

**EARLY PREDICTION OF BRAIN STROKE USING MACHINE
LEARNING TECHNIQUES**

BY

**SHARZAN RAHMAN
ID: 191-15-12451**

AND

**AFSANA ALAM MIM
ID: 191-15-12234**

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

Supervised By

Mst. Eshita Khatun
Lecturer(Senior Scale)
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled "**EARLY PREDICTION OF BRAIN STROKE USING MACHINE LEARNING TECHNIQUES**", submitted by Sharzan Rahman, ID No: 191-1512451 & Afsana Alam Mim, ID No: 1991-15-12234 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 January 23, 2023.


BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan

Professor and Head

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



Internal Examiner

Dr. Md. Zahid Hasan

Associate Professor

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

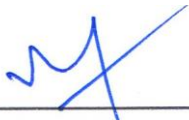


Internal Examiner

Fahad Faisal

Assistant Professor

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



External Examiner

Dr. Ahme Wasif Reza

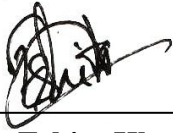
Associate Professor

Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mst. Eshita Khatun, Lecturer (Senior Scale), 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:

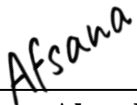


Mst. Eshita Khatun
Lecturer (Senior Scale),
Department of CSE
Daffodil International University

Submitted by:



Sharzan Rahman
ID: 191-15-12451
Department of CSE
Daffodil International University



Afsana Alam Mim
ID: 191-15-12234
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for Her divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mst. Eshita Khatun, Lecturer (Senior Scale)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. Her 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 project.

We would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan, Professor, and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

Brain stroke is the second-leading cause of death and the third-leading cause of disability worldwide. A stroke occurs when the blood circulation in the brain is obstructed or when a blood vessel in the brain ruptures and leaks. A stroke is a medical emergency that must be treated as soon as possible. Early intervention can help to prevent brain damage and other complications. Machine learning and data science play an important role in medical science. Using technology, we can predict a disease based on the symptoms of the human body. In this paper, we propose an intelligent system that can predict potential brain strokes with only twenty-three (23) features. In addition, we apply six (10) well-known machine learning algorithms to Bangladeshi datasets collected from various hospitals in Bangladesh to assess prediction accuracy. In our work, the accuracy of gradient boosting classification is 96.09%, and it is consistent. Gradient Boosting's accuracy is higher than other classifiers such as Random Forest, Bagging, Logistic Regression, SVM, K Neighbors, Decision Tree, Gaussian Naïve Bayes, XG Boost, and Ada Boost. We have a data shortage because we only collected data from 385 patients. We can get a better result if we can manage more patients' data.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 The rationale of the study	2
1.4 Research Questions	3
1.5 Expected Output	3-4
1.6 Layout of the report	4
CHAPTER 2: BACKGROUND STUDY	5-12
2.1 Introduction	5
2.2 Related Works	5-10
2.3 Research Summary	11
2.4 Challenges	12
CHAPTER 3: METHODOLOGY	13-28
3.1 Introduction	13

3.2 Dataset	13
3.2.1 Dataset Features	13-20
3.3 Machine Learning Algorithms	20-27
3.3.1 Logistic Regression	21
3.3.2 Gaussian Naïve Bayes	21-22
3.3.3 K Neighbors	22-23
3.3.4 Random Forest	23
3.3.5 Decision Tree	23-24
3.3.6 XG Boost	24-25
3.3.7 SVM	25
3.3.8 Ada Boost	26
3.3.9 Bagging	26-27
3.3.10 Gradient Boosting	27
3.4 Proposed Method	28
3.5 Working Procedure	29
CHAPTER 4: RESULTS ANALYSIS	30-39
4.1 Experimental Results Analysis	30-39
CHAPTER 5: CONCLUSION	40-41
5.1 Summary	40
5.2 Discussion	40
5.3 Future Work	40
5.4 Conclusion	40-41

APPENDIX

REFERENCES

42-43

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Proposed Model Method	28
Figure 3.2: Working Procedure	29
Figure 4.1: Comparison bar graph of precision	35
Figure 4.2: Comparison bar graph of Recall	36
Figure 4.3: Comparison bar graph of F1 Score	37
Figure 4.4: Comparison bar graph of Accuracy	38
Figure 4.5: Comparison of Different Classifiers	39

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Research Paper Summary	11
Table 3.1: Dataset Features	14-15
Table 4.1: Confusion Matrix	30
Table 4.2: Confusion Matrix of Decision Tree	32
Table 4.3: Confusion Matrix of Random Forest	32
Table 4.4: Confusion Matrix of Random Forest	32
Table 4.5: Confusion Matrix of K Neighbors	32
Table 4.6: Confusion Matrix of Ada Boost	33
Table 4.7: Confusion Matrix of Gaussian Naïve Bayes	33
Table 4.8: Confusion Matrix of SVC	33
Table 4.9: Confusion Matrix of Logistic Regression	33
Table 4.10: Confusion Matrix of XG Boost	33
Table 4.11: Confusion Matrix of Gradient Boosting	33
Table 4.12: Performance Analysis of Different Machine Learning Classifiers	34

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cerebral stroke, also known as cerebrovascular accident (CVA), is a serious health issue that happens when the flow of blood to the brain is interrupted, resulting in damage to brain cells. Brain strokes are a major cause of death and disability globally, and early detection and therapy of the condition are crucial for improving the chances of recovery and reducing the risk of long-term complications. Traditionally, the diagnosis of brain stroke has relied on the evaluation of clinical signs and symptoms and the use of diagnostic tests like imaging techniques such as CT scans or MRIs may be used in the treatment of strokes. However, these methods can be time-consuming and may not always provide a definitive diagnosis. In addition, brain stroke can have a range of different causes and presentations, making it challenging to accurately predict the condition. Recently, machine learning algorithms have gained increasing attention as a potential tool for predicting brain stroke. Machine learning algorithms are computer programs that are designed to learn from data and improve their performance over time. They can analyze large amounts of data and identify patterns that may not be apparent to human analysts. Therefore, the use of machine learning algorithms for predicting brain stroke could lead to more accurate and reliable predictions and facilitate earlier detection and treatment of the condition. In this research, we aim to investigate the apply machine learning algorithms to predict brain strokes and evaluate their performance compared to traditional methods. We will collect data from patients with brain stroke from multiple hospitals and apply different machine learning algorithms to the data to develop a prediction model. We will then assess the model's performance using various evaluation metrics and compare the result to traditional methods. The results of this study will have the potential to improve the early identification and treating brain strokes and reduce the risk of long-term complications.

1.2 Motivation

Brain stroke is a stroke disease that everyone cannot understand for instance. For this reason, we have to go to the hospital and check for a which stroke that has occurred. From our research, we have gotten – Early detection and treatment of brain stroke are crucial for improving the chances of recovery and reducing the risk of long-term complications. A machine learning model for brain stroke prediction could help doctors identify the condition in its early stages and initiate appropriate treatment more quickly. Machine learning algorithms have the potential to analyze large amounts of data and identify patterns that may not be apparent to human analysts. This could enable more accurate and reliable predictions of brain stroke. The development of a brain stroke prediction model using machine learning could lead to improved efficiency in the healthcare system by reducing the need for unnecessary testing and enabling more targeted use of resources. Working on brain stroke prediction using machine learning could contribute to the overall advancement of the field of artificial intelligence and its application in healthcare. Finally, the creation of a machine learning model for predicting brain strokes could have significant societal benefits, as strokes are a major cause of disability and death globally. These are some reasons for us to work on brain stroke.

1.3 The rationale of the study

Brain stroke is a serious medical condition that requires prompt treatment to lower the risk of long-term problems, and increase the chances of recovery. The ability to predict brain stroke in its early stages could lead to earlier detection and treatment of the condition. Machine learning algorithms have the potential to analyze large amounts of data and identify patterns that may not be apparent to human analysts. This could enable more accurate and reliable predictions of brain stroke. The development of a brain stroke prediction model using machine learning could lead to improved efficiency in the healthcare system by reducing the need for unnecessary testing and enabling more targeted use of resources. Studying brain stroke prediction using machine learning could contribute to the overall advancement of the field of artificial intelligence and its application in healthcare.

1.4 Research Questions

- a) Which machine learning algorithms are the best at predicting brain strokes?
- b) How do machine learning models for predicting brain strokes compare to traditional methods in terms of accuracy?
- c) Which factors have the greatest impact on the accuracy of machine learning algorithms in predicting brain strokes?
- d) What can be done to enhance the performance of machine learning models for predicting brain strokes?
- e) To what extent do the results of machine learning models for predicting brain strokes vary among different patient groups and healthcare settings?
- f) What ethical considerations should be taken into account when using machine learning algorithms for predicting brain strokes?
- g) What is required for the implementation of machine learning models for predicting brain strokes in clinical settings?
- h) How do the predictions of machine learning models for brain strokes compare to those made by human experts?
- i) What can be done to increase the interpretability of machine learning models for predicting brain strokes?
- j) What are the possible long-term effects of using machine learning algorithms for predicting brain strokes on patient outcomes and the healthcare system?

1.5 Expected output

For our study, we will get these output after the experimental results:

- a) An evaluation of the accuracy, precision, and recall of various machine learning algorithms for predicting brain strokes.
- b) An understanding of the factors that have the greatest influence on the accuracy of machine learning algorithms in predicting brain strokes, such as age, medical history, and lifestyle choices.

- c) Suggestions for enhancing the performance of machine learning models for predicting brain strokes, such as choosing particular algorithms or incorporating additional features in the training data.
- d) An examination of the possible advantages and difficulties of using machine learning models for predicting brain strokes in clinical settings.
- e) A contrast of the predictions made by machine learning models for brain strokes with those made by human experts.

