
**HAWASSA UNIVERSITY**

**INSTITUTE OF TECHNOLOGY**

**FACULTY OF INFORMATICS**

**DEPARTMENT OF COMPUTER SCIENCE**

# ENSEMBLE LEARNING-BASED PREDICTION OF STROKE RISK

**M.Sc. Thesis**

**SELAMAWIT TADESSE CHACHAMO**

**HAWASSA UNIVERSITY, HAWASSA, ETHIOPIA**

**NOVEMBER, 2024**

# ENSEMBLE LEARNING-BASED PREDICTION OF STROKE RISK

SELAMAWIT TADESSE CHACHAMO

A THESIS SUBMITTED TO THE
DEPARTMENT OF COMPUTER SCIENCE,
FACULTY OF INFORMATICS, SCHOOL
OF GRADUATE STUDIES
HAWASSA UNIVERSITY
HAWASSA, ETHIOPIA

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE IN COMPUTER SCIENCE

NOVEMBER, 2024

## Declaration

I hereby declare that this thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been acknowledged appropriately.

Name: Selamawit Tadesse Chachamo

Signature: _____

Date: _____

## SCHOOL OF GRADUATE STUDIES

## HAWASSA UNIVERSITY ADVISORS' APPROVAL SHEET

This is to certify that the thesis entitled "Ensemble Learning –based Prediction of Stroke Risk" submitted in partial fulfillment of the requirements for the degree of Master of Science(MSc) with specialization in Computer Science, the Graduate Program of the Department of Computer Science in the faculty of informatics, and has been carried out by Selamawit Tadesse Chachamo under our supervision. Therefore, we recommend that the student has fulfilled the requirements and hence hereby can submit the thesis to the department.

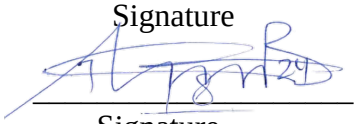
Dr. Varagantham Anitha _____ 10-10-2024

Name of major advisor Signature Date

## SCHOOL OF GRADUATE STUDIES

## HAWASSA UNIVERSITY EXAMINERS' APPROVAL SHEET

We, the undersigned, members of the Board of Examiners of the final open defense by **Selamawit Tadesse** have read and evaluated her thesis entitled "Ensemble Learning–based Prediction of Stroke Risk", and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirements for the degree.

| | | |
|---|---|---|
| Dr. Varagantham Anitha |  | 10-10-2024 |
| Name of Major Advisor | Signature | Date |
| | | |
| Name of Internal Examiner-I | Signature | Date |
| | | |
| Name of Internal Examiner-II | Signature | Date |
| Asrat Mulatu (Ph.D.) |  | 28-11-2024 |
| Name of External Examiner | Signature | Date |
| | | |
| SGS Approval | Signature | Date |

Final approval and acceptance of the thesis is contingent upon the submission of the final copy of the thesis to the School of Graduate Studies (SGS) through the Department/School Graduate Committee (DGC/SGC) of the candidate's department.

Stamp of SGS Date: _____

## ACKNOWLEDGMENTS

First and foremost, I want to sincerely thank Almighty God for always being there for me and for keeping me well so that I could finish my studies. Next, I would like to sincerely thank my advisor, Dr. Varagantham Anitha (PhD), for her great and helpful suggestions, remarks, and remarkable follow-up in every aspect of this research. I will always be grateful for her support and sincerity.

I appreciate my parents' unwavering love and support, which keep me motivated and confident in my research endeavors. They had faith in me, and that's why I succeeded. My classmates for providing suggestions for possible study topics and for being there for me during all of my experiences, as well as my instructors at Hawassa University's Department of Computer Science, are the recipients of my next gratitude. In addition, I want to express my gratitude to my friends for helping me with many facets of my study.

Finally, I would like to thank Jinka University which supports me financially and the workers at Yirgalem General Hospital, Hawassa Referral Hospital, and Yanet Internal Medicine Specialty Center for helping to collect the data that is useful for my study.

# Table of Contents

# List of Tables

# List of Figures

# List of Acronyms and Abbreviations

| | |
|---|---|
| BMI | Body Mass Index |
| CDSS | Clinical Decision Support Systems |
| CPU | Central Processing Unit |
| CSV | Comma Separated Value |
| DALY | Disability Adjusted Life Years |
| FN | False Negative |
| FP | False Positive |
| GPU | Graphics Processing Unit |
| HTN | Hypertension |
| LightGBM | Light Gradient Boosting Machine |
| ML | Machine Learning |
| RAM | Random Access Memory |
| RBC | Red Blood Cell count |
| RF | Random Forest |
| TN | True Negative |
| TP | True Positive |
| WHO | World Health Organization |
| XGBOOST | eXtreme Gradient Boosting |

# List of Equations

# ABSTRACT

A stroke is a potentially fatal illness that results from insufficient blood flow to a portion of the brain or bursts of arteries**.** It is the leading cause of disability and ranks second globally in terms of causes of death. Stroke is currently one of the most common reasons for hospital admission in many healthcare facilities and has become a serious public health concern in Ethiopia. Early prediction is necessary to reduce death and disability. Additionally, as risk factors for stroke include where you live, your lifestyle, your diet, the temperature, the environment, and socioeconomic issues, it is important to investigate the risk of stroke in different geographic places. The study aims to predict stroke risk using three ensemble learning models.Random Forest, XGBoost, and LightGBM are used in this study to predict stroke risk across the study area.The collected data is integrated, cleaned, normalization, the missing data is handled, and Synthetic Minority Over-sampling Technique (SMOTE) is used to handle a class imbalance in the data before evaluation started, Grid search technique is also used to find best performances of the models. The model is evaluated with accuracy, precision, recall, F1-score, and confusion matrix, and a correlation graph is also used to capture the relationship of the attributes. Random Forest had the maximum accuracy of 97.6% among models, followed by XGBoost at 96.1% and LightGBM at 92.9%.The study found that Discontinuation of Anti Hypertensive drug is the major risk factor for Stroke in the study.



***Keywords:*** *Stroke; Stroke Risk Prediction; Ensemble learning; Random Forest; XGBoost; LightGBM*

# CHAPTER ONE

# INTRODUCTION

## 1.1. Background

A stroke occurs when an artery leading to the brain clogs or bursts[1]. The brain requires a steady flow of nutrients and oxygen to function properly. An interruption in blood flow might cause a problem known as stroke[2]. It occurs suddenly. There are two categories of strokes: ischemic stroke and hemorrhagic stroke. In a hemorrhagic stroke, a relatively weak blood artery bursts and bleeds all over the brain; in an ischemic stroke, the blood vessel is blocked by the pressure of the clot [3].

According to World Health Organization (WHO), Stroke ranks second in the world in terms of causes of mortality and is the main cause of disability[1]. According to the 2022 version of the Global Stroke Factsheet, the lifetime risk of having a stroke has risen by 50% in the past 17 years, with an estimated 1 in 4 people expected to experience one at some point in their lives. Between 1990 and 2019, there was a rise of 70% in the incidence of stroke, 43% in the number of stroke-related deaths, 102% in the frequency of stroke, and 143% in Disability Adjusted Life Years (DALY)[1]. The most notable aspect is that lower- and lower-middle-income countries account for the majority of the world's stroke burden (86% of stroke-related fatalities and 89% of DALYs). For families with fewer resources, this disproportionate burden felt by lower- and lower-middle-income nations has created an unprecedented challenge[1].

Around the world, 15 million people get strokes each year. Five million of them pass away, and a further five million become permanently disable, burdening their families and communities[4]. The World Health Organization (WHO) lists stroke as the fifth most common cause of death for those between the ages of 15 and 59 and as the second most

common cause of death for those over 60. it is the leading cause of long-term disability regardless of age, sex, nationality, or race[1].

Stroke has an annual incidence rate of up to 316 per 100,000 in Africa, a frequency of up to 1,460 per 100,000, and a 3-year death rate of more than 80%, according to data released within the last ten years. Furthermore, a large number of Africans experience a stroke between the ages of 40 to 60, which can have a major impact on the victim, their family, and society[5].

Among non-communicable illnesses, stroke is one of the most frequent causes of disability and death. Over the past few decades, Sub-Saharan Africa has seen a notable increase in its incidence. It has been demonstrated that the mortality rate in this region is greater than in wealthy nations. Nonetheless, Ethiopia has a significant knowledge vacuum about stroke[3].The health system, which was formerly focused on communicable diseases with acute periods of illness, is facing a challenge from the rising burden of chronic disease in low-income nations like Ethiopia[6]. Ethiopia has a 6.23 percent stroke-related fatality rate overall and an 89.82 percent age-adjusted stroke mortality rate per 100,000 people[7].

Stroke mortality is very high. Paralysis, disorientation, and loss of speech and vision are possible symptoms for humans [4]. Since these are the negative effects of stroke, prediction of the condition will lessen these effects. As a result, employing ensemble learning facilitates the acquisition of more accurate and precise predictions.

Ensemble learning is a generic meta-approach to machine learning that aims to improve predictive performance, by merging the predictions from several models. Nowadays, numerous studies have employed ensemble approaches for disease prediction [8]. Ensemble learning methods consist of three main classes. These are bagging, stacking,

and boosting[8]. Currently, it become an interesting research area in healthcare. To get highly accurate prediction this study used Ensemble learning for prediction of stroke risk.

For this reason, the researcher motivated to ensemble learning prediction of stroke risk aid in resource allocation, early intervention, and healthcare planning, potentially saving lives and reducing the burden on the healthcare system. Moreover, it will contribute to the development of locally relevant healthcare solutions and improve the overall quality of stroke care in the study area.

### 1.2. Statement of the Problem

The primary research issue that drove this investigation is that stroke is the second most common cause of death and disability worldwide, particularly in low-income nations the death rate has increased [1]. Studies on the subject are lacking, and the knowledge about the disease is less in the population of Ethiopia[9].Although Stroke has emerged as a major public health concern in Ethiopia and is one of the most frequent causes of hospital admission in many healthcare facilities nowadays[9]. Because of this, it needs to be predicted early to decrease death and disability. The study which was taken in Hawassa University Referral Hospital, Sidama Region states that because of stroke risk over half of the stroke patients were released from the hospital With severe disabilities[10]. Also, It is necessary to research the risk of stroke in various geographic locations since risk factors for stroke include where you live, your lifestyle, your diet, the temperature, the climate, and socioeconomic variables[11]. For this reason in this study, Sidama regional state is selected.

