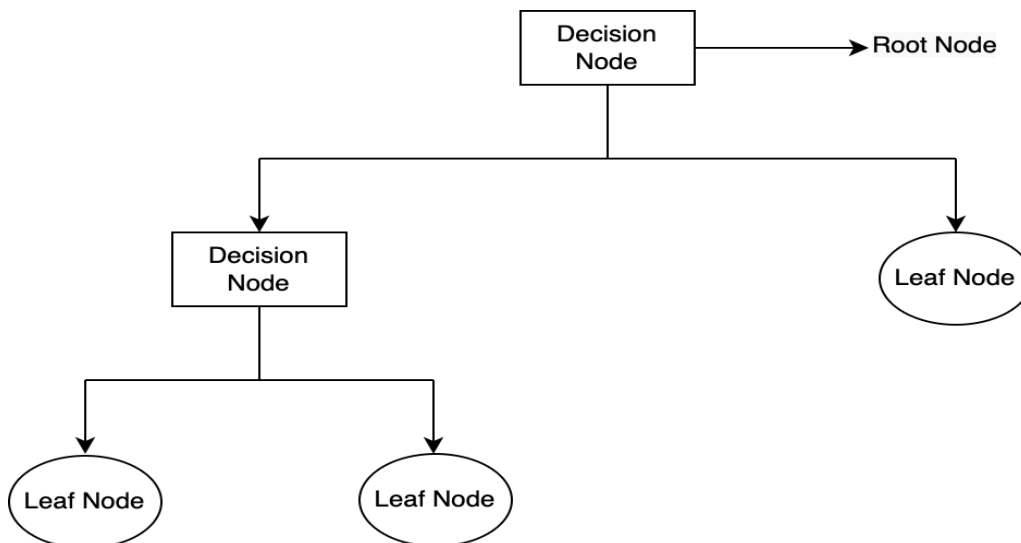


Figure 3.3.1: General Structure of Decision Tree

3.3.1.2 Random Forest

Another supervised machine learning algorithm, generally used to solve classification and regression problems is Random forest. It generates decision trees from several samples, using the majority vote for classification and the average for regression. One of the most important characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. For classification problems, it produces better results.

3.3.1.3 Naïve Bayes

The supervised machine learning algorithm or classifier Naive Bayes is well-known. To categorize data, Nave Bayes uses the Bayes theorem, assuming that the probability of one characteristic A is completely independent of the likelihood of another attribute B. The Bayes theorem is a theory that explains how to determine the probability of a hypothesis based on prior information.

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

The likelihood is the probability of predictor given class, here posterior means the posterior probability of class/target given the predictor. Evidence is the prior probability of a prediction, whereas Prior means the prior probability of a class.

3.3.1.4 K-Nearest Neighbor

The distance function is used by K-Nearest Neighbors. A class is categorized by a majority vote of its neighbors. For estimating the distance to the nearest neighbor, the Manhattan, Minkowski, Euclidean, and Hamming distance formulae are usually utilized.

The K Nearest Neighbor model may be implemented using the methods below.

1. Load the dataset you've gathered.
2. Initialize the k value
3. Do the following where $n = 1$ to the sum of the training dataset,
 - Measure the distance between each row of the training dataset and the test data.

- For that use any distance measuring formula
- Sort the estimated distances by distance values in ascending order measured earlier
- Take the first k rows of the sorted array
- Determine the most common class of those rows and return the predicted class

3.4 Proposed Model

The following is a description of the work's suggested model (see Figure 3.4):

- Handle missing data and balance the data set
- Using the feature importance score, choose the most important features.
- Using split techniques, split the dataset
- Analyze the dataset by using several machine learning and data mining techniques for improved accuracy and prediction.