1.6 Layout of the report

- i. Chapter 1 is all about the Introduction of this research work.
- ii. In chapter 2, there will be a Background Study of this work.
- iii. In chapter 3, there will be Research Methodology.
- iv. In chapter 4, there will be Results Analysis.
- v. In chapter 5, there will be the Conclusion and Future Work of this thesis work.

CHAPTER 2

BACKGROUD STUDY

2.1 Introduction

This section will go over related works, research summaries, and research challenges. This section on previous research covers other papers and their findings, methods, and accuracy that are related to our work. We will provide an overview of these related works in the research summary section. In the challenges section, we will discuss how we enhanced our accuracy.

2.2 Related work

The authors of [1] proposed a method for detecting and classifying intracranial hemorrhage strokes using a microwave imaging system (MIS) and machine learning (ml). They demonstrated a system for detecting and localizing hemorrhagic strokes in a layered human head phantom. As a dataset, they used images of healthy and unhealthy brain tissue. To put this dataset to the test, they created a human-like head system. Their proposed method produced 97% accurate results on a core i7 processor computer with 12 extracted feature numbers. It is only effective against cerebrovascular targets.

To predict brain stroke, the authors of [2] proposed an improved random forest machine learning algorithm. They gathered the data from NIHSS-compliant medical records. The dataset includes 4799 patients, 3128 of whom are male and 1676 of whom are female. Gender, blood pressure, glucose levels, age, paralysis, smoking, BMI, cholesterol, and stroke record were among the dataset's characteristics. ML algorithms such as Gaussian Naïve Bayes, K-Means, Linear Regression, Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression were used. They have also employed the random forest and Ada boost algorithms. Then, in order to improve accuracy, they proposed an improved random forest model. With an overall accuracy of 94.23%, the results were as follows: decision tree (93.12%), Naive Bayes (76.77%), logistic regression (82.5%), linear SM (23.22%), poly SM (83.43%), and PBG SVM (81.97%). As a result, they created a

model called "improvised random forest," which had a 96.97% accuracy rate. And this model had a 0.03% error rate. They hoped to propose derived methods for various types of strokes from an image dataset in the future.

The authors of [3] used machine-learning approaches to predict brain stroke and compared them to the Cox model. They gathered data through medical checks, questionnaires, and phone calls. They made use of 16 dataset characteristics. They combined the performance matrix and some feature selection, as well as forward feature selection, L1 regularized logistic regression, and conservative mean feature selection. As machine learning algorithms, they employed supported vector machines and margin-based censored regression. With a concordance index of 0.770, using a combination of CM feature selection and MCR for prediction resulted in the best outcomes. The Conservative definition performs very well in feature selection for the CH'S dataset. However, feature selection method may not function well in other datasets. This strategy may be used to uncover possible risk factors for illnesses without having to conduct clinical trials.

The authors of [4] proposed an automated system for early detection of ischemic strokes using a CNN deep learning algorithm. For this strategy, they employed data augmentation. The collection includes CT pictures of brain strokes. They employed a total of 256 patch images with a 32x32 picture size. They divided the data in half, using half for training and half for testing. To obtain additional data, they employed data augmentation. They used the shift of affine transformation approach, which shifts one point at a time, to collect more meaningful data. CNN has also been utilized in machine learning. The training stage has an accuracy rate of 97.66%, while the testing stage has an accuracy rate of 92.969%. As a result, CNN's identification rate exceeds 90%. They want to gather additional brain stroke photos in the future to improve the overall system identification rate.

The authors of [5] suggest a ten-classifier machine-learning technique for stroke prediction. They gathered the information from medical facilities in Bangladesh. They have acquired information from around 5110 people. Age, gender, hypertension, employment type,

housing type, heart disease, age, glucose level, BMI, marital status, smoking habits, and stroke history were all factors. They employed following machine learning classifiers: Linear Regression, SGD, DTC, Ada Boost, Gaussian Naïve Bayes, QDA, MLP, K Neighbors, GBC, and XG Boosting. Testing data is 20% and training data is 80%. To obtain a better result, they create weighted voting models. Weighted voting had the best accuracy (97%), GBC and XGB had the second highest (96%), and the SGD classifier had the lowest accuracy (65%). When compared to other machine learning algorithms, weighted voting produced the greatest results in this study. In the future, they hoped to focus on deep learning-based imaging.

Four machine learning techniques were utilized by the authors of [6] to diagnose stroke disease. Their key contributions were data collection and preparation for 77 the dataset. They utilized WEKA for their project. They gathered 1058 individual patient records. There were 412 male patients and 646 female patients from 1059 onwards.

The authors of [7] used eight machine learning techniques to identify strokes illnesses using a image dataset of CT scan. For the picture processing, they employed data augmentation. They obtained CT scan data from 102 patients at Hajj Hospital in Surabaya, Indonesia. They had a total of 233 CT scan pictures, 226 of which were ischemic stroke data and 7 of which were hemorrhagic stroke data. They had gone through six processes to process the photograph. They have INN, Gaussian Naïve Bayes, Logistic Regression, Random Forest, NN-MLP, Decision Tree, Deep Learning, and SVM weighted ml algorithms. From above this, Random Forest achieved the highest accuracy of 95.67% among these algorithms.

The authors of [8] proposed a machine learning-based approach for detecting brain strokes. They demonstrated a system architecture in which they constructed a model utilizing ml techniques as SVM, Random Forest, Decision Tree, XG Boost, and SUD. The dataset will be run through these models, yielding prediction results. They also add people to the dataset upload. Users may verify their stroke forecast by utilizing this system.

Author's of [9] have proposed predicting stroke outcomes using NLP-based machine learning on the radiology report of a brain MRI. They have collected the data of 1840 acute ischemic stroke (A75) patients, among whom 646 had an outcome. They used a poor MPI text report as a dataset when they were first admitted. They had split the dataset into 70 and 30 reactions for training and testing, respectively. They used ML-based natural language processing along with deep learning, CNN & LSTM, CNN Max, and Multi-CNN. Among these MI algorithms, Multi-CNN obtained the best result. Deep learning showed superior performance over machine learning methods.

The author of [10] used machine learning approaches to investigate Acute Ischemic Stroke Neuroimaging. Unsupervised machine learning was utilized. Neuroimaging characteristics were used to train the dataset. They have a dataset size limitation. Because neuroimaging-based deep learning requires a large quantity of data, which is not available. Another constraint was the requirement to label each photograph. In the future, they hoped to work with numerous universities to create a robust dataset.

Authors of [11] have proposed & three machine learning methods to detect strokes within 4.5 hours, they have used MRI for this work. They have collected 1836 stroke patients' data. They have finally collected 355 stroke patients data from 1830 patients using diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR). They analyzed DW7 and FLAIR images. Then they applied three machine-learning methods to this dataset. ML classifiers were Logistic regression, SVM, and random forest. Among these methods, logistic regression and random forest both obtained the same result, the best result. For future work, they wanted to evaluate the applicability of these ML algorithms to other patients.

The paper "Deep into the brain: artificial intelligence in stroke imaging" by Lee, Eun-Jae et al. (2017) presents an overview of the use of artificial intelligence (AI) in the imaging of stroke. The authors provide a comprehensive review of current AI-based methods for stroke imaging, including their applications, limitations, and future directions. The authors

first introduce the different types of stroke and the imaging modalities used for diagnosis and treatment. They then discuss the various AI-based methods for stroke imaging, including deep learning, machine learning, and computer-aided diagnosis (CAD) systems. They also review the use of AI-based methods for the analysis of imaging data, such as magnetic resonance imaging (MRI) and computed tomography (CT) scans. The authors found that AI-based methods have the potential to improve the diagnostic accuracy, efficiency, and automation of stroke imaging. These methods can also provide valuable information for treatment planning and follow-up. However, the authors also note that there are limitations to the current AI-based methods, such as the lack of large and diverse datasets and the need for further validation in clinical settings. The authors suggest that the future directions for AI-based methods in stroke imaging include the development of more sophisticated algorithms and the integration of multiple imaging modalities. They also suggest the importance of data sharing and collaboration among researchers to improve the performance of AI-based methods.

The paper "Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke" by Kim, Chulho et al. (2019) presents a method for automatically identifying acute ischemic stroke (AIS) in brain MRI reports using natural language processing (NLP) and machine learning techniques. The authors aim to improve the efficiency and accuracy of the diagnostic process for AIS by automating the identification of relevant information in brain MRI reports. The authors first preprocessed the data by extracting the text from the reports and removing irrelevant information. They then applied NLP techniques to extract features such as named entities, negation, and modality. These features were used as input to various machine learning algorithms, including support vector machines (SVMs), random forests, and gradient boosting. The authors found that the combination of NLP and machine learning techniques was able to accurately identify AIS in brain MRI reports. Their best performing model, a gradient boosting algorithm, achieved an F1-score of 0.94. This indicates that the model was able to accurately identify AIS in the majority of the cases. The authors also performed experiments to evaluate the effect of different feature sets and different machine learning

algorithms on the performance of the model. They found that the gradient boosting algorithm performed the best overall, and that the inclusion of negation and modality features improved the performance of the model. The paper provides a promising approach to automating the identification of AIS in brain MRI reports. The use of NLP and machine learning techniques allows for the efficient and accurate identification of relevant information in the reports, which can improve the diagnostic process for AIS. However, more research is needed to evaluate the performance of the model on larger and more diverse datasets.

2.3 Research Summary

We are provided a brief synopsis of a research article pertaining to our work.