Not taking antihypertensive medicine as prescribed or skipping doses in patients with hypertension has an increased risk of stroke[12].but, discontinuation or not taking anti-HTN medication is not listed as a risk factor in related studies so, as a prediction of the

models performed based on its features adding discontinuation of anti-HTN drug as input feature is helpful to achieve the researchers goal. Consequently, this study considers it a risk factor.

As there are many risk factors for Stroke, it needs to be predicted before the risk affects the individuals. For this purpose, it is advantageous to use the Ensemble learning method which is a type of ML. Ensemble learning techniques are the best for prediction purposes because Machine learning has several benefits over traditional methods, including the ability to recognize patterns that are too complex for humans to notice, the ability to make predictions based on a much larger data set, forecasting that is not influenced by human emotions or subjective opinions, the ability to adapt to changes in the data set, and the ability to become more accurate over time and less manipulable [13].Ensemble learning is a generic meta-approach to machine learning that aims to improve predictive performance, by merging the predictions from several models[8]. So, this learning is more advantageous than the traditional way. It combines many weak learners and helps them to become strong learners[14].Using ensemble learning for predicting stroke risk is more effective than the traditional method in the study area.

To create well-informed and suitable intervention measures, it is essential to generate accurate and timely information about the risk of stroke. Despite the fact that stroke risk is one of the main issues in the field of study, not much research has been done to support stroke risk prediction. In order to produce location-specific information on the stroke risk prediction in the study area, this research will be conducted.

## 1.3. Research Objectives

### 1.3.1. General objective

The general objective of this research is to develop an Ensemble learning technique for predicting stroke risk in the study area.

### 1.3.2. Specific objectives

1. To collect the datasets from the study area
2. To identify suitable Ensemble learning algorithms to be applied on the datasets in the study area
3. To identify the most common features or risk factors associated with stroke in the study area.

## 1.4. Research Questions of the study

1. Which Ensemble learning model is suitable to be applied to the dataset in the study area?
2. What is the performance of selected Ensemble learning models in predicting stroke risk within the study area?
3. What are the most common risk factors or attributes of stroke in the study area?

## 1.5. Significance of the Study

This study is focused on predicting stroke risk by Ensemble learning techniques in the Sidama Region, its goal is to minimize the incidence of stroke and increase health of the community. Stroke is known as global cause of disability and mortality, Ensemble learning approach is necessary to identify high-risk of stroke in the study area and implement preventive method.By focusing on the selected risk factors in the study develop the good performing approach .This research helps other researchers and policy makers to have comprehensive datasets in stroke risk .It identifies predictive features and

offers indicates into the effectiveness of Ensemble learning models, possibly serving as a model for advanced technology adoption in healthcare risk management for case of stroke.The results can shows healthcare policies, resource allocation, and strategies to reduce health problem, finally improving healthcare practices and reducing stroke frequency in the Regional state of Sidama.

### 1.6. Scope and Limitation of the study

Prediction models can be developed in a variety of methods. The researchers' main focus in this study is on ensemble learning algorithms for stroke risk prediction. The original dataset's 9203 instances with 13 variables including target class were utilized to predict the risk factors that contribute to stroke using ensemble learning methods.

The limitation of the study is it restricted to the Regional state sidama and analysis of real Stroke related data from the fixed years (2010–2015 E.C.).The other limitation of this study is Stroke risk is varies depend on demographic factors, lifestyle, and regional disparities this finding of the study not allow for to generalize for other regions of the country[11].Because of this the regional focus and the restricted risk factors considered, the study's findings and predictive Ensemble learning technique may have limited coverage for other regions. The other limitation of this study is most of the hospitals in the study area use manual way for recording patients' information and these data are not getting well-organized data as data recorded in electronic format for prediction purposes this limit the amount of data used for this study. And also there is no research conducted with the same datasets for prediction purpose. In addition to this smoking, genetics, Excessive alcohol use, Illegal drugs are risk factors for stroke there is no well-collected record in these risk factors[11] and also the data collection is limited to only three

hospitals in the Region such as Hawassa Referral Hospital, Yanet Internal Medicine Specialty Center, and Yirgalem General Hospital.

## 1.7 Methodology

The main aim of this study is to develop a prediction model for Stroke risk using ensemble learning techniques. The following methods are conducted to achieve the objectives stated in section 1.3.

### 1.7.1 Research Design

Making decisions about every detail of the research process and concentrating on the researcher's approach is part of the research design methodology. In research design, a variety of broad methodologies can be applied. This research is conducted using an experimental design. Experimental research is defined as a study that mostly adheres to a scientific research design.

Since the research topic needs several experiments and domain experts' opinions, the researcher used a mixed (both quantitative and qualitative) research approach. The researcher has collected the dataset for Ensemble learning purposes from Hawassa Referral Hospital, Yanet Internal Medicine Specialty Center, and Yirgalem General Hospital which are found in the Region.

### 1.7.2 Literature Review

To gain a conceptual understanding of the issue at hand, the researchers read a variety of research publications and journal papers that have been written about stroke and ensemble learning. To fill in the gap and develop the study's topic and research questions, the researchers also examined comparable works.

### 1.7.3 Data collection

The researchers employed data from secondary sources to achieve the study's goal. Secondary data, arranged from 2010 to 2015 E.C (six years of data), is the primary source of information for this study. The three hospitals in the region: - Yanet Internal Medicine Specialty Center, Yirgalem General Hospital, and Hawassa Referral Hospital—were the sources of the data.

### 1.7.4 Implementation tools

To conduct this research study, the researcher used different environments, tools, and libraries. The developing tool used in this research is colab using python.Colab is a free online cloud-based Jupyter notebook environment by using it we can train ML and DL models[15]. Notably, Python is the main programming language used in Google Colab. For many Colab users, the language of choice is Python due to its ease of use, large library, and broad use in the data science and machine learning communities[16].And there are different Libraries used in this developing tool like Numpy, Panda, Matplotlib, Seaborn which are pre-installed in colab[17]. Those software and designing tools have been selected using the most-fit strategy with different parametric analyses. The researchers used open-source tools for data analysis, designing a model, and validating it. In Ensemble learning, there are many different tools for the analysis and design so the researcher uses these tools, and also a programming language selected by the researchers is Python.

### 1.7.5 Evaluation Metrics

During classification training, the evaluation metric is essential to getting the best classifier. Therefore, choosing an appropriate assessment metric is crucial to differentiating and achieving the best classifier[18].From different types of evaluation

metrics the confusion matrix, accuracy, precision, recall, and F1-score, are used to determine the best ensemble learning model and predict stroke risk.

### 1.8 Ethical consideration

The research does not require the name of the Patient to protect the privacy and confidentiality of the patient treated in the given health center during data collection. The research is fully used for academic purposes and the data is collected with the help of Nurses and Medical Students who are interns. After discussing the purpose and method of the study, written permission was given from the medical directors of each hospital, the researcher collected the data before the data analysis carry out the study.

### 1.9. Organization of the Study

This research proposal organized into five chapters. The first chapter deals with introduction including background, statement of the problem, objectives of the study, research questions addressed the significance of the study, scope or delimitation of the study, and limitations of the study. The second chapter deals with review of literature includes introduction, overview of Ensemble Learning, an overview of stroke risk, and related Works. While the third chapter presents the research methodology including a description of the study area, the study design, the proposed stroke risk model, dataset collection and Source, dataset description, data preprocessing, ensemble learning model, and performance evaluation metrics followed by the study. The fourth chapter focuses result and discussion including introduction, prepared dataset for model, dataset splitting, ensemble learning model selection, hyperparameter tuning, Result, and Analysis, Compare the Results of the Models, and discussion. The fifth and the last chapter focused and the conclusion, recommendation and contribution of the study.

# CHAPTER TWO

# LITERATURE REVIEW AND RELATED WORKS

## 2.1. Introduction

In this section, a technical review of stroke risk and current methodologies for predicting stroke risk variables is provided, along with an analysis of related works that are relevant to the goal of this study. This helps to clearly describe and further analyze the research topic. Began by providing an overview of Ensemble learning, Ensemble learning algorithms, and an overview of stroke risk and its risk factors. Furthermore, a great deal of related research has been conducted which is helpful for our study.

## 2.2 Conceptual Literature Review

The objective of the conceptual literature review is to sort, characterize, and outline the link between concepts that are relevant to the study or topic—including relevant theory and empirical research[19].In this research before anything started the researcher reviewed different research papers, journals, books, and articles that are related and give relevant information for our research domain, the review is based on the quality of journals, articles, and research papers and focused on recent year works, because these are more helpful for this research.

## 2.3 Overview of Ensemble Learning

Ensemble learning is a machine learning techniques that perform better than traditional algorithms[20]. It describes techniques that combine several base models into a single framework to produce a more powerful model that performs better than the individual models. An ensemble method's performance is contingent upon a number of parameters, including the training and combination of baseline models[20]. In order to improve

performance on machine learning tasks, we employ ensemble learning[14]. When predicted accuracy is crucial for a variety of machine learning applications, ensemble learning is used. Finance, healthcare, image recognition, weather forecasting, sales forecasting, traffic prediction, and other fields are among the frequently used applications[21].Boosting, Bagging, and Stacking are the three most widely used methods for ensemble learning. Every one of these methods provides a different way to increase the accuracy of predictions[22].

### 2.3.1 Boosting

Boosting is an ensemble strategy that improves future predictions by learning from the mistakes made by the predictor in the past. By combining multiple weak base learners into a single strong learner, the approach greatly increases the predictability of models. Boosting involves placing underperforming learners in a sequential order so that underperforming learners can improve their predictive models by picking up tips from the next learner. There are various types of boosting, such as Boost (Extreme Gradient Boosting), gradient boosting, and adaptive boosting (AdaBoost)[23].