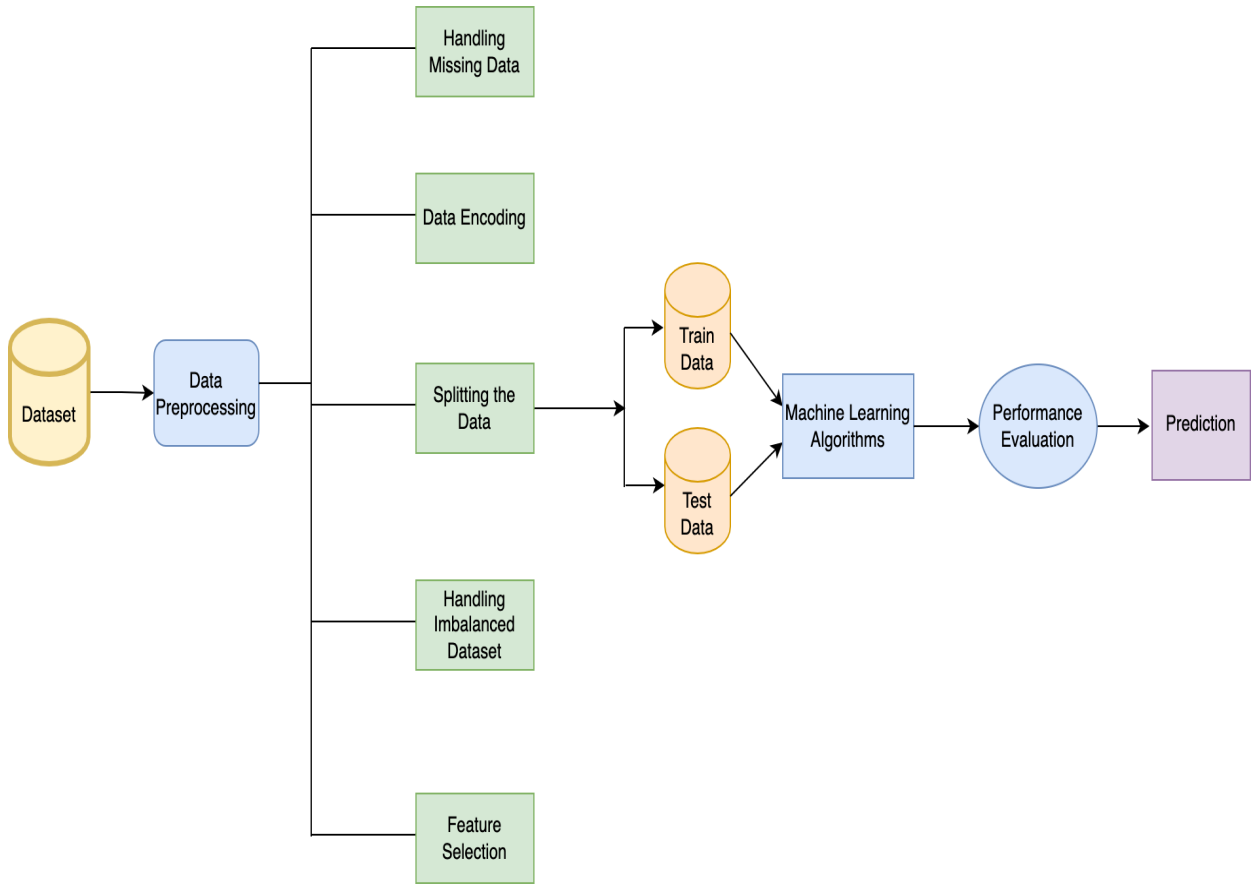


Figure 3.4: Flow diagram of the proposed model

Chapter 4 Results

4.1 Correlation Results

The Pearson correlation results reveal the impact of feature qualities on target attributes. The relationship between stroke qualities and other properties is depicted in Figure 4.1. As can be seen from the graph, no one feature has a significant impact on stroke. Age, hypertension, heart disease, avg glucose level are among the metrics that have a significant impact on stroke. The least effective factors are work type, residence type.