Table 2.1: Research paper summary

Serial	References	Used Methodology	Results
01	The author of paper [2]	Improvised Random Forest	94.23%
02	The author of paper [3]	Machine Learning vs Cox Model	0.770
03	The author of paper [4]	CNN Deep Learning	Above 90%
04	The author of paper [5]	Eight machine learning classifiers	95.67%
05	The author of paper [7]	Machine Learning Classifiers	95.97%
06	The author of paper [12]	CT Scan Images	0.805
07	The author of paper [13]	Brain MRI Images	0.744

A stroke is a health issue that happens when the flow of blood to the brain is interrupted. This can be due to a blockage in the blood vessels that supply the brain (ischemic stroke) or bursting of a blood vessel in the brain (hemorrhagic stroke). When brain does not receive enough blood and oxygen, the cells in the affected area can become damaged or die. The symptoms of a stroke can vary widely depending on which part of the brain is affected and how severe the damage is. Possible symptoms include weakness or numbness on one side of the body, difficulty speaking or understanding speech, confusion, vision problems in one or both eyes, difficulty walking or loss of balance or coordination, and a severe headache with no known cause. Treatment for a stroke may include medications to dissolve blood clots or stop bleeding, surgery to repair damaged blood vessels, and rehabilitation to help the person recover function and independence.

2.4 Challenges

One of the major obstacles in predicting accuracy is data collecting. It cannot predict without data. Preprocessing is the next obstacle. After preprocessing, the data set does not contain any null values, which allows to make an accurate prediction. After that, feature scaling helps to standardize all feature values to the same scale. As a result, a different algorithm has been applied to the suggested design. Lastly, the application procedure had created in order to obtain accurate anticipated value. According to the working technique, numerous obstacles arose.

CHAPTER 3

METHODOLOGY

3.1 Introduction

In our work, we attempted to collect the data from several medical hospitals in Bangladesh. We also attempted to make this study distinctive by making correct predictions. We discovered several missing values after collecting data, which we corrected. We finished the feature scaling procedure to acquire the correct forecast. Datasets are used for training and testing algorithms such as Decision Tree, Logistic Regression, k Neighbors, support vector machines (SVM), Gaussian Naive Bayes, Ada Boost, XG Boost, Bagging, and Gradient Boosting.

3.2 Dataset

Dataset is collected from different hospitals in Bangladesh. Dataset contains 385 patients data. This dataset contains both stroke and non-stroke patients data. We have collected the data from a brain stroke patient after one was diagnosed with the disease and after that admitted. This data was collected from a medical registry book. Our data set contains 385 cases, with males accounting for 209 and females accounting for 176. Our data set also includes data from 234 brain stroke patients and 151 non-brain stroke persons.

3.2.1 Dataset Features

This resulting data set includes 23 characteristics and a predict class. Gender, Age, Residence Area, Work Type, Marital Status, Hypertension, Smoking Habit, Heart Disease, Average Glucose Level, Weight, BMI, RBS, Height, Serum Creatinine, Serum Cholesterol, HDL, Triglyceride, LDL, HbA1c, Hb, WBC, ESR, RBC are the important characteristics.

Table 3.1: Dataset features

Serial	Feature	Feature Details
1	Gender	Male: 0 Female: 1
2	Age	Years
3	Work Type	Government Job: 1 Private Job: 2 Farmer: 3 Businessman: 4 Politics: 5 Teacher: 6 No Job: 7 Self Employed: 8
4	Residence Area	Urban: 0 Rural: 1
5	Marital Status	No: 0 Yes: 1
6	Hypertension	No: 0 Yes: 1
7	Heart Disease	No: 0 Yes: 1
8	Smoking	No: 0 Yes: 1
9	Average Glucose Level	>6.0mmol/l
10	Weight	kg
11	Height	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	Triglyceride	<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/h
24	Class: Stroke	No: 0 Yes: 1

Gender: Strokes are more likely to affect men. However, after a certain age, women are at greater risk. Males make up 209 of the 385 total occurrences we have gathered, while females make up 176.

Age: Age is a major risk factor for stroke. The risk of stroke increases with age, and older people are more prone to experience a stroke and to have more severe symptoms. According to WHO, the incidence of stroke increases exponentially after the age of 55, with the highest rates occurring in people over the age of 75.

Work Type: Some studies have suggested that certain types of work may be associated with an increased risk of stroke. For example, research has shown that people who work in physically demanding or sedentary jobs may be at higher risk of stroke. Jobs with physical demands may increase the risk of stroke by increasing the risk of high blood pressure, obesity, and other conditions that are known risk factors for stroke. Sedentary work, on the

other hand, may increase the risk of stroke by contributing to a lack of some physical activity, which is also associated with an alarming risk of stroke and other health problems. Other factors related to work, such as stress, long working hours, and shift work, may also be associated with an increased risk of stroke. However, more research work is needed to understand exact relationship between work-type and stroke risk.

Residence Area: According to several research, persons who live in cities may be at those who have a higher risk of stroke who reside in rural regions. Living in an urban location may raise the risk of stroke for a variety of reasons. Air pollution in cities is greater, which has been related to an increased risk of stroke and other cardiovascular disorders. In addition, urban locations may have greater levels of stress and social isolation, which might increase the risk of stroke. Furthermore, cities may have greater prevalence of harmful habits such as smoking and poor eating, which can raise the risk of stroke.

Marital Status: Marital status may be related to an individual's risk of stroke, although the relationship which is complex and not fully understood. Some studies have suggested that being married or being in a long-term committed relationship may be associated with a lower risk of stroke compared to being single, divorced, or widowed.

Hypertension: Hypertension, or stroke is a major risk that is associated with high blood pressure. When blood pressure is consistently high, it can cause damage to the blood vessels which leading to strokes. According to the WHO, hypertension is a leading cause of stroke, responsible for up to 50% of stroke cases worldwide. The risk of stroke increases with higher blood pressure levels, and people with uncontrolled hypertension are at particularly high risk of stroke.

Heart Disease: Heart disease, particularly conditions such as coronary artery disease and stroke risk can be elevated by a condition called atrial fibrillation. Coronary artery disease which is a condition in which a heart attack can occur when the arteries that bring blood to the heart become narrowed or blocked, resulting in reducing blood flow, and an increased

risk of heart attack. Atrial fibrillation is a type of irregular heartbeat that raise the likelihood of blood clots forming, which can lead to a stroke. People with heart disease are at an increased risk of both ischemic and hemorrhagic stroke. Ischemic stroke is caused by a blockage in the blood vessels that supply the brain, and can be triggered by a blood clot that forms in the heart and travels to the brain. A Hemorrhagic stroke is caused by bleeding in the brain and can be caused by an aneurysm, which is a bulge in the wall of a blood vessel that can burst and cause bleeding in the brain.

Smoking: Smoking is a major risk factor for stroke. According to the World Health Organization, smoking is a leading cause of stroke, responsible for up to 25% of stroke cases worldwide. Smoking can increase the risk of stroke in several ways. First, smoking damages the blood vessels, making them more prone to blockages and damage. This can lead to an increased risk of ischemic stroke, which is caused by a blockage in the blood vessels that supply the brain. Second, smoking can increase the risk of hemorrhagic stroke, which is caused by bleeding in the brain. Smoking can cause the walls of the blood vessels to become weak and prone to rupture, leading to bleeding in the brain. Third, smoking can increase the risk of other conditions such as high blood pressure, diabetes, heart disease, and other known risk factors for stroke.

Average Glucose Level: High blood sugar levels, also known as hyperglycemia, can increase the risk of stroke. Diabetes is a condition that is characterized by high blood sugar levels, and people with diabetes are at an increased risk of stroke.

Weight: Being overweight or obese can increase the risk of stroke. Obesity is a condition characterized by excess body fat, and having high blood pressure is linked to a higher likelihood of several health issues, including stroke. Obesity can increase the risk of stroke in several ways. First, obesity can increase the risk of conditions that are known risk factors for stroke, such as high blood pressure, diabetes, and heart disease. Second, obesity can cause inflammation in the body, which can damage the blood vessels and increase the risk of stroke.

Height: There is some evidence to suggest that taller people may be at an increased risk of stroke compared to shorter people. A number of studies have found that taller people have a higher risk of ischemic stroke, Stroke occurs when the blood flow to the brain is interrupted due to a blockage in the blood vessels. However, relationship between height and stroke risk is complex and not fully understood.

BMI: Body mass index (BMI) is a calculation that uses a person's weight and height to estimate their body fat percentage. High BMI is associated with an increased risk of stroke and other health problems. Obesity, which is defined as a BMI of 30 or higher, significantly increases the likelihood of stroke occurring. Obesity can increase the risk of stroke in several ways, including by increasing the risk of conditions such as diabetes, high blood pressure, and heart disease. Obesity can also cause inflammation in the body, which can damage the blood vessels and increase the risk of stroke.

RBS: RBS stands for random blood sugar, which is a measure of blood sugar levels at a specific point in time. High blood sugar levels, also known as hyperglycemia, can increase the risk of stroke. Diabetes is a condition that is characterized by high blood sugar levels, and people with diabetes are at an increased risk of stroke. In people with diabetes, high blood sugar levels can cause damage to the blood vessels and nerves, leading to an increased risk of stroke. High blood sugar levels can also increase the risk of other conditions that are known risk factors for stroke, such as high blood pressure and heart disease.

Serum Creatinine: Serum creatinine is a laboratory test that measures the level of creatinine in the blood. Creatinine is a waste product that is produced by the muscles and filtered out of the blood by the kidneys. Elevated serum creatinine levels may be a sign of kidney disease, which can increase the risk of stroke. People with kidney disease may be at higher risk of stroke due to a number of factors, including high blood pressure, inflammation, and an increased risk of blood clots.