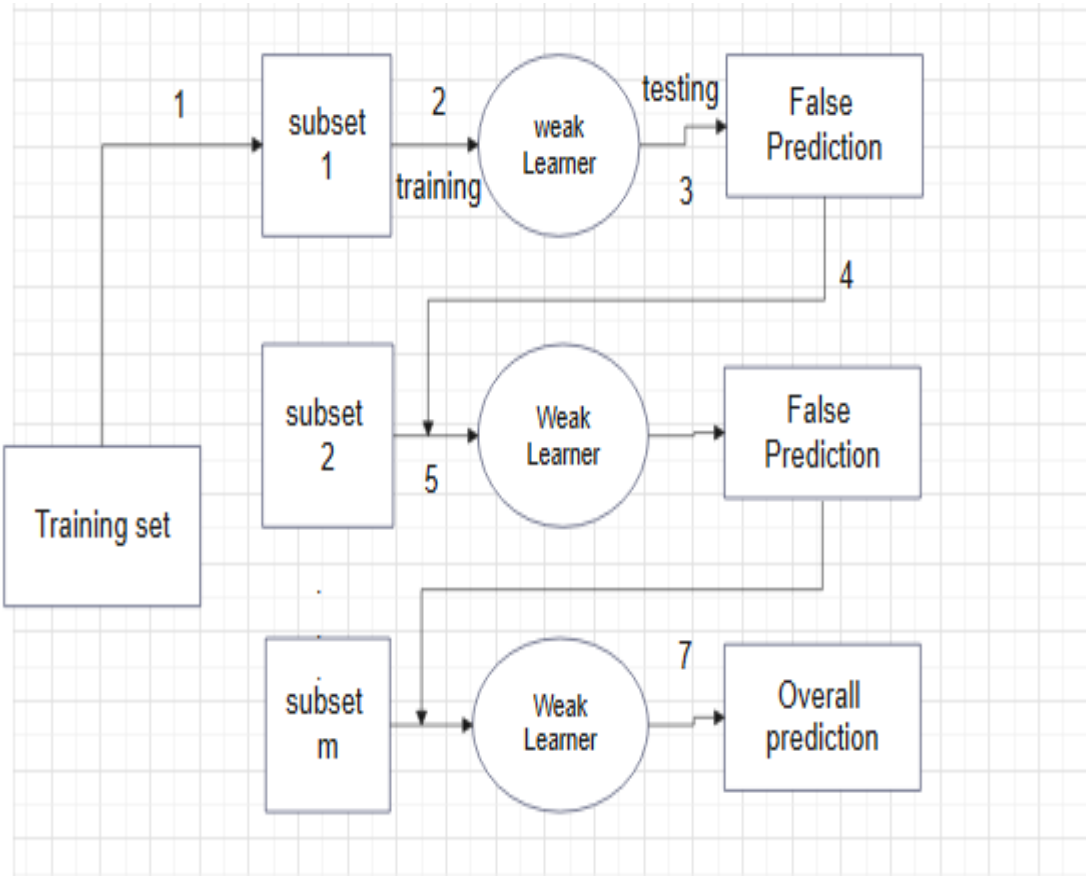



**Figure 1.1:**Boosting process

### 2.3.2 Bagging

Bagging or Bootstrapping aggregating is a term that is mostly used in regression and classification. Through the use of decision trees, it improves model accuracy and significantly lowers variance. By removing overfitting, which is a problem for many predictive models, variance reduction improves accuracy. A pair of bagging techniques are distinguished: bootstrapping and aggregation. The computational cost of bagging is one of its drawbacks. Ignoring the correct bagging technique can therefore result in increased bias in models[23].

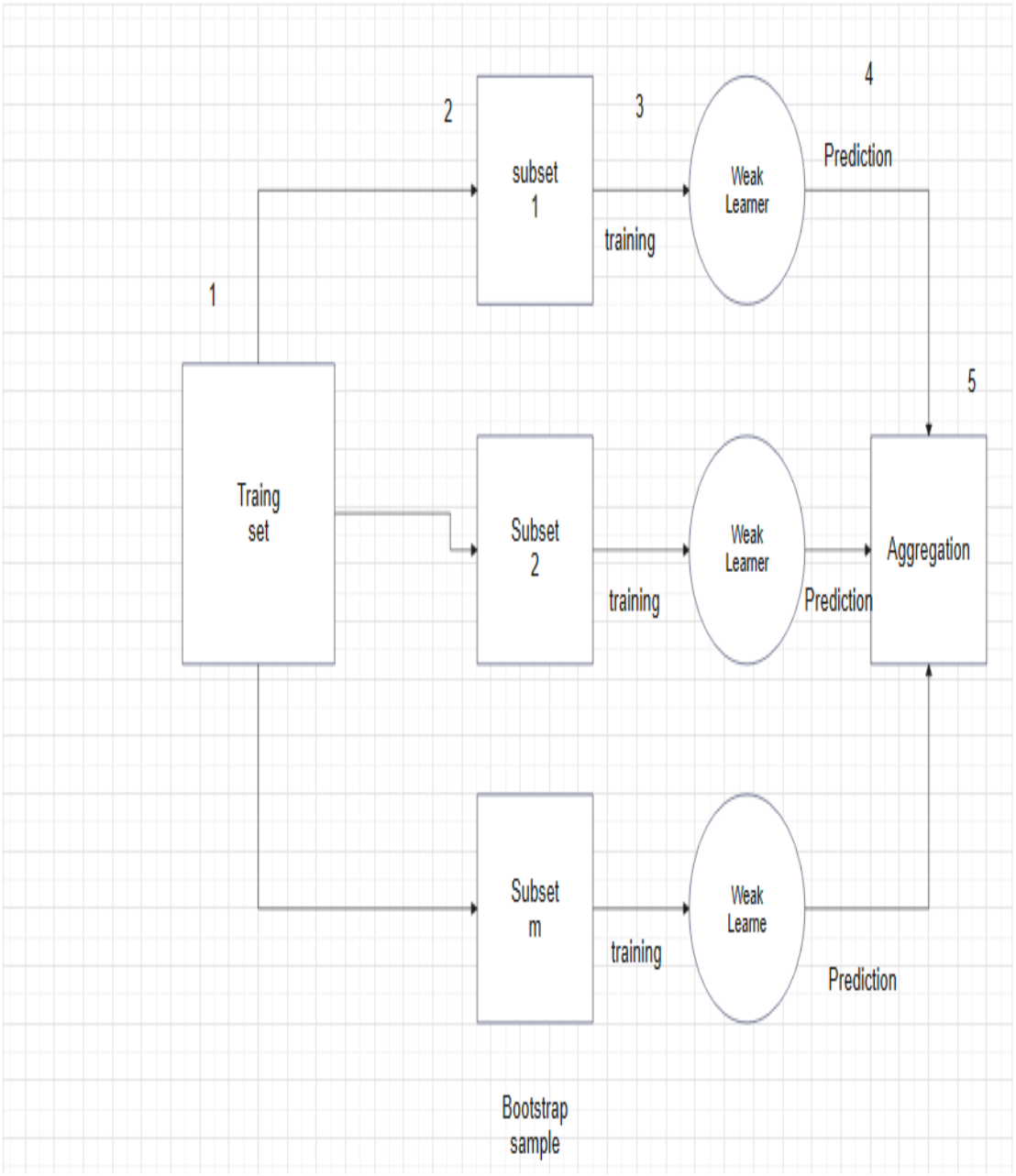



**Figure 1.2:**Bagging process

### 2.3.3 Stacking

Another ensemble technique is stacking, which is also known as layered generalization. This method functions by enabling a training algorithm to group together many predictions from similar learning algorithms. Regression, density estimations, distance learning, and classifications have all effectively used stacking[23].