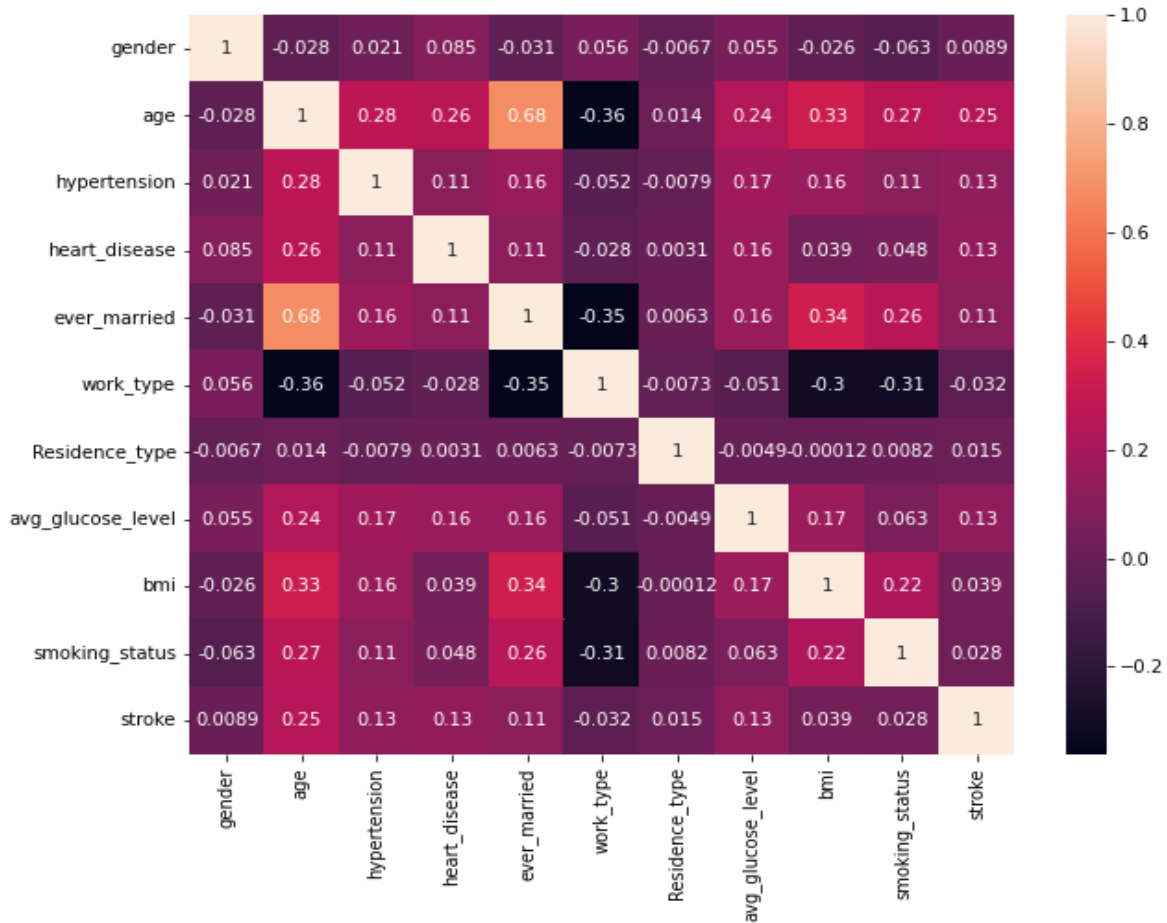


Figure 4.1: Correlation matrices among different features

4.2 Experimental Results

The classification and prediction models are evaluated with the use of four evaluation metrics. From the dataset, stroke patients and non-stroke patients have been labeled as the Predicted No Stroke and the Predicted Stroke class respectively. The following are some equations of performance evaluation metrics:

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

Where, TP , FP , FN , and TN represent True Positive, False Positive, False Negative, and True Negative respectively.

By using seven machine learning classifiers called Naive Bayes, Logistic Regression, AdaBoost Classifier, XGB Classifier, K-Nearest Neighbors, Decision Tree, and Random Forest we have trained. The classification tasks were done by using Python. From the dataset, 80% of the data was used for training purposes and the rest of the data was used for validation and testing. K-fold(K=10) cross-validation is used to estimate the accuracy.

In the analysis of the accuracy of these classification algorithms, the accuracy of the Random Forest classifier is the highest and the value is 94.704% and it performs better than any other classifier

such as Naive Bayes(79.228%), Logistic Regression(80.154%), AdaBoost Classifier(85.244%), XGB Classifier(87.197%), K-Nearest Neighbors(89.820%), Decision Tree(91.773%).

The confusion matrix of XGBoost Classifier, Ada Boost Classifier, Logistic Regression(LR), Naive Bayes, and Random Forest are shown in Figure 4.2 to Figure 4.6 respectively.

The result of the experiments is displayed in Table 4.2. Also, Figure 4.7 shows the comparison graph of the classification accuracy of all the classifiers.