Serum Cholesterol: Elevated serum cholesterol levels can increase the risk of stroke. Cholesterol is a type of fat that is found in the blood, and high levels of cholesterol can contribute to the build-up of plaque in the blood vessels, which can lead to a stroke.

LDL: Low-density lipoprotein (LDL) cholesterol, also known as "bad" cholesterol, can increase the risk of stroke. LDL cholesterol can contribute to the build-up of plaque in blood vessels, which can lead to a stroke. High LDL cholesterol levels are a major risk factor for stroke, along with other conditions such as heart disease and high blood pressure. Reducing LDL cholesterol levels can help to reduce the risk of stroke and other health problems.

HDL: High-density lipoprotein (HDL) cholesterol, also known as "good" cholesterol, may be associated with a lower risk of stroke. HDL cholesterol helps to remove plaque from the blood vessels and can help to protect against the build-up of plaque that can lead to a stroke. While low levels of HDL cholesterol are a risk factor for stroke and other health problems, high levels of HDL cholesterol may be protective. However, the relationship between HDL cholesterol and stroke risk is complex and not fully understood.

Triglyceride: High levels of triglycerides, a type of fat found in the blood, can increase the risk of stroke. Triglycerides are a risk factor for stroke, along with other conditions such as high blood pressure and heart disease. High triglyceride levels can contribute to the build-up of plaque in the blood vessels, which can lead to a stroke. In addition, high triglyceride levels may be associated with other conditions that increase the risk of stroke, such as diabetes and obesity.

HbA1c: HbA1c, also known as glycated hemoglobin, is a laboratory test that measures the average blood sugar levels over the past two to three months. High HbA1c levels may be a sign of uncontrolled diabetes, which can increase the risk of stroke.

Hb: Hb, also known as hemoglobin, is a protein found in red blood cells that carries oxygen from the lungs to the rest of the body. Low levels of Hb, also known as anemia, may be associated with an increased risk of stroke. Anemia can increase the risk of stroke in several ways. First, anemia can cause the heart to work harder to pump blood, leading to an increased risk of heart disease and stroke. Second, anemia can cause a decrease in oxygen delivery to the brain, which may increase the risk of stroke.

RBC: Red blood cells (RBCs) are cells in the blood that carry oxygen to the body's tissues. Low levels of RBCs, also known as anemia, may be associated with an increased risk of stroke.

WBC: White blood cells (WBCs) are cells in the blood that help to fight infections and protect the body from illness. High levels of WBCs may be a sign of inflammation in the body, which can increase the risk of stroke. Inflammation can damage the blood vessels and increase the risk of stroke. In addition, high levels of WBCs may be associated with other conditions that increase the risk of stroke, such as high blood pressure and heart disease.

ESR: Erythrocyte sedimentation rate (ESR) is a laboratory test that measures the rate at which red blood cells (RBCs) settle to the bottom of a tube in a laboratory setting. Elevated ESR levels may be a sign of inflammation in the body, which can increase the risk of stroke.

3.3 Machine Learning Algorithms

Machine learning is a type of artificial intelligence that allows computers to learn and improve their performance on a specific task without being explicitly programmed. Machine learning algorithms are able to learn from data, identify patterns, and make predictions or decisions based on that data. There are different types of machine learning algorithms, including supervised learning algorithms, which are trained on labeled data and can be used to make predictions about new, unseen data, and unsupervised learning algorithms, which are used to discover patterns in data. Machine learning is used in a wide

range of applications, including image and speech recognition, natural language processing, and predictive modeling. It is a rapidly growing field that is transforming many industries and has the potential to revolutionize the way we interact with and make sense of data. We have used 10 machine learning classifiers for this work.

3.3.1 Logistic Regression

Logistic regression is a statistical method used for predicting binary outcomes, such as whether an individual will have a certain disease or not. It is a type of supervised learning algorithm that is used to model the relationship between a dependent variable and one or more independent variables by fitting a logistic curve to the data.

In logistic regression, the relationship between the dependent variable (Y) and the independent variables (X) is modeled using the following equation:

$$\log(p/(1-p)) = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

In this equation, p is the probability of the dependent variable occurring (e.g., the probability of having a certain disease), b₀ is the intercept, and b₁, b₂, ..., b_n are the coefficients for each independent variable. The values of the independent variables are X₁, X₂,..., X_n.

The coefficients in the equation are estimated using maximum likelihood estimation, which involves finding the values of the coefficients that maximize the likelihood of the data given the model. The resulting equation can then be used to predict the probability of the dependent variable occurring for a given set of values for the independent variables.

3.3.2 Gaussian Naïve Bayes

Gaussian naive Bayes is a type of supervised machine learning algorithm that is used for classification tasks. It is based on the idea of Bayes' theorem, which states that the probability of an event occurring is related to the prior probability of the event occurring and the likelihood of the event given certain observations.

In Gaussian naive Bayes, the probability of a class (C) given a set of features (X) is calculated using Bayes' theorem:

$$P(C|X) = P(X|C) * P(C) / P(X)$$

In this equation, $P(C|X)$ is the probability of the class (C) occurring given the features (X), $P(X|C)$ is the probability of the features (X) occurring given the class (C), $P(C)$ is the prior probability of the class (C) occurring, and $P(X)$ is the prior probability of the features (X) occurring.

The probability of the features (X) occurring given the class (C) is estimated using the normal (Gaussian) distribution, which is defined by the mean (μ) and standard deviation (σ) of the features in the class. The probability density function of the normal distribution is given by the following equation:

$$f(x; \mu, \sigma) = (1 / (\sigma * \sqrt{2\pi})) * \exp(-((x - \mu)^2) / (2 * \sigma^2))$$

In this equation, x is a feature value, μ is the mean of the feature values in the class, and σ is the standard deviation of the feature values in the class.

The probability of the class (C) occurring given the features (X) is then calculated using the probabilities of the individual features (X) occurring given the class (C), which are estimated using the normal distribution. The class with the highest probability is chosen as the predicted class.

3.3.3 K Neighbors

K-nearest neighbors (KNN) is a type of supervised machine learning algorithm that is used for classification and regression tasks. It is based on the idea of finding the K number of data points in the training set that are closest to a given data point and using those data points to make a prediction.

The K-nearest neighbors (KNN) algorithm does not have a specific formula as it does not involve training a model or estimating coefficients. Instead, it relies on the distance between data points to make predictions.

To make a prediction using KNN, the distance between the data point to be predicted (X) and each data point in the training set (X1, X2, ..., Xn) is calculated using a distance measure, such as Euclidean distance. The K data points in the training set that are closest to the data point to be predicted are then identified, and the prediction is made based on these K nearest neighbors.

3.3.4 Random Forest

Random forest is a type of ensemble learning algorithm that is used for classification and regression tasks. It is based on the idea of building a collection (ensemble) of decision trees, each of which is trained on a random subset of the training data. The predictions made by the individual decision trees are then combined to make a final prediction.

To make a prediction using a random forest, the algorithm first makes a prediction for each decision tree in the ensemble. The final prediction is then made by combining the predictions of the individual trees, either by taking the majority vote for classification tasks or by taking the average for regression tasks.

3.3.5 Decision Tree

A decision tree is a type of machine learning algorithm that is used for classification and regression tasks. It is based on the idea of building a tree-like model of decisions, where an internal node represents a feature or attribute, the branches represent decisions based on that attribute, and the leaves represent the final prediction or classification.

The Gini index is a measure of impurity that is used to split the data at each node in a decision tree. It is calculated using the following formula:

$$\text{Gini} = 1 - \sum(p(i|t)^2)$$

In this equation, $p(i|t)$ is the probability of class i occurring at a given node t , and the sum is taken over all classes. The Gini index ranges from 0 (pure) to 1 (impure), and the goal is to split the data in a way that maximizes the reduction in impurity.

The entropy is another measure of impurity that is used to split the data at each node in a decision tree. It is calculated using the following formula:

$$\text{Entropy} = -\sum(p(i|t) * \log(p(i|t)))$$

In this equation, $p(i|t)$ is the probability of class i occurring at a given node t , and the sum is taken over all classes. The entropy ranges from 0 (pure) to $\log(n)$ (impure), where n is the number of classes, and the goal is to split the data in a way that maximizes the reduction in impurity.

3.3.6 XG Boost

XG Boost (eXtreme Gradient Boosting) is a type of gradient boosting algorithm that is used for classification and regression tasks. It is a powerful and widely used machine learning algorithm that is known for its efficiency, flexibility, and predictive performance.

In XG Boost, the objective function to be minimized is defined as follows:

$$\text{Loss}(y, f) = \sum(L(y, f)) + \Omega(f)$$

In this equation, $L(y, f)$ is the loss function, y is the true label, f is the predicted label, and $\Omega(f)$ is the regularization term. The loss function measures the discrepancy between the true label and the predicted label, and the regularization term helps to prevent overfitting by adding a penalty to the objective function for large values of f .

The gradient of the loss function is then calculated using the following formula:

$$\partial \text{Loss}(y, f) / \partial f = \partial L(y, f) / \partial f + \partial \Omega(f) / \partial f$$

The gradient is used to update the weights of the weak learners at each iteration, and the process is repeated until the loss function is minimized or a pre-defined number of iterations is reached.

3.3.7 SVM

Support vector machines (SVMs) are a type of supervised machine learning algorithm that is used for classification and regression tasks. It is based on the idea of finding the hyperplane in an N-dimensional space that maximally separates the data points of different classes.