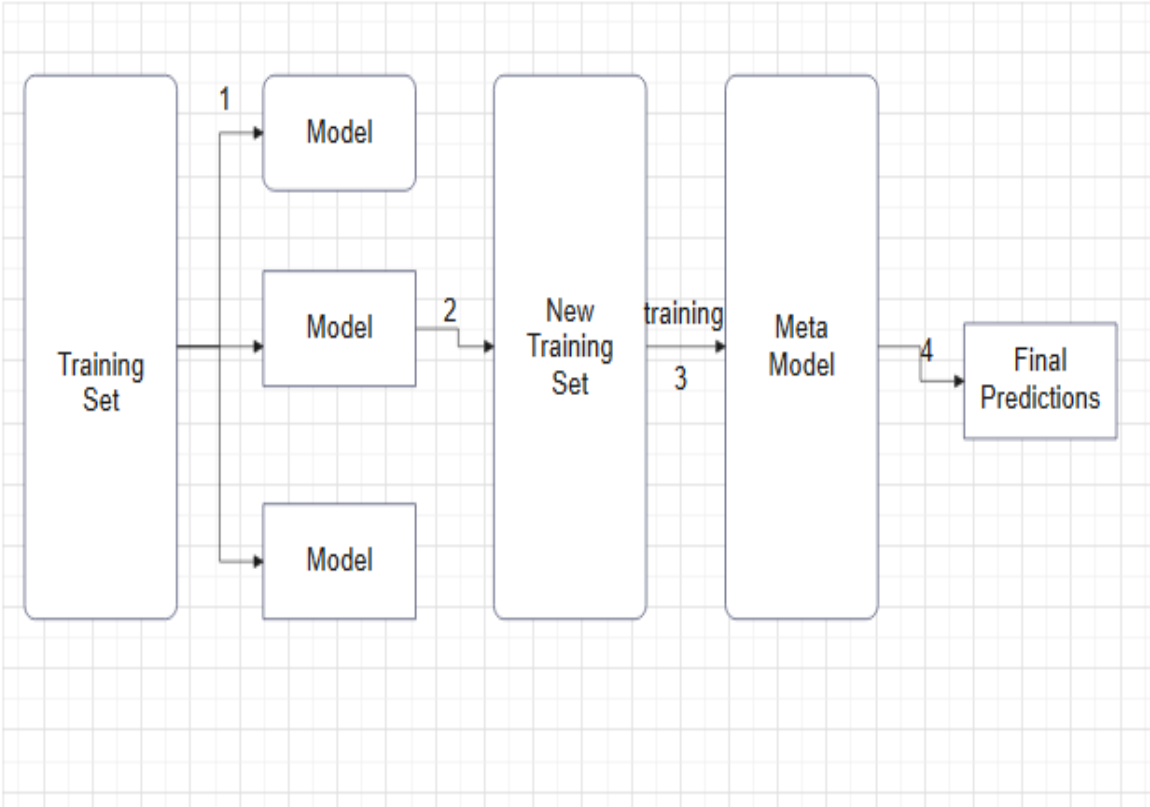



**Figure 1.3:Stacking process**

### 2.4 Overview of stroke risk

A stroke occurs when the blood supply to the brain is cut off. There is an urgent circumstance for the brain to function properly, it requires a steady flow of nutrients and oxygen. An interruption in the blood supply, even for a brief period, may result in complications. In just a few minutes without blood or oxygen, brain cells start to die. Brain function is lost when brain cells die[11]. There are two types of strokes which is defined next. Ischemic strokes are caused by a blockage of blood flow to the brain, while hemorrhagic strokes are caused by abrupt bleeding in the brain. Stroke risk is increased by numerous factors[11] .The risk factors classified in to two group: modifiable and non-modifiable and also there are some other factors which are not categorized in these. For both ischemic and hemorrhagic stroke, age, sex, and race/ethnicity are non-modifiable

risk factors. However, some of the more widely known modifiable risk factors include hypertension, smoking, diet, and physical inactivity.

Stroke is a common health issue that affects a lot of people globally, but its prevalence and risk factors vary depending on where you live. In Ethiopia, stroke deaths accounted for 6.23% of all deaths in 2017, according to WHO data released. Furthermore, the nation's age-adjusted stroke death rate is 89.82 per 100,000 people[24].

### 2.5 Stroke risk factors

A stroke can happen to anyone at any age, but the likelihood of getting one rises if you have specific risk factors; some of these risk factors are modifiable or controlled, while others cannot[11].

#### 2.5.1 Modifiable Risk Factors

Stroke risk factors that are modifiable risk factors includes[11]:

**High blood pressure:** Blood vessels (arteries) that feed blood to the brain can be harmed by blood pressure of 140/90 or greater.

**Heart disease:** The leading cause of mortality for stroke survivors and the second most significant risk factor for stroke is heart disease. Numerous risk factors for heart disease and stroke are similar.

**Diabetics:** Individuals who have diabetes have a higher risk of stroke than those who do not.

**Smoking**: An ischemic stroke nearly doubles your risk if you smoke.

**High level of red blood cells:** An appreciable rise in red blood cell count thickens the blood and increases the risk of clotting. Stroke risk is increased as a result.

**Cholesterol levels:** Elevated cholesterol levels have been linked to atherosclerosis, or the hardening or thickening of the arteries as a result of plaque accumulation. Deposits of calcium, cholesterol, and fats make up plaque.

**Excessive use of Alcohol:** Blood pressure rises when you consume more than two drinks each day. Alcohol abuse can result in stroke.

**Illegal drug:** Abuse of intravenous (IV) drugs increases the risk of stroke due to cerebral embolisms, or blood clots. Drugs like cocaine have been strongly associated with heart attacks, strokes, and numerous other cardiovascular issues.

**An irregular heartbeat:** Your risk of stroke may increase if you have certain heart disease types. Atrial fibrillation, an irregular heartbeat, is the most potent and manageable heart risk factor for stroke.

**Discontinuation of Anti-HTN drug:** If the HTN is not controlled it is forbidden to stop the medication.

### 2.5.2 Non-modifiable Risk Factors

Unchangeable risk factors for stroke include[11]:

**Age:** Your risk of stroke more than doubles for every ten years of life after the age of fifty-five.

**Race:** Compared to White people, African Americans have a significantly higher risk of stroke-related death and disability. This is partially due to the higher prevalence of high blood pressure among African Americans.

**Gender:** Men are more likely than women to have a stroke, but more women than men die from strokes.

**History of stroke:** Once you have experienced a stroke, your chances of getting another one are increased.

**Genetics:** Those who have a family history of stroke are more likely to experience one themselves.

### 2.5.3 Other Risk Factors

Stroke risk factors other than the above both risk factors are[11]:

**Geographical Area:** This could be due to variations in nutrition, smoking habits, race, and way of life between regions.

**Climate, season, and temperature:** Extreme temperatures are associated with a higher risk of stroke mortality.

**Economic and social aspects:** There is evidence to suggest that persons with poor incomes are more likely to have strokes.

## 2.6 Review of Related Research Works/Literature

Studies have examined different aspects of stroke risk. The researcher has reviewed various journal articles related to machine learning model prediction of stroke risk to achieve the study's goal and improve this research paper. To show how previous studies relate to one another, identify any gaps in prior related work, and determine what material is now available in the study domain, the researcher looked through the following literature relevant to the problem domain. A few of these articles are included below.

In this research [25], the researcher attempted to create a classifier model using machine learning classification techniques that can predict stroke. The machine learning method achieved the highest accuracy, precision, recall, and F1-score of 94%-95% based on the Random Forest classifier, as per the experiment results. Furthermore, the correlation

values were computed to determine the extent to which a given factor influences the target feature (having a stroke) and whether it also impacts other features. Nevertheless, ensemble models must be used by considering other accuracy metrics to increase accuracy and decrease error.

By using algorithms like the AdaBoost classifier, artificial neural network, decision tree classifier, k-nearest neighbor (KNN) classifier, random forest, stochastic gradient descent (SDG), support vector machine (SVM), and XGBoost classifier to predict the risk possibility of stroke, this study[26] aims to develop a classifier model to predict the occurrence of stroke disease. Next, these eight conventional classifiers are subjected to a voting classifier implementation. The researcher discovered 98% accuracy in predicting the risk factor for stroke after evaluating several machine learning algorithms with voting classifiers, which is superior to other models. Although the discontinuation of anti-HTN drug is not included as an input feature of Stroke in this study, it is related to the planned study because it is one of the risk factors for stroke.

To create a classifier model that can predict the occurrence of stroke, research was done[27]. Six different models were trained for accurate prediction using machine learning algorithms, including Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, and Naïve Bayes Classification. Naïve Bayes is the algorithm that did the best on this task, with an accuracy of about 82%. The proposed research is pertinent to the problem domain of the research study. Employing ensemble-based machine learning algorithms in this work may help the researchers boost accuracy because these algorithms may forecast outcomes by averaging the performance of the decision tree's weak learners.

The purpose of this work[28] was to create machine learning models for the prediction of stroke in an older Chinese population using unbalanced data. The unbalanced data in this study were processed using data balancing approaches such as synthetic minority over-sampling technique (SMOTE), random under-sampling technique (RUS), and random over-sampling technique (ROS). To predict stroke, machine learning techniques such as random forest (RF), support vector machine (SVM), and regularized logistic regression (RLR) were applied. The sensitivity and AUC significantly increased with reasonable precision and specificity after applying data balancing procedures; the maximum values for sensitivity and AUC were 0.78 (95% CI, 0.73–0.83) for RF and 0.72 (95% CI, 0.71–0.73) for RLR. The AUCs of all three machine learning techniques significantly improved ($p < 0.05$) in the balanced data sets compared to those for RLR, SVM, and RF in the imbalanced data set. Even though 1131 participants—56 stroke patients and 1075 non-stroke participants—were included in the prospective cohort from which the data were collected, more data is needed to achieve greater precision.

This study[50] used demographic and diagnostic data from Hallelujah and Zewditu hospitals to develop stroke risk prediction models with Logistic Regression, SVM, and Random Forest (RF). After data preparation and cleaning, the models were evaluated, with Random Forest achieving the highest accuracy (99.3%). Based on these results, RF was the optimal model.Since stroke risk varies by region, similar studies are needed in other parts of the country. This study missed key risk factors, such as discontinuation of antihypertensive medication and residency, while including height, which lacks a direct link to stroke.

**Table 2.1 Summary of Related Works**

| Authors | Technical Description | Contribution &conclusion | Critical Remark |
|---|---|---|---|
| H. Al-Zubaidi, M. Dweik, *et al*. [26] | Random Forest classifier outperforms the Voting classifier, Decision Tree, Logistic Regression, and SVM. | The experiment findings showed that the machine learning approach produced the highest accuracy, precision, recall, and F1-score of 94%-95% based on the Random Forest classifier. | Evaluation metrics for the Random Forest classifier are regularly in the 94%–95% range, indicating its strong performance. This small range, however, points to areas that could use improvement, including adjusting hyperparameters, include pertinent information, or investigating sophisticated ensemble techniques. Despite its effectiveness, the model's performance begs for improvement in order to overcome the present barrier. |
| R. Islam, S. Debnath, *et al*. [27] | The voting classifier performs better than AdaBoost, ANN, RF, SDG, SVM, XGBoost, and KNN | After comparing other machine learning algorithms using voting classifiers, the researcher found that 98% accuracy in identifying the stroke risk factor is better than other models. | The Voting classifier's ability to enhance model performance through ensemble learning was examined in this study. It highlights that the study failed to include the recognized stroke risk factor of stopping antihypertensive medications as an input feature. This gap restricts the model's applicability and points to areas where its forecast accuracy could be raised. |

| | | | |
|---|---|---|---|
| G. Sailasya ,*et al*. [28] | Naïve Bayes outperforms Logistic regression, decision tree, RF, KNN, and SVM | With an accuracy of almost 82%, Naïve Bayes is the algorithm that performed the best on this task. | With an accuracy of 82%, Naïve Bayes beats a number of conventional algorithms, yet this performance is still deemed inadequate for real-world uses. Ensemble-based machine learning techniques, which have demonstrated better results in comparable tasks and potentially improve the model's accuracy and robustness, could be incorporated into the study. |
| Y. Wu , *et al*. [29] | By applying machine learning techniques in balanced data the researcher got a better result | The researcher got better results in balanced data than unbalanced data using balancing techniques | Although the researcher used balancing approaches to obtain better results with balanced data than with unbalanced data, the performance might be further improved by include a larger and more varied dataset. In order to overcome any potential shortcomings of the current dataset, expanding the sample size would probably enhance model generalization and overall accuracy. |

# CHAPTER THREE

# PROPOSED SOLUTION

In this chapter, the research methodology used to reach the study goal of this research is discussed. The main objective of this study is to find a predictive model for Stroke risk using an Ensemble learning model. Below, the study area, research material and approach, and methods used are discussed in detail.