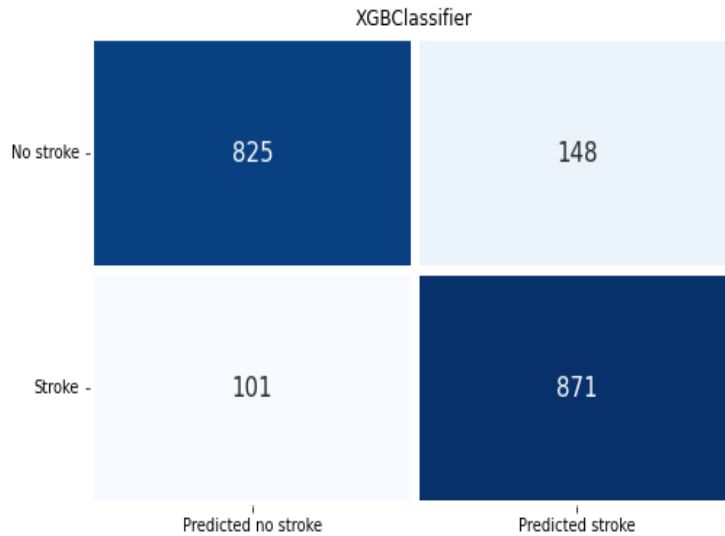


Figure 4.2: Confusion matrix for XGBoostClassifier

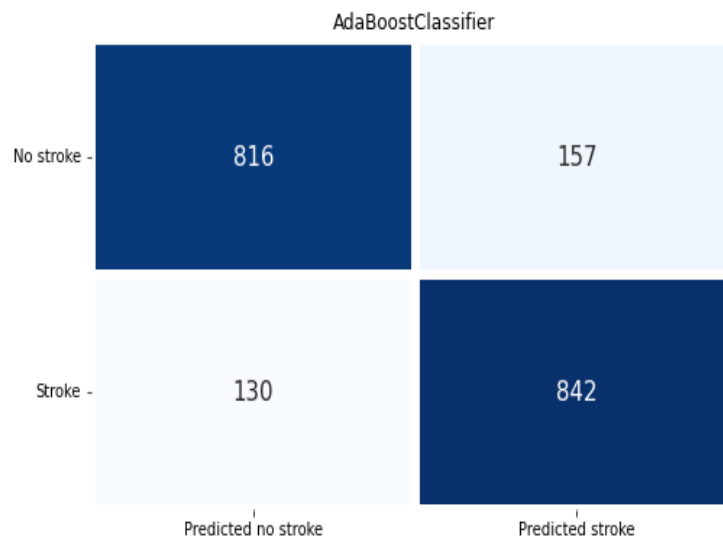


Figure 4.3: Confusion matrixfor for AdaBoostClassifier

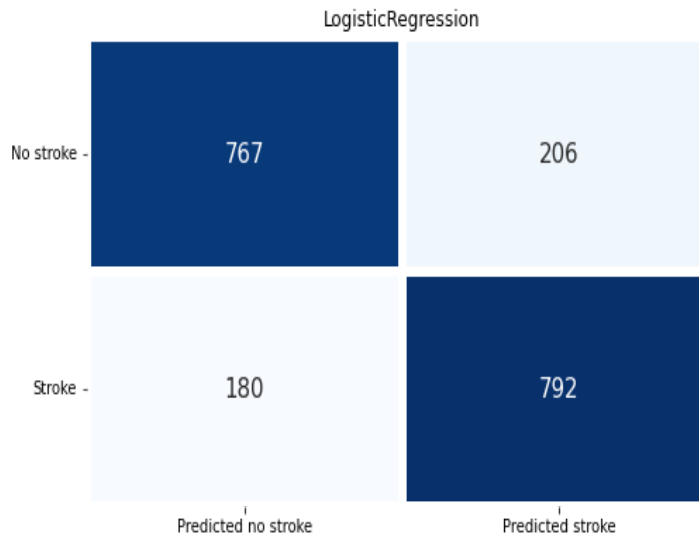


Figure 4.4: Confusion matrix for LogisticRegression

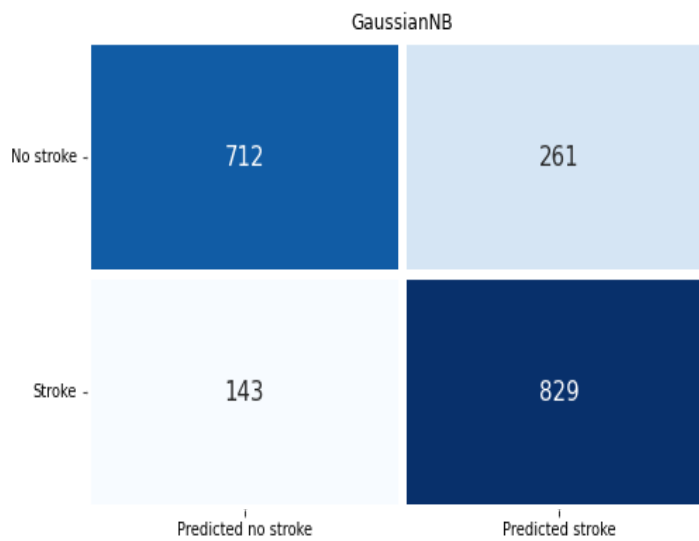


Figure 4.5: Confusion matrix for Naive Bayes

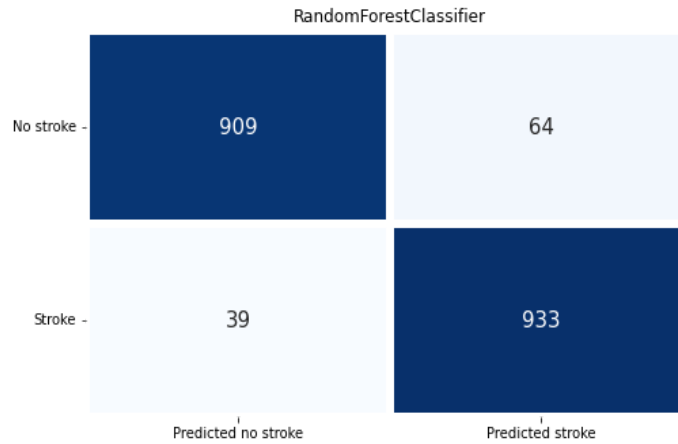


Figure 4.6: Confusion matrix for Random Forest

Table 4.2: Analysis of Performance for various classifiers

Model	Accuracy	K-Fold Mean Accuracy	Std. Deviation	ROC_AUC	Precision	Recall	F1 Score
Random Forest	94.704	93.827787	0.472856	0.947050	0.935807	0.95987	0.94768
Decision Tree	91.773	90.073002	0.954405	0.917743	0.909274	0.92798	0.91853
K-Nearest Neighbors	89.820	88.877083	1.302270	0.898242	0.843085	0.97839	0.90571
XGB Classifier	87.197	86.884034	1.157278	0.871992	0.854760	0.89609	0.87493
AdaBoost Classifier	85.244	83.219174	1.577011	0.852449	0.842843	0.86625	0.85438
Logistic Regression	80.154	80.030587	1.521142	0.801549	0.793587	0.81481	0.80406
Naive Bayes	79.228	78.744793	1.570666	0.792319	0.760550	0.85288	0.80407