In SVM, the data points are represented as vectors in an N-dimensional space, and the goal is to find the hyperplane that maximally separates the data points of different classes. The hyperplane is defined by a weight vector (w) and a bias term (b), and the decision boundary is given by the equation $wx + b = 0$. The data points that are closest to the hyperplane are called support vectors and play a crucial role in determining the position of the hyperplane.

In SVM, the optimization problem is defined as follows:

$$\begin{aligned} &\text{minimize } (1/2) * \|w\|^2 \\ &\text{subject to } y(i) * (w * x(i) + b) \geq 1 \text{ for } i = 1, 2, \dots, n \end{aligned}$$

In this equation, $\|w\|$ is the Euclidean norm of the weight vector w , $y(i)$ is the true label of data point i , $x(i)$ is the feature vector of data point i , and n is the number of data points. The goal is to find the weight vector and bias term that minimize the Euclidean norm of w while ensuring that the margin between the classes is maximized.

3.3.8 Ada Boost

Ada Boost (Adaptive Boosting) is a type of ensemble learning algorithm that is used for classification and regression tasks. It works by building a collection (ensemble) of weak learners (e.g., decision trees) and combining them to create a strong learner. The algorithm adjusts the weights of the weak learners at each iteration based on their performance, with the goal of increasing the accuracy of the ensemble.

In Ada Boost, the weak learners are trained sequentially, and at each iteration, the weights of the misclassified data points are increased to give them more importance. The weak learners are then trained on the weighted data, and the process is repeated until the desired number of iterations is reached or the error rate reaches a pre-defined threshold.

The final prediction of Ada Boost is given by the following formula:

$$f(x) = \sum(\alpha(i) * h(i)(x))$$

In this equation, $\alpha(i)$ is the weight of the i -th weak learner, $h(i)(x)$ is the prediction of the i -th weak learner, and the sum is taken over all weak learners. The final prediction is made by combining the predictions of the individual weak learners using the weights assigned by Ada Boost.

3.3.9 Bagging

Bagging (Bootstrap Aggregating) is a type of ensemble learning algorithm that is used for classification and regression tasks. It works by building a collection (ensemble) of weak learners (e.g., decision trees) and combining them to create a strong learner. The goal of bagging is to reduce the variance of the ensemble, which helps to improve the generalization error and reduce overfitting.

In bagging, the weak learners are trained on different subsets of the training data, which are created using bootstrapping. Bootstrapping is a sampling technique that involves

sampling with replacement from the original dataset to create multiple new datasets (bootstrapped samples). The weak learners are then trained on the bootstrapped samples and combined to create the final prediction.

The final prediction of bagging is given by the following formula:

$$f(x) = 1/M * \sum(h(i)(x))$$

In this equation, M is the number of weak learners, $h(i)(x)$ is the prediction of the i-th weak learner, and the sum is taken over all weak learners. The final prediction is made by averaging the predictions of the individual weak learners.

3.3.10 Gradient Boosting

Gradient Boosting is a type of ensemble learning algorithm that is used for classification and regression tasks. It works by building a collection (ensemble) of weak learners (e.g., decision trees) and combining them to create a strong learner. The algorithm adjusts the weights of the weak learners at each iteration based on the gradient of the loss function, which helps to reduce the error and improve the prediction accuracy.

In Gradient Boosting, the weak learners are trained sequentially, and at each iteration, the prediction of the ensemble is updated based on the gradient of the loss function. The weak learners are then trained on the residual errors (difference between the true label and the predicted label) and combined to create the final prediction.

The final prediction of Gradient Boosting is given by the following formula:

$$f(x) = f(x) + h(x)$$

In this equation, $f(x)$ is the prediction of the ensemble, and $h(x)$ is the prediction of the weak learner. The final prediction is made by adding the prediction of the weak learner to the prediction of the ensemble.

3.4 Proposed Model

The proposed model (Figure 3.4) of this paper, describe below:

- i. Data collect from different hospitals in Bangladesh
- ii. Then we need to pre-processing the data.
- iii. Then analyze the preprocessed data using some machine learning methods to get the prediction value.
- iv. Check also the confusion matrix results.

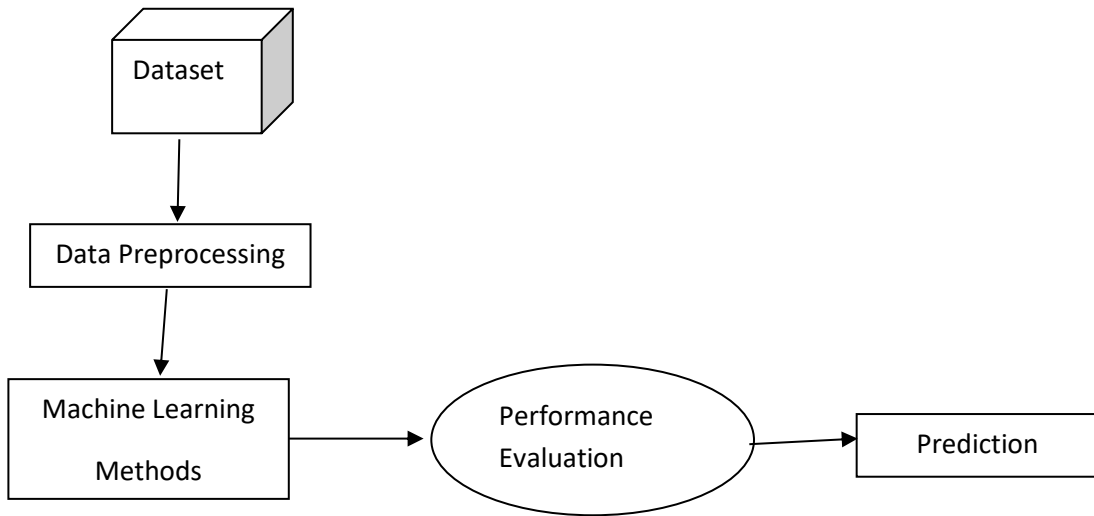


Figure: 3.1 Proposed model method

3.5 Working Procedure

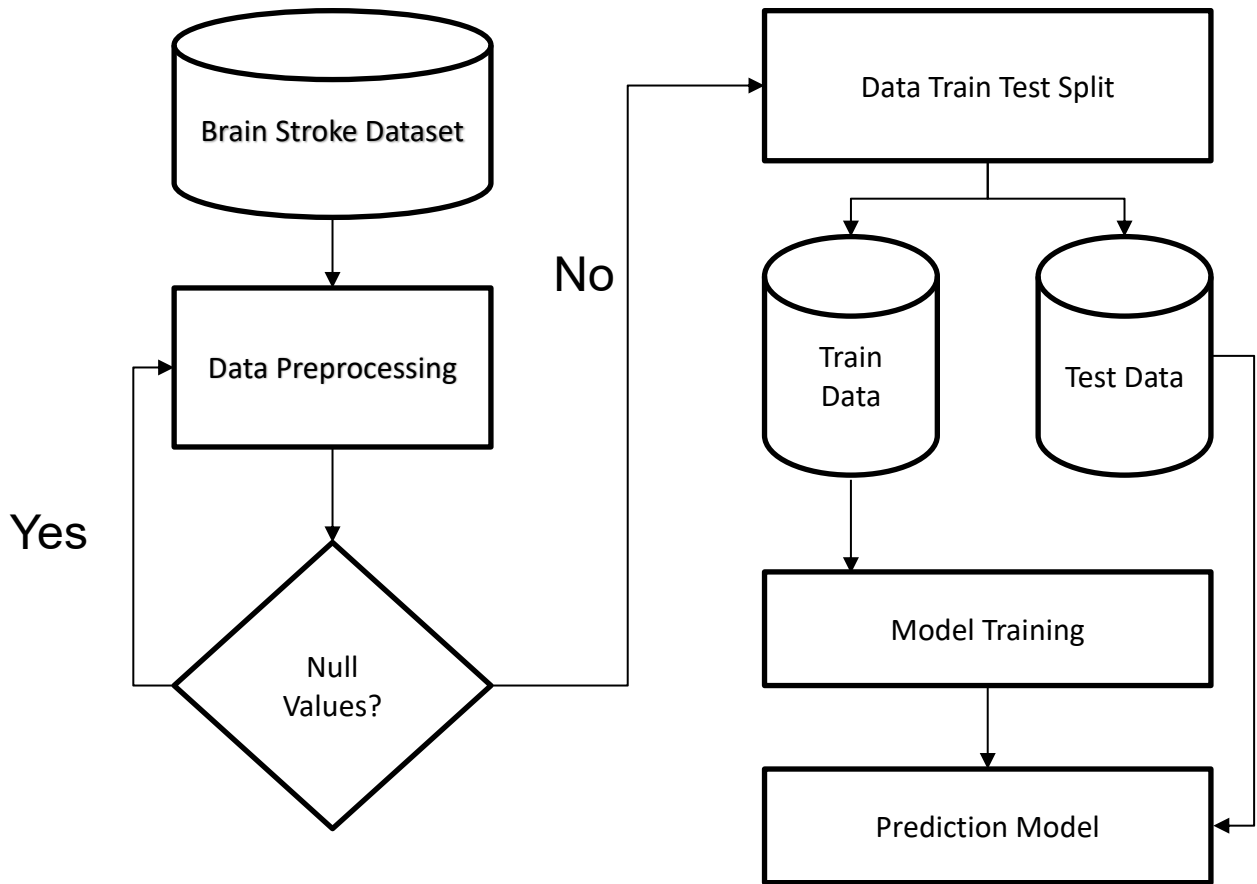


Figure: 3.2 Working Procedure

CHAPTER 4

RESULTS ANALYSIS

4.1 Experimental Results

To assess performance, we used Accuracy, Precision, Recall, and F1-score. The proportion of correct predictions to total predictions made by the model is known as classification accuracy. The ratio of true positive predictions to true positive and false positive predictions is defined as precision. The recall is the ratio of true positives to true positives and false negatives. The F1-score, often known as the F-measure, is a balancing measure that is intended to represent performance in a single figure. It is the mean of precision and recall.