**3.1:** Research process of the proposed study

### 3.1. Description of the Study Area

This study is conducted in Sidama regional state of Ethiopia. Sidama is one of the newest regions in Ethiopia since June 18, 2020 previously it was one of the 100 and more zones of the country. The Area of this region estimated using Arc GIS Pro measurement gave as geographic extent of approximately 7,700 square kilometers within 5°45′–6°45′N latitudes and 38°5′–39°41′E longitudes. In September 2020, The population of Sidama Region has an estimated of 4.3 million and is bordered in the West Woliyta Zone which is separated by Bilate river, Oromia Region in the north and east border and in the south bordered with Oromia Region, and Gedeo Zone. To study the risk of stroke in the Sidama region the study selects three hospitals based on the number of patients visiting the hospitals such as Hawassa Referral Hospital, Yanet Internal Medicine Speciality Center, and Yirgalem General Hospital. The first two Hospitals were founded in the capital of the region Hawassa and the third was founded in Yirgalem City. As stroke risk depends on the demographic area and the lifestyle of the population it needs to study the major risk factors of stroke in the Sidama Region[29].

### 3.2. The Study Design

Research design is considered the blue-print and cornerstone of any study since it facilitates various research operations. This study uses a mixed method (Qualitative and Quantitative) research approach which is often considered the most powerful since it combines the power of quantitative and qualitative research approaches. The quantitative approach helps to gather quantitative data from different hospitals, including patient demographics, medical records, diagnostic tests, treatments, and outcomes, and qualitative research approaches help to identify and consult with research experts in the study area or domain, such as clinicians, researchers, and healthcare professionals. Many

researchers use mixed methods approaches as a way to increase the validity of their research process[30]. The study was aimed at predicting stroke risk as well as finding the major risk factors of the study. A cross-sectional survey was administered to collect quantitative and qualitative data used for the study.

In this study, experimental research is conducted using the research design. To explain the research questions and gain an understanding of the problem, a literature review is conducted. Data collection happens after the formulating research question is developed. The process of formulating and designing a solution resolves the issue. The suggested remedies are put into practice, and assessed, and the evaluation's findings are evaluated. Lastly, present the outcome of our solution.

### 3.3 The Proposed Stroke Risk Model

In the Sidama Region, where there is a risk of stroke, the suggested model is used to predict the risk of stroke after data collection, models are created using ensemble learning algorithms such as XGBoost, Random Forest, and LightGBM. They are trained on a training set that includes the training dataset. using Model Evaluation like F-measure metrics, precision, recall, and accuracy are used to assess the models. The best prediction model was chosen using the evaluation results. Lastly, the Prediction model is assessed and chosen by the outcomes determined using the model evaluation metrics. A prediction model for stroke risk in the Sidama region was developed using the best model that was chosen and found the major stroke risk of the study area.

### 3.4 Dataset Collection and Source

In this research, to predict the risk of stroke, we must gather, preprocess, and construct an Ensemble learning model. The most valuable aspect of research when developing

Ensemble learning models is the data. The suggested model needs to be trained on a sizable number of records with complete information to accurately predict the risk of stroke. To create a prediction model for the region, the researchers in this study gathered data from three hospitals in the Sidama region: Yirgalem General Hospital, Hawassa Referral Hospital, and Yanet Internal Medicine Speciality Center, which are located in various parts of the region. The data collected from years between 2010 to 2015 E.C and uses 12 features. To make the original data relevant for ensemble learning, the researchers encoded it into digital formats. The information gathered from hospital records is used in this study.

The data collected from the first two hospitals Yirgalem General Hospital, and Hawassa Referral Hospital was recorded on paper but, the data collected from Yanet Internal Medicine Speciality Center was recorded in digital form. Both the data in paper and digital format were selected based on the selected risk factors, then stored in Excel format and changed into CSV format to make it suitable for the proposed purpose.

The data collection method starts from a primary source and proceeds to a secondary source. The actual source of the data is a secondary source. However, to decide the risk factors interviewing with the specialists in the area was essential, and based on the data obtained from the interview the data is collected from secondary sources which are recorded in the hospitals in both paper and digital form.

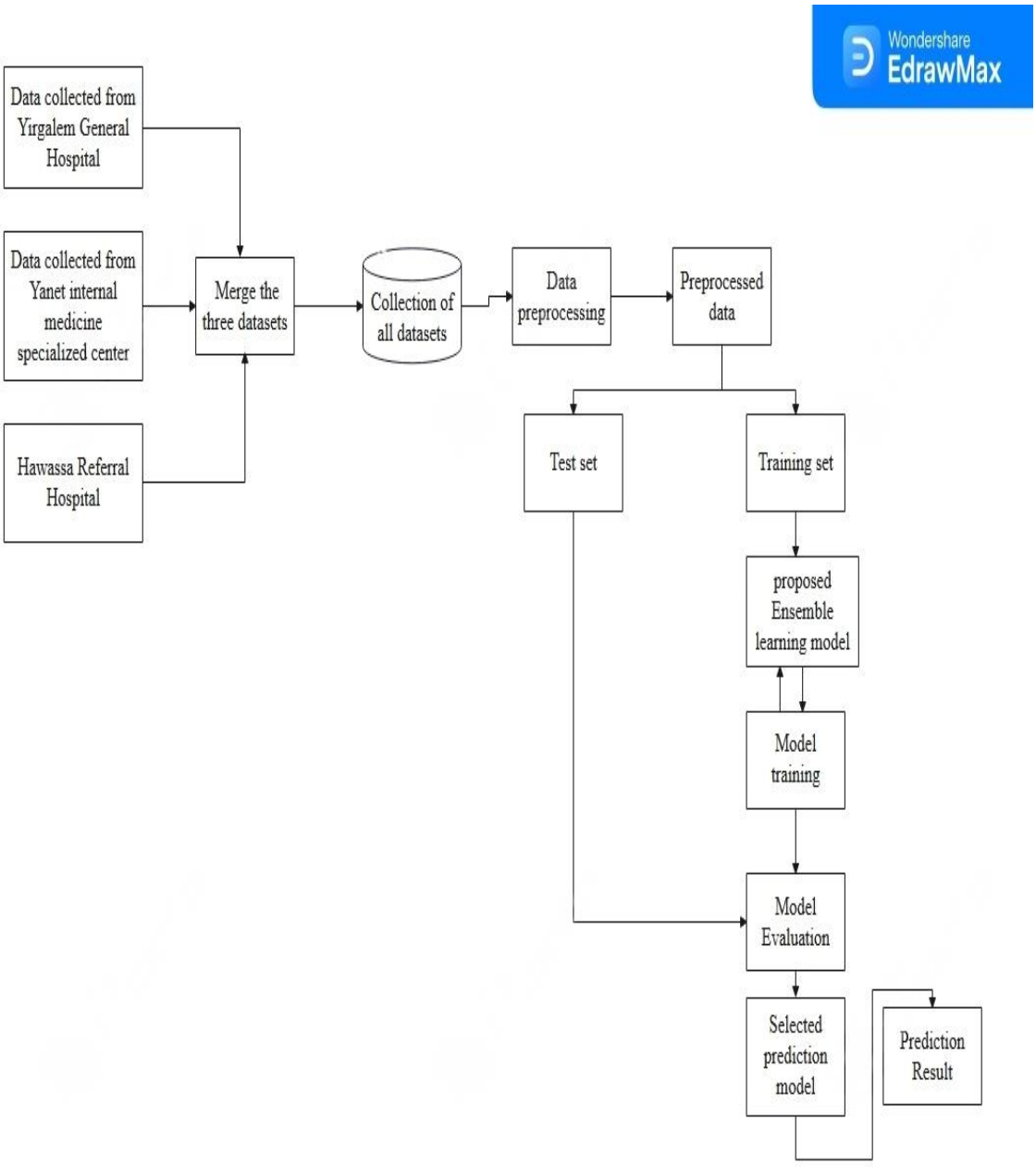



**Figure 3.2: Data Collection and Evaluation Method**

Large datasets, such as those found in healthcare records, are challenging to evaluate and create a machine-learning model for since loading them would cause our system to run out of memory. Additionally, a number of limitations, including financial, material, and scheduling constraints as well as the availability of the data itself, limit the amount of data

that can be collected for the planned study. Purposive sampling techniques were thus used to allow us to get accurate information about the population based on data from a sample, or subset of the population, instead of having to investigate each and every individual.

### 3.5 Dataset Description

The table below describes different attributes used in this dataset and describes their meaning and the data type that is expressed.

**Table 3.1: Dataset descriptions**

| S.No | Feature Name | Description |
|---|---|---|
| 1 | Age | Age of the Patient |
| 2 | Sex | Sex of Patient |
| 3 | Residence | Living place of the patient |
| 4 | Blood Pressure | Blood Pressure of the patient |
| 5 | RBC | Red blood cell count of the patient |
| 6 | Pulse rate | The heartbeat of the patient |
| 7 | Hypertension | Hypertensive or non-Hypertensive |
| 8 | Diabetics | Diabetic or non-diabetic |
| 9 | Heart Disease | Patients with heart disease or free from heart disease |
| 10 | BMI | The body Mass Index of the patient |
| 11 | Discontinuation of Anti-HTN Drug | The patient discontinued the HTN drug or continued taking the Anti-HTN drug |
| 12 | Cholesterol | The blood Cholesterol level of the patient |
| 13 | Stroke | Stroke case and no stroke case (target class) |

### 3.6 Data Preprocessing

As the data is collected from different sources and stored in different formats it needs to be preprocessed before training the model. To train a model with high-quality data, preprocessing procedures such as data integration, data transformation and standardization, data normalization and scaling ,missing value extraction from a dataset, balancing the datasets, feature selection based on significance, data training, and test splitting were carried out.