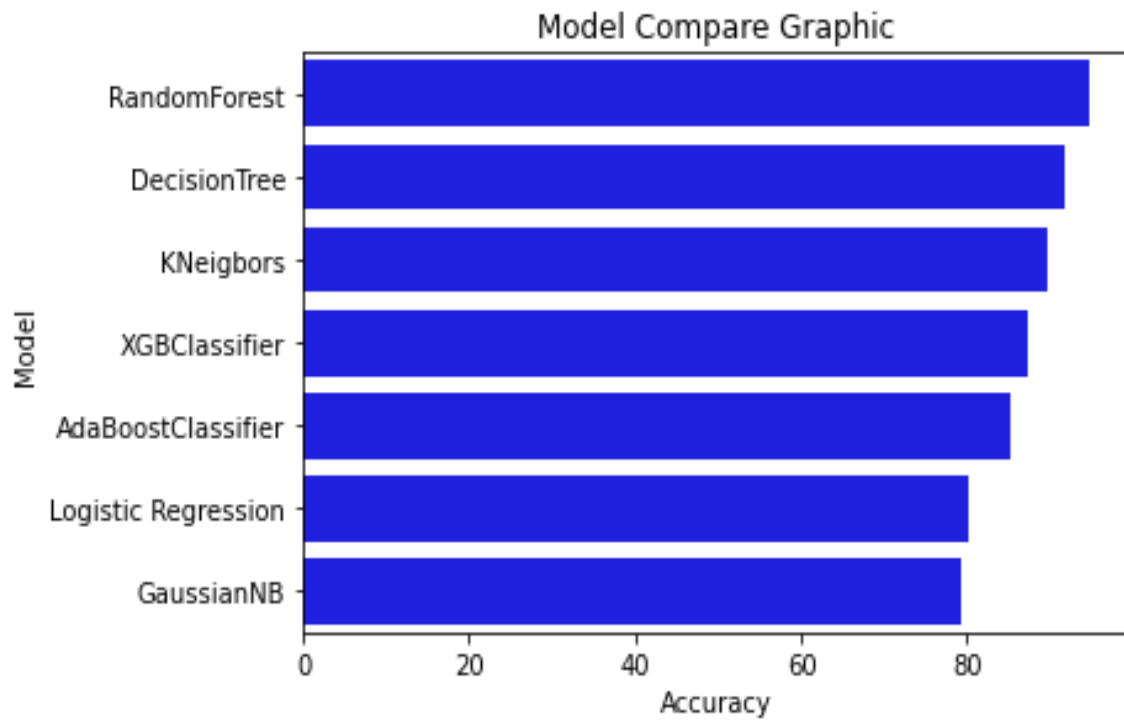


Figure 4.7: Comparison of classification accuracy

Chapter 5

Conclusion

5.1 Summary

To detect the presence of a stroke in a person, the suggested study effort utilized seven classifiers. To predict stroke, the suggested Random Forest classifier took into account gender, age, hypertension, heart disease, average glucose level, BMI, and smoking status feature characteristics. In comparison to the other machine learning algorithms, the Random Forest classifier offered the greatest accuracy of around 94 percent, according to the performance evaluation. As a result, the Random Forest classifier is a viable option for stroke prediction. The link between these disorders and the risk of having a stroke in a human being has been investigated. The characteristics employed in this study are thought to be important in the early therapeutic treatment of stroke patients. So, if this condition is effectively detected and managed from the start, it can help to lower the risk of stroke in our lives.

5.2 Future Work

For future work, I would like to focus on a few things. One would be preparing a repository database based on stroke patients of Bangladesh. Another objective is to create an easy-to-use internet or mobile-based program for relative clinical outcome prediction based on demographics and lifestyle.

References

1. Pahus SH, Hansen AT, Hvas AM (2016) Thrombophilia testing in young patients with Ischemic stroke. *Thromb Res* 137:108–112
2. Types of stroke (2021) Cdc.gov. Available at: https://www.cdc.gov/stroke/types_of_stroke.htm (Accessed: December 10, 2021).
3. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart disease and stroke statistics—2021 update: a report from the American Heart Association external icon. *Circulation*. 2021;143:e254–743
4. Nikola Kasabov, Valery Feigin, Zeng-Guang Hou, Yixiong Chen, Linda Liang, Rita Krishnamurthi, Muhaini Othman, Priya Parmar, “Evolving spiking neural networks for personalized modeling, classification and prediction of spatio-temporal patterns with a case study on stroke”, *Neurocomputing*, Volume 134, Pages 269-279, 2014.
5. 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.
6. 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.
7. Arslan AK, Colak C, Sarihan ME, “Different medical data mining approaches based prediction of ischemic stroke”, *Computer Methods and Programs in Biomedicine*, 2016.
8. 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.

9. 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.

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

Plagiarism Report

Stroke Prediction Using Machine Learning Techniques

ORIGINALITY REPORT

30% SIMILARITY INDEX	21% INTERNET SOURCES	20% PUBLICATIONS	17% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	dspace.uiu.ac.bd Internet Source	9%
2	link.springer.com Internet Source	1%
3	Submitted to Ghana Technology University College Student Paper	1%
4	Md. Azizul Hakim, Md. Zahid Hasan, Md. Mahabur Alam, Md. Mehadi Hasan, Mohammad Nurul Huda. "An Efficient Modified Bagging Method for Early Prediction of Brain Stroke", 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2019 Publication	1%
5	Ferdib-Al-Islam, Mounita Ghosh. "An Enhanced Stroke Prediction Scheme Using SMOTE and Machine Learning Techniques", 2021 12th International Conference on	1%