A confusion matrix is a table that is used to evaluate the performance of a machine learning model, particularly for classification tasks. It shows the number of true positive, true negative, false positive, and false negative predictions made by the model.

Table 4.1: Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

Accuracy is a measure of how well a machine learning model is able to predict the true values of the data. It is calculated as the proportion of correct predictions made by the model. For example, if a model makes 90 correct predictions out of 100 total predictions, the accuracy would be 90%.

Accuracy is calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{\text{Total Prediction}}$$

Precision is a measure of how accurate a machine learning model is in making positive predictions. It is calculated as the proportion of correct positive predictions made by the model out of all the positive predictions made by the model.

Precision is calculated as follows:

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Recall is a measure of how sensitive a machine learning model is to positive cases. It is calculated as the proportion of correct positive predictions made by the model out of all the actual positive cases.

Recall is calculated as follows:

$$\text{Recall} = \frac{TP}{(TP+FN)}$$

The F1 score is a measure of the balance between precision and recall in a machine learning model. It is calculated as the harmonic mean of precision and recall, with a higher score indicating a better balance between the two.

$$\text{F1 Score} = 2 * \frac{\text{Precision*Recall}}{(\text{Precision+Recall})}$$

We employed ten distinct machine learning techniques, as well as an ensemble model. Logistic Regression Classifier, Gaussian Naive Bayes Classifier, K-Neighbors Classifier, Random Forest Classifier, Decision Tree Classifier, XG Boost Classifier, SVC Classifier, Ada Boost Classifier, Bagging Classifier, Gradient Boosting Classifier, and an ensemble model are among the approaches. Python was picked for this categorization. 70% of the data was used for training, whereas 30% was used for testing. Random Forest Classifier

has the highest accuracy of these ten classifiers, at 96.87%. Gradient Boosting Classifier accuracy is thus 96.09%. The confusion matrix of Decision Tree, Random Forest, K Neighbors, Ada Boost, Gaussian Naïve Bayes, SVC, Logistic Regression, XG Boost, Bagging, and Gradient Boosting are showed in Table 4.2 to Table 4.11 respectively. Performance analysis of different machine learning classifiers are showed in Table 4.12. Comparison bar graph of precision, recall, f1 score, and accuracy are showed in Figure 4.1 to Figure 4.4 respectively. Comparison of all classification is showed in Figure 4.5.

Table 4.2: Confusion Matrix for Decision Tree

	Actual Positive	Actual Negative
Predicted Positive	45	3
Predicted Negative	6	46

Table 4.3: Confusion Matrix for Random Forest

	Actual Positive	Actual Negative
Predicted Positive	47	1
Predicted Negative	2	50

Table 4.4: Confusion Matrix for K Neighbors

	Actual Positive	Actual Negative
Predicted Positive	44	4
Predicted Negative	34	18

Table 4.5: Confusion Matrix for Ada Boost

	Actual Positive	Actual Negative
Predicted Positive	48	0
Predicted Negative	5	47

Table 4.6: Confusion Matrix for Gaussian Naïve Bayes

	Actual Positive	Actual Negative
Predicted Positive	47	1
Predicted Negative	3	49

Table 4.7: Confusion Matrix for SVC

	Actual Positive	Actual Negative
Predicted Positive	48	0
Predicted Negative	52	0

Table 4.8: Confusion Matrix for Logistic Regression

	Actual Positive	Actual Negative
Predicted Positive	46	2
Predicted Negative	9	43

Table 4.9: Confusion Matrix for XG Boost

	Actual Positive	Actual Negative
Predicted Positive	47	1
Predicted Negative	1	51

Table 4.10: Confusion Matrix for Bagging

	Actual Positive	Actual Negative
Predicted Positive	47	1
Predicted Negative	4	48

Table 4.11: Confusion Matrix for Gradient Boosting

	Actual Positive	Actual Negative
Predicted Positive	47	1
Predicted Negative	0	52

Table 4.12: Performance Analysis of Different Machine Learning Classifiers

Classifiers	Precision	Recall	F1 Score	Accuracy
Logistic Regression	91.25%	87.95%	89.57%	86.72%
Gaussian Naïve Bayes	96.39%	96.39%	96.39%	95.31%
K Neighbors	81.25%	62.65%	70.75%	66.41%
Random Forest	96.39%	96.39%	96.39%	95.31%
Decision Tree	90.12%	87.95%	89.02%	85.94%
XG Boost	95.29%	97.59%	96.43%	95.31%
SVC	64.84%	1.00%	78.67%	64.84%
Ada Boost	95.24%	96.39%	95.81%	94.53%
Bagging	97.47%	92.77%	95.06%	93.75%
Gradient Boosting	96.43%	97.59%	97.01%	96.09%