Python libraries are required in order to process data. To conduct experiments, import Numpy, pandas, Matplotlip, Scikit-Learn, and a few Python methods and classes. The resulting Python code is as follows.

#### 3.6.1 Data Integration

Data Integration is the process of merging data from several sources to present in a uniform format[31]. In our case, the data is collected from three different hospitals and stored in various formats and also some attributes or features are registered in different names to combine the data from different sources we need to select the same feature names for the attributes. While the data from Yirgalem General Hospital and Hawassa Referral Hospital were recorded on paper, the data from Yanet Internal Medicine Speciality Center was captured digitally. Thus, using the risk factors we have chosen for our study, we arrange the data that has been gathered from the three hospitals in an MS Excel format.

#### 3.6.2 Data Standardization and Transformation

Before training a model, the data must be standardized at the same scale to prevent model overfitting. In our case, some features have a different scale or standard, so that, there is a

need for standardization. Some features like Blood pressure, RBC, Pulse rate, BMI, and cholesterol have different numeric values so, it needs to be scaled.

All of the dataset's data types should be converted to numeric values because this allows us to represent the data as 1 if a symptom is associated with a particular risk factor and 0 otherwise. In other words, all of the columns that include symptoms are filled with 1 if a symptom is associated with a particular disease and 0 otherwise. To train a model, certain of the categorical data type columns, such as those related to sex, residency, hypertension, diabetes, heart disease, Discontinuation of anti-HTN drugs, and disease type, were also converted to numeric values. Since handling categorical data is a common practice, labels with data types of categorical are transformed into numeric using a label encoder. Using this system, each label is assigned a unique number. Therefore, the values of Non-Stroke and Stroke in the disease type (stroke class) column are changed to 0 and 1, respectively as shown in Table 3.2. Additionally, the values for the sex column—male and female—were changed to 1 and 0, respectively. Additionally, the Yes and No columns for residency, hypertensive, diabetics, heart disease, and Discontinuation of anti-HTN Drugs were changed to 1 and 0, respectively as shown in Table 3.3.

**Table 3.2: Data Transformation Target class (Stroke)**

| Disease type | Transformed value |
|---|---|
| Non-Stroke | 0 |
| Stroke | 1 |

**Table 3.3: Data Transformation for Categorical Values**

| Sex | | Risk Factors (Residency, Hypertensive, Diabetics, Heart disease, and Discontinuation of anti-HTN drug) | |
|---|---|---|---|
| Male | Female | Yes | No |
| 1 | 0 | 1 | 0 |

### 3.6.3 Finding Missing Value

A missing value in the training data can cause an error and prevent the model from being able to learn from the data. Thus, in order to train a model and get acceptable performance on it, it is necessary to identify the missing value and either replace it with the appropriate data or remove these records. We must remove the two columns labeled **"Smoking"** and **"Alcoholic"** right away because they contain values that are unknown. There are **273** missing values in this dataset besides these columns, and it has been determined to drop all entries that have missing values. Additionally, the remaining original dataset has been forwarded for additional handling.

## 3.7 Ensemble Learning Model

This Research aims to develop an Ensemble learning-based prediction of stroke risk based on hospital-recorded datasets three Ensemble learning algorithms were used. In this study to compare the model building based on each algorithm's performance. Based on a thorough experimental comparison and their popularity in research done on stroke risk prediction, these algorithms are chosen. To create an ensemble learning prediction of stroke risk, this study chose three ensemble learning algorithms: Random Forest, XGBoost, and LightGBM.

### 3.7.1 Random Forest

A random forest is a well-known ensemble learning Algorithm that is used to predict and classify the given dataset. As we all know, a forest is made up of lots of trees, and if it has more trees, it is more robust. Similarly, In the Random Forest Algorithm, as the number of trees increases, accuracy and capacity for solving problems also increase. A Random Forest classifier uses multiple decision trees on different dataset subsets and averages

them to increase the dataset's predictive accuracy. Its foundation is the idea of ensemble learning, which is the process of merging several classifiers to solve a challenging issue and enhance the model's functionality. One well-liked machine learning technique for classifying and predicting data is the random forest [47].

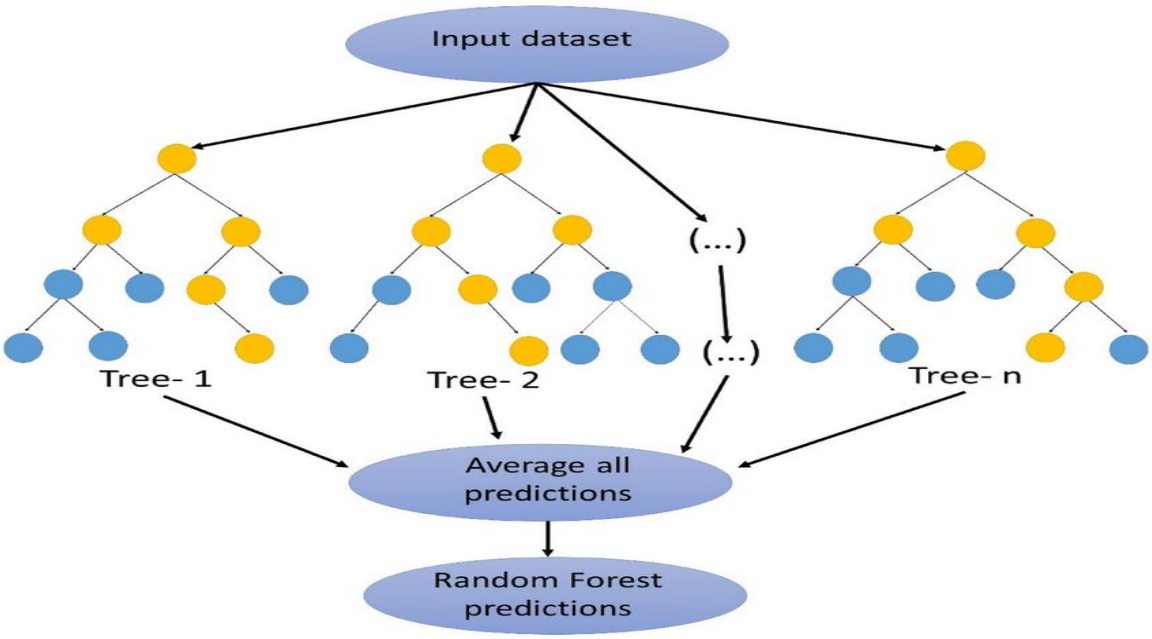



**Figure3.3:** Random Forest[32]

### 3.7.2 XGBoost

The ensemble learning algorithm XGBoost makes predictions by combining gradient boosting with an ensemble of decision trees. It has won multiple machine learning competitions and is widely used in the field of data science. XGBoost builds a powerful learner from weaker learners. It gradually adds more models. To obtain an optimal solution, the models in the chain rectify the errors made by the weaker models. We call this ensembling. The ones we call weaker models are the base model [48].

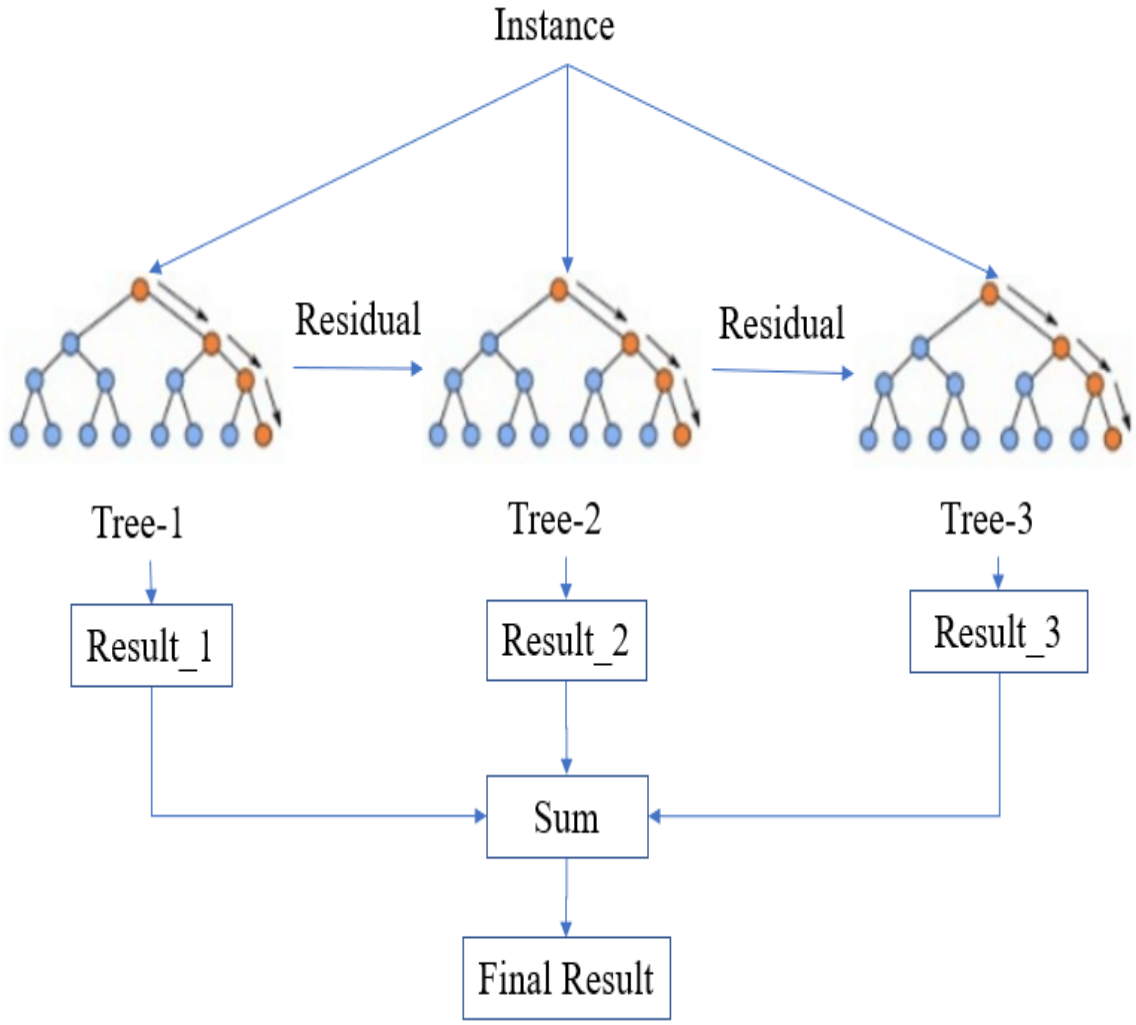



**Figure3.4:** XGBoost[34]

### 3.7.3 LightGBM

This is the third type of ensemble learning algorithm used for this study. A gradient boosting technique called LightGBM is an ensemble learning framework that builds a strong learner by gradually adding weak learners in a gradient descent fashion. Gradient-based One-Side Sampling is one of the strategies used to improve memory utilization and training time[33]. LightGBM uses decision trees as weak learners specially classification trees for categorical outcome as the case of this study.