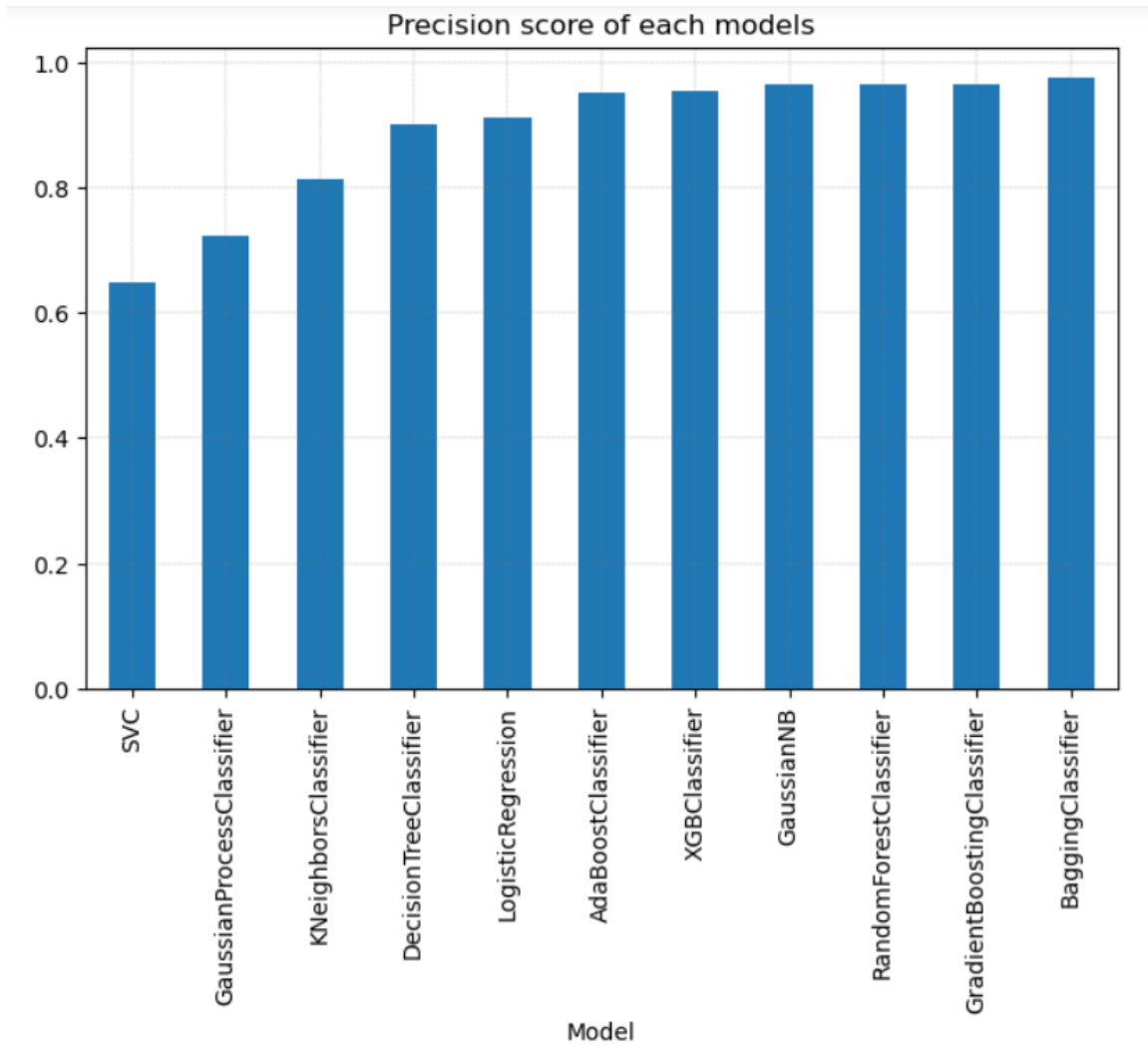


Figure 4.1: Comparison bar graph of Precision

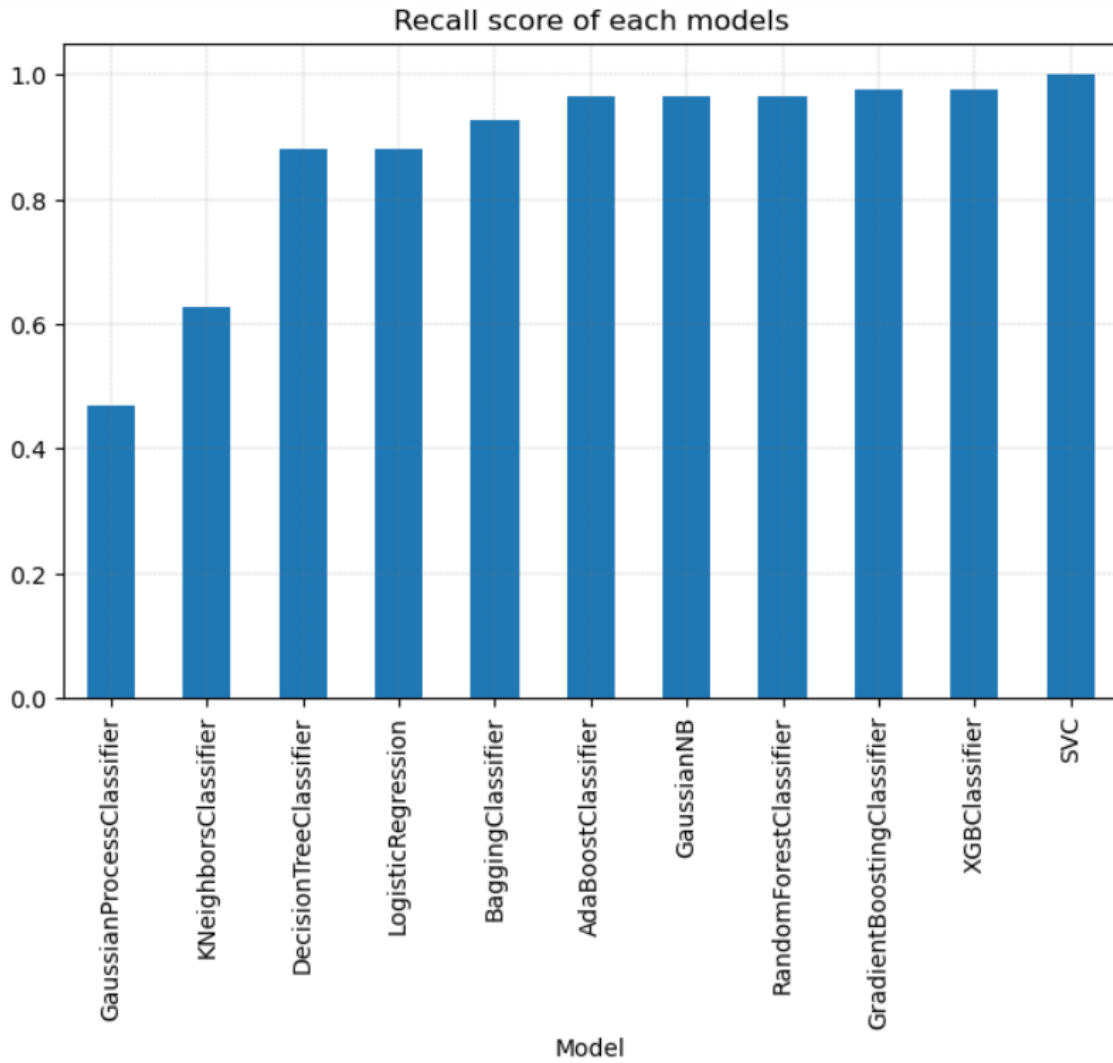


Figure 4.2: Comparison bar graph of Recall

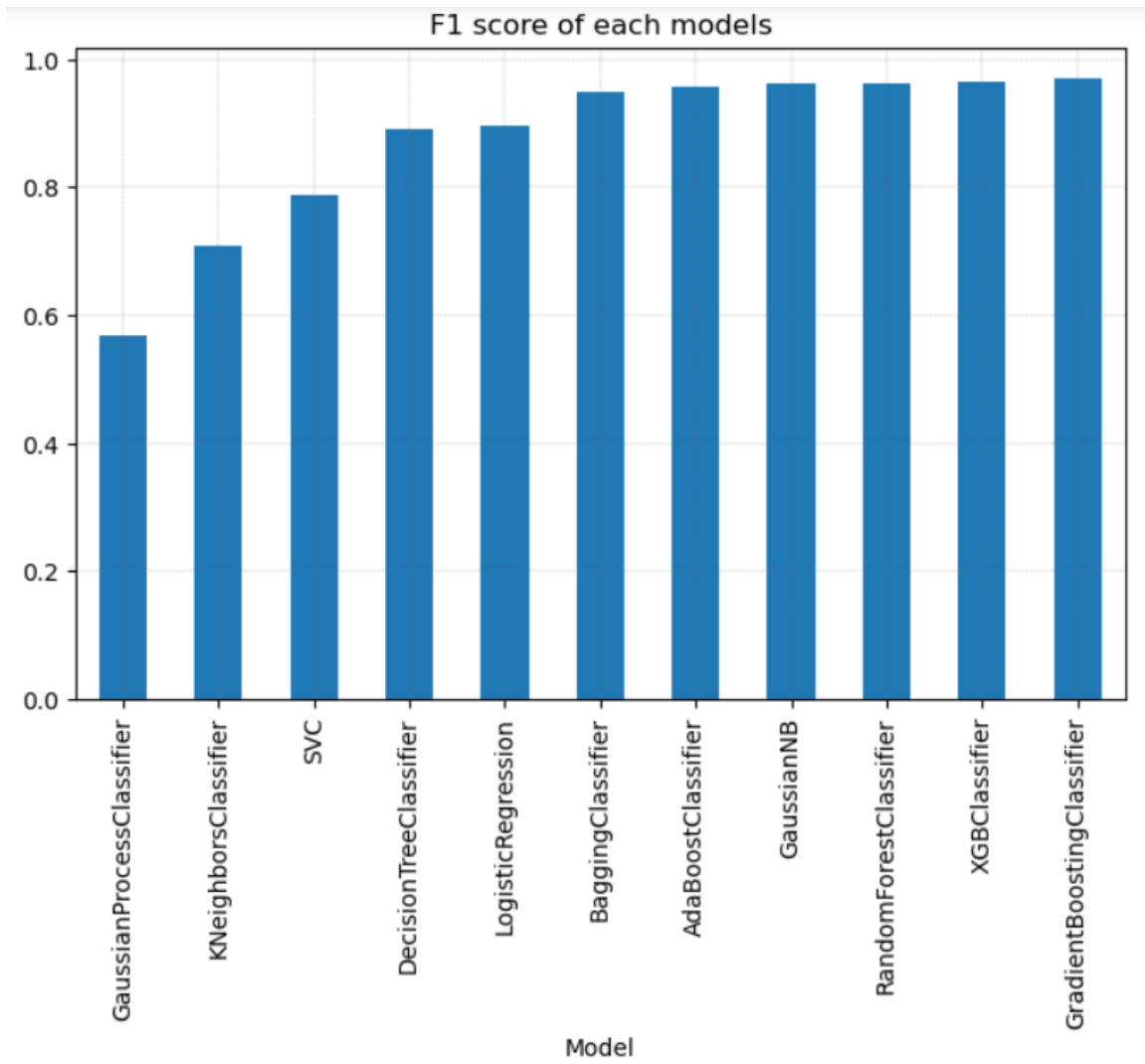


Figure 4.3: Comparison bar graph of F1 Score

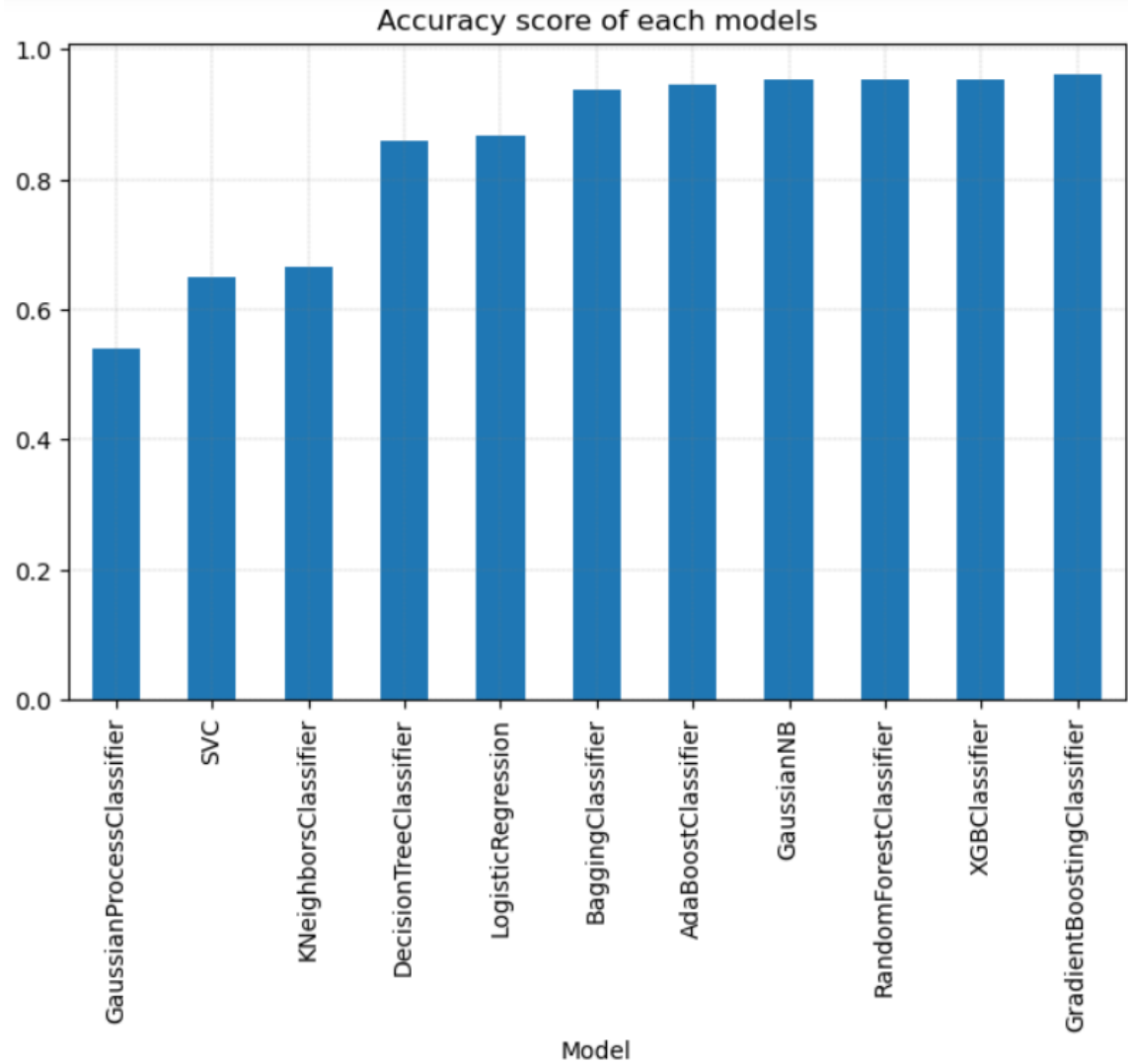


Figure 4.4: Comparison bar graph of Accuracy

	Model	Score_Test	Time_Train	Test_Score_Accuracy	Test_Score_Recall	Test_Score_Precision	F1_Score
0	LogisticRegression	0.863846	0.617535	0.867188	0.879518	0.912500	0.895706
1	GaussianNB	0.933692	0.105165	0.953125	0.963855	0.963855	0.963855
2	KNeighborsClassifier	0.653231	0.520108	0.664062	0.626506	0.812500	0.707483
3	RandomForestClassifier	0.922000	20.516938	0.953125	0.963855	0.963855	0.963855
4	DecisionTreeClassifier	0.836308	0.133844	0.859375	0.879518	0.901235	0.890244
5	XGBClassifier	0.933538	1.101872	0.953125	0.975904	0.952941	0.964286
6	SVC	0.587692	0.199655	0.648438	1.000000	0.648438	0.786730
7	AdaBoostClassifier	0.918462	2.199080	0.945312	0.963855	0.952381	0.958084
8	BaggingClassifier	0.898769	0.714102	0.937500	0.927711	0.974684	0.950617
9	GradientBoostingClassifier	0.941538	8.620927	0.960938	0.975904	0.964286	0.970060
10	GaussianProcessClassifier	0.610923	3.669499	0.539062	0.469880	0.722222	0.569343

Figure 4.5: Comparison of Different Classifiers.

CHAPTER 5

CONCLUSION

5.1 Summary

In this study, a brain stroke prediction model was developed using data from five hospitals in Bangladesh to aid doctors in predicting brain stroke in patients based on their clinical data. The model was trained on a dataset of 385 instances with 23 features and ten different machine learning algorithms were evaluated. The results showed that the Gradient Boosting algorithm had the highest accuracy, above 96%.

5.2 Discussion

In our dataset, we have 385 patient's data along with 23 features. We have applied 10 machine learning algorithm to get the best result. Our main limitation is data shortage. If we get more data, we can get even better result from our prediction model. The classification accuracy of Logistic Regression (86.72%), Gaussian Naïve Bayes (95.31%), K Neighbors (66.41%), Random Forest (95.31%), Decision Tree (85.94%), XG Boost (95.31%), SVC (64.84%), Ada Boost (94.53%), Bagging (93.75%), and Gradient Boosting (96.09%). In our work, Gradient Boosting gave best result among these classifications.

5.3 Future Work

In this work, we had shortage of data. We want to collect more data for this dataset and also want to work on this model. And we want to add GUI for user experience.

5.4 Conclusion

The purpose of this study was to develop a brain stroke prediction model that could identify the condition in its early stages and create a dataset of brain stroke cases in Bangladesh. To achieve this goal, we collected data from several hospitals in the country. Ten different machine learning classification techniques, including Logistic Regression, K Neighbors, SVC, Gaussian Naïve Bayes, Bagging, Ada Boost, XG Boost, Random Forest, Gradient Boosting, and Decision Tree, were applied to the data. The results showed that the Gradient Boosting technique was the most effective, with an accuracy of above 96%. This work

represents an important contribution to the field of brain stroke prediction and to the development of the first brain stroke data repository in Bangladesh.

REFERENCES

- [1] Roohi, Majid, et al. "Machine learning approaches for automated stroke detection, segmentation, and classification in microwave brain imaging systems." *Progress In Electromagnetics Research C* 116 (2021): 193-205.
- [2] Bandi, Vamsi, Debnath Bhattacharyya, and Divya Midhunchakkravarthy. "Prediction of Brain Stroke Severity Using Machine Learning." *Rev. d'Intelligence Artif.* 34.6 (2020): 753-761.
- [3] Khosla, Aditya, et al. "An integrated machine learning approach to stroke prediction." *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010.
- [4] Chin, Chiun-Li, et al. "An automated early ischemic stroke detection system using CNN deep learning algorithm." *2017 IEEE 8th International conference on awareness science and technology (ICAST)*. IEEE, 2017.
- [5] Emon, Minhaz Uddin, et al. "Performance analysis of machine learning approaches in stroke prediction." *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2020.
- [6] Shoily, Tasfia Ismail, et al. "Detection of stroke disease using machine learning algorithms." *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2019.
- [7] Badriyah, Tessy, et al. "Machine learning algorithm for stroke disease classification." *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*. IEEE, 2020.
- [8] Kumar, G. Ravi, et al. "Brain Stroke Detection Using Machine Learning." *International Journal of Research in Engineering, Science and Management* 5.3 (2022): 34-36.
- [9] Heo, Tak Sung, et al. "Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI." *Journal of personalized medicine* 10.4 (2020): 286.
- [10] Kamal, Haris, Victor Lopez, and Sunil A. Sheth. "Machine learning in acute ischemic stroke neuroimaging." *Frontiers in neurology* 9 (2018): 945.
- [11] Lee, Hyunna, et al. "Machine learning approach to identify stroke within 4.5 hours." *Stroke* 51.3 (2020): 860-866.

[12] Lee, Eun-Jae, et al. "Deep into the brain: artificial intelligence in stroke imaging." *Journal of stroke* 19.3 (2017): 277.

[13] Kim, Chulho, et al. "Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke." *PloS one* 14.2 (2019): e0212778.

Turnitin Originality Report

Document Viewer

Processed on: 2023#01#04# 09:49 +06
ID: 1988400118
Word Count: 8448
Submitted: 1

Similarity Index 25%	Similarity by Source	
	Internet Sources:	20%
	Publications:	11%
	Student Papers:	14%

A COMPARATIVE ANALYSIS ON PREDICTION OF BRAIN... By Mst. Eshita Khatun

[exclude quoted](#) [exclude bibliography](#) [exclude small matches](#) mode: [quickview \(classic\) report](#) [print](#) [refresh](#) [download](#)

1% match (Internet from 20-Nov-2022) http://dspace.daffodilvarsity.edu.bd:8080	❏
1% match (Internet from 18-Oct-2021) http://repository.aust.edu.ng	❏
1% match (student papers from 15-Dec-2022) Submitted to Coventry University on 2022-12-15	❏
1% match (student papers from 28-Nov-2022) Submitted to Liverpool John Moores University on 2022-11-28	❏
1% match (Internet from 06-Mar-2022) https://vsip.info/bma-illustrated-medical-dictionary-3rd-edition-pdf-free.html	❏
1% match (student papers from 15-Dec-2022) Submitted to Bahcesehir University on 2022-12-15	❏
1% match (student papers from 04-Dec-2022) Submitted to Taylor's Education Group on 2022-12-04	❏
1% match (student papers from 17-Dec-2022) Submitted to University of East London on 2022-12-17	❏
1% match ("Transient Ischemic Attacks", Wiley, 2004) "Transient Ischemic Attacks", Wiley, 2004	❏
1% match (Internet from 25-Oct-2021) https://www.cambridge.org/core/journals/animal-health-research-reviews/article/review-of-traditional-and-machine-learning-methods-applied-to-animal-breeding/03DAAED77E8525B07B0E66A622B7CF0D	❏
1% match (Internet from 20-Mar-2022) https://www.globe.gov/documents/10157/10754848/Final_IVSS_Submission_Argus_Jain_Mittelman_Rawashdeh.pdf/f55bad22-3a73-4c87-07e0-86c78570488a?download=true&t=1647013879247&version=1.0	❏
<1% match (student papers from 24-Aug-2022) Submitted to Liverpool John Moores University on 2022-08-24	❏
<1% match (student papers from 16-Dec-2022) Submitted to University of Greenwich on 2022-12-16	❏
<1% match (student papers from 16-Dec-2022) Submitted to University of Greenwich on 2022-12-16	❏
<1% match (Internet from 26-Jul-2019) https://www.coursehero.com/file/36373547/L14pdf/	❏
<1% match (Internet from 01-Oct-2022) https://www.coursehero.com/file/49558362/MC2-UNIT3docx/	❏
<1% match (Internet from 23-Dec-2022) https://www.coursehero.com/file/79227984/Basic-diagnostic-grid-unit-1docx/	❏
<1% match (student papers from 12-Dec-2022) Submitted to The Robert Gordon University on 2022-12-12	❏
<1% match (student papers from 22-Dec-2022) Submitted to University of Lancaster on 2022-12-22	❏
<1% match (student papers from 28-Dec-2022)	❏