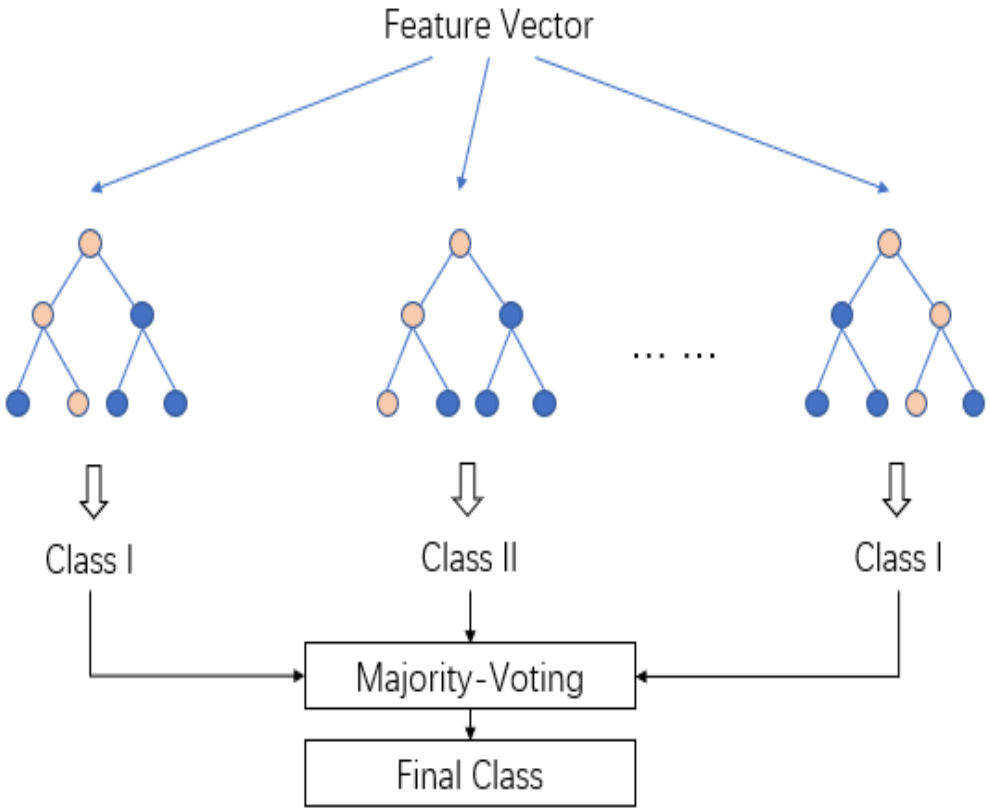



**Figure 3.5:** LightGBM[34]

In this study, these three algorithms such as XGBoost, Random Forest, and LightGBM are selected for prediction purposes and ensemble by using stacking.As we know, these are ensemble algorithms based on trees for supervised machine-learning problems [35]. When compared to other algorithms, ensemble algorithms especially those that use decision trees as weak learners, have several advantages, as their algorithms are simple to comprehend and illustrate, can manage both categorical and ordered data, are powerful opposed over fitting well, Scaling of inputs is not necessary, faster than algorithms like neural networks or support vector machines and When compared to boosting and bagging algorithms decision trees high variance and overfitting will result in decreased accuracy[35]. Several machine learning algorithms are combined in ensemble techniques

to produce predictions that are more accurate than those produced by a single[36]. So, using these algorithms leads to more accurate predictions.

### 3.8 Performance Evaluation Metrics

Evaluation metrics are quantitative measures that are used to evaluate the efficacy and performance of a statistical or machine-learning model; they aid in the comparison of various models or algorithms and offer insights into the model's performance[37]. In the study, different evaluation metrics are used accuracy, precision, recall, f1-score, and confusion matrix.

**Confusion matrix:**

Confusion matrix is one of the evaluation metrics used in this study. An n X n matrix, where n is the number of expected classes, is a confusion matrix. In this study the target group called Stroke consists of 2 classes so n=2 for the current case, we obtain a 2 X 2 matrix. It is a metric used to assess how well ensemble learning tasks perform while producing two or more classes as an output. The figure below having four distinct combinations of expected and actual values is called a confusion matrix.

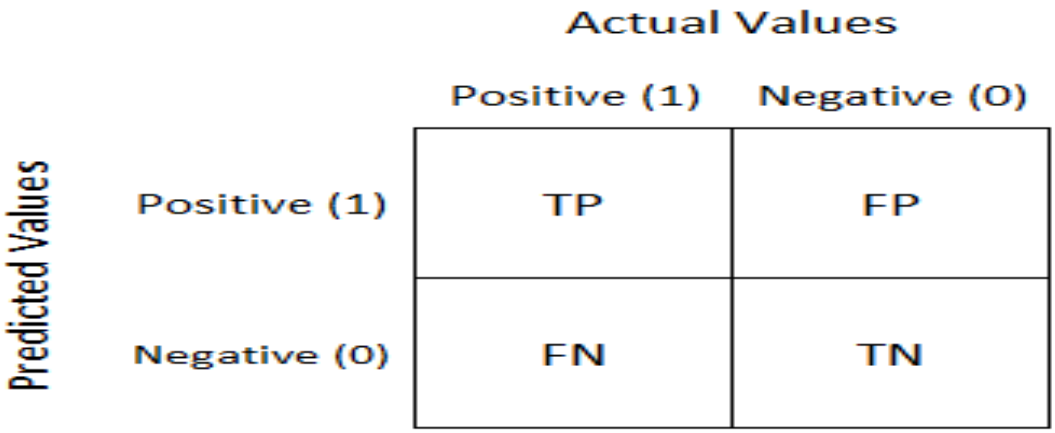



**Figure 3.6** Confusion matrix

**True Positive (TP):** - indicates both the actual class and the predicted class values are positive

**False Positive (FP):** - indicates the actual class is negative and the predicted class value is positive.it is also called a Type 1 error.

**True Negative (TN)**: -indicates both the actual class and the predictive class value are negative

**False Negative (FN)**: -indicates the actual class is positive and the predicted class value is negative.it is also called a Type 2 error.

**Accuracy**

One of the most well-known and widely used measures for assessing a classification model is accuracy. It is calculated by dividing the fraction of corrected predictions by the total number of samples[38]. The value for accuracy is calculated by using the equation below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad \text{.....Equation 4.1 Accuracy}$$

**Precision**

The precision metric provides the ratio of true positive results to the total number of positive results predicted by the model. It provides a response to the query, "How many of our optimistic predictions came true?"[39]. Its value is calculated using the equation below.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{.....Equation 4.2 Precision}$$

**Recall**

Recall, also known as true positive rate, addresses the issue of how well the model finds all the positives by providing an answer to the following query: "Out of all the data points that should have been predicted as true, how many did we correctly predict as true?"[39].Its value is calculated by using the equation below.

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{..... Equation 4.3 Recall}$$

**F1-Score**

Recall and precision are combined into one metric called the F1 Score. F1 can be used to gauge how well our models make the trade-off between precision and recall, as we have discovered to be necessary[39]. Its value is calculated by using the equation below.

$$\text{F1-Score} = 2.\frac{Precision\ x\ Recall}{\text{Precision+Recall}} \quad \text{.....Equation 4.4 F1-Score}$$

### 3.9 Libraries Used for Proposed Model Development

Many Python libraries were employed to facilitate and expedite the construction of the suggested model through Python programming. Instead of writing the same code for every program, we have imported the relevant library for the one we want to use. The below table shows the libraries used for developing the models.

**Table 3.4 Important libraries used in the proposed study**

| No. | Name of library | Use |
|---|---|---|
| 1 | NumPy | Library used for working with arrays. |
| 2 | Pandas | Used to load and analyse data. |
| 3 | Sklearn | Used to preprocess the data and develop a machine learning Model. |
| 4 | Seaborn | A library that provides a Python visualization of a library based on matplotlib |
| 5 | Matplotlib | Used to visualize the data |

# CHAPTER FOUR

# EXPERIMENTAL RESULTS AND DISCUSSION

## 4.1 Introduction

This chapter covers how the models work and the execution of experiments using the suggested framework. Furthermore, this experiment and evaluation exhibit the chosen ensemble learning models, including Random Forest, Extreme Boost (XGboost) and Light Gradient Boosting Machine (LightGBM), before the model training starts data preprocessing completed and demonstrate the realization of the architecture described in Chapter 3. With the help of Python's Sklearn Library, various performance evaluations, including Precision, Recall, F-measure, accuracy, confusion matrix, and correlation, are considered for each of the chosen algorithms.

## 4.2. Experimental Setup

To conduct the experiment work for this study, the researcher used a machine that has the following qualifications:

✓ Laptop computer with Intel(R) Core (TM) i5

✓ 4GB RAM

✓ CPU with 2.6 GHz speed

✓ Hard Disk storage-500 GB

✓ Window 10

✓ System type- x64

✓ Software -Google Colab

✓ Edraw Max

### 4.3 Prepared Dataset for Model

After collecting data on strokes from three hospitals: Hawassa Referral Hospital, Yanet Internal Medicine Specialty Center, and Yirgalem General Hospital, it merged. The data collected from these hospitals are generally counted as 9203 with 15 attributes including the target group. From these data 6294 instances were grouped in “Non-Stroke (0)”, and 2909 instances were grouped in “Stroke (1)”. This information is shown in Figure 4.1 below.

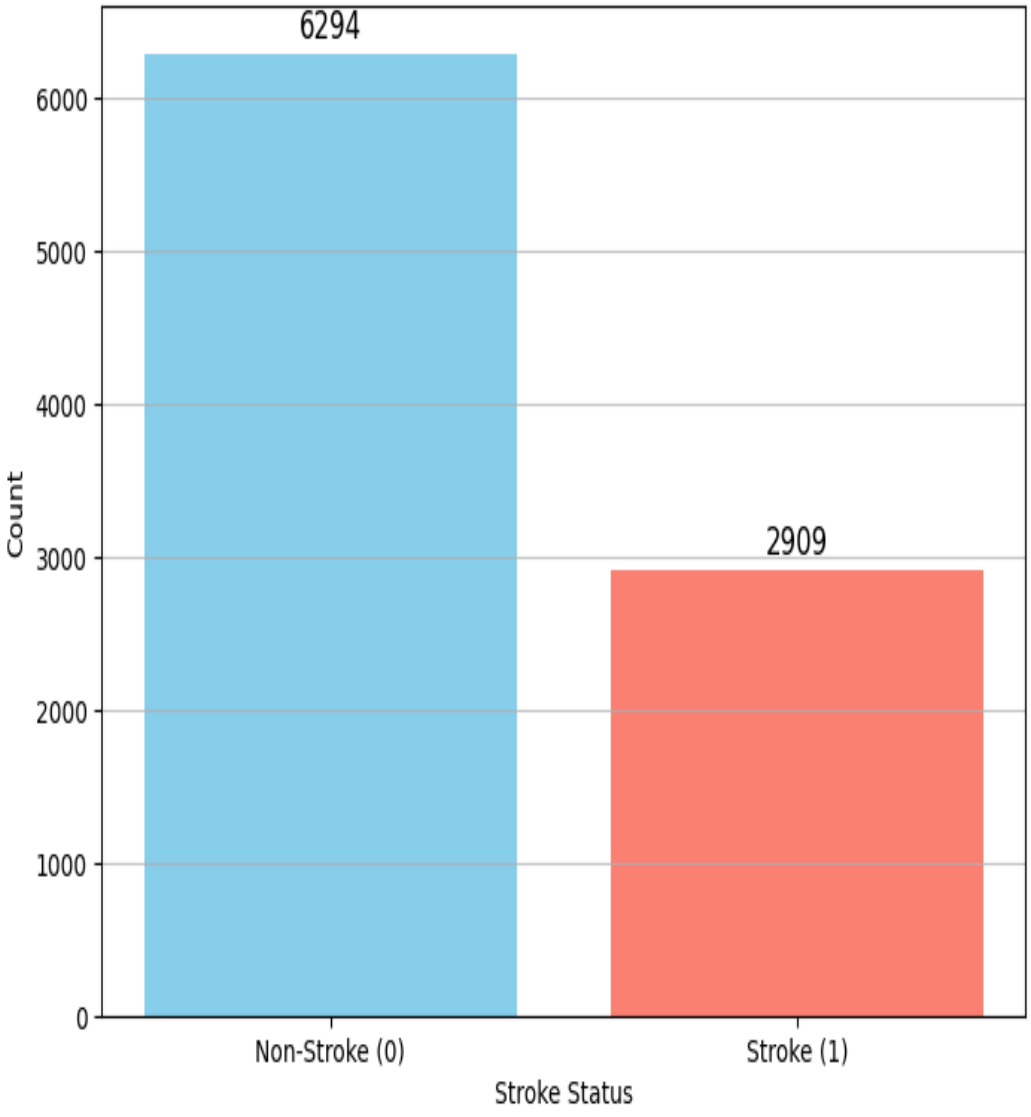



**Figure 4.1:** Distribution of Stroke

During the data collection process at these three hospitals, the researcher was unable to get organized information about two risk factors: the patients' alcoholism and smoking habits. These two risk factors were therefore not employed for prediction in this study. Furthermore, NULL values were present in  records, which were also ignored. The dataset had 8,930 instances after the entries with NULL values were eliminated, which meant that 273 records were no longer included. There are 13 attributes in the dataset (alcohol and smoking status are omitted). 2,838 occurrences are classed as "Stroke (1)" and 6,092 instances as "Non-Stroke (0)." The distribution is depicted in Figure 4.2 that follows.

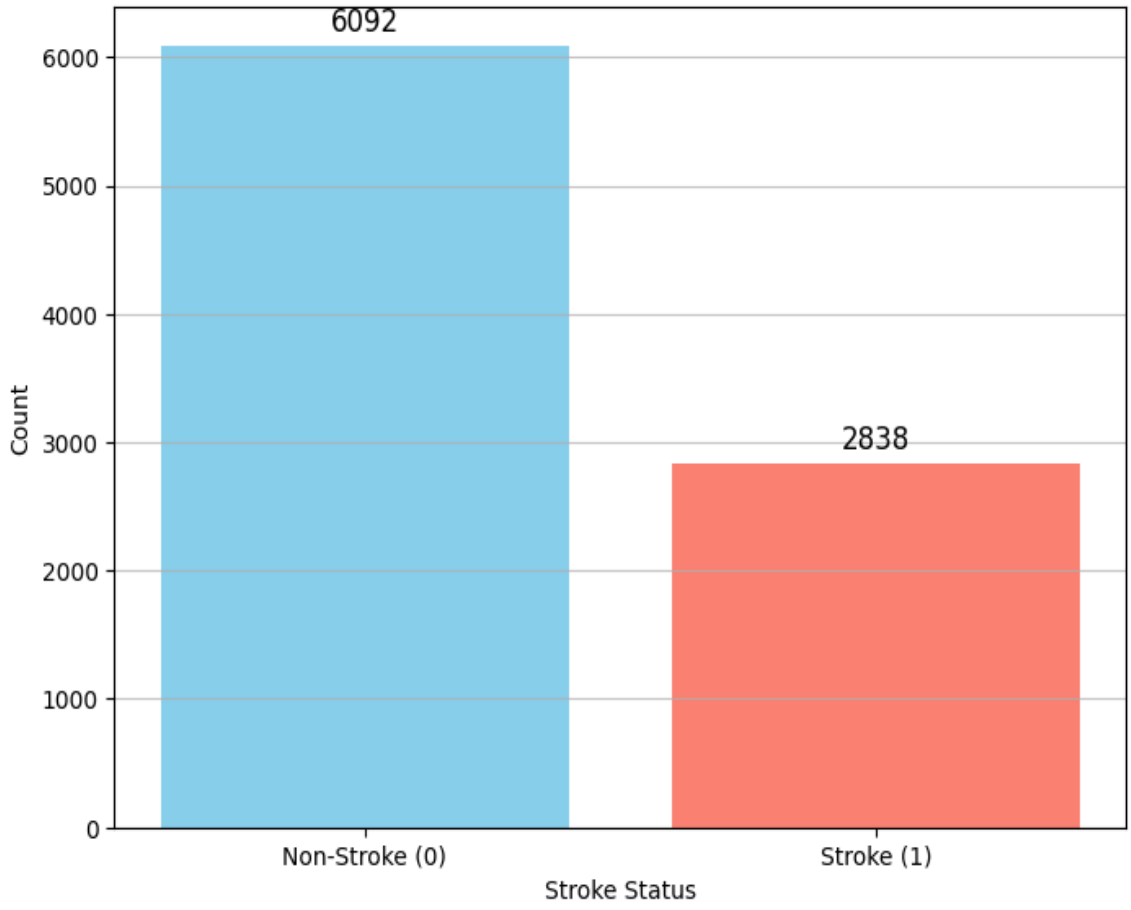



**Figure 4.2:** Dataset after Ignoring NULL Values

### 4.4 Data Preprocessing

Data preprocessing is an important stage that prepares the collected dataset for training the models.In this preprocessing various stages take place starting from importing the useful libraries followed by loading the collected data, data cleaning, normalization, and balancing which are used to increase the accuracy and efficiency of the model.

#### 4.4.1 Importing Libraries

In this stage, the researcher imported the relevant library for data processing tasks and the Python code for importing libraries is shown below.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from imblearn.over_sampling import SMOTE
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import QuantileTransformer, LabelEncoder
```

#### 4.4.2 Loading dataset

After importing the important libraries the data must loaded.

```
df = pd.read_csv('/content/patients record of disease.csv')
print("Shape of the data:", df.shape)
df.head()
```

After the above stage, as mentioned in Chapter 3 the input features ‘smoking’ and ‘alcoholic’ contain known data so, we can’t use these two features. By using the below code the researcher removes these attributes.

```
columns_to_ignore = ['Smoking', 'Alcholic']
df = df.drop(columns=columns_to_ignore)
```

And other missing values are handled.

#### 4.4.3 Data Normalization

Normalization is the process of bringing all of a dataset's columns into alignment with a common scale. It is not necessary to normalize every dataset for machine learning. Only

in cases where the ranges of features differ is it necessary[40]. Therefore, to prevent bigger numeric feature values from predominating over smaller numeric feature values, a data normalization technique is used to transform dataset characteristics in a common range. The adjusted features value distribution is transformed to a normal distribution using the quantile transform, and the negative impact of marginal values is lessened using min-max scaling. Min-Max Scaling is Divide the result by the range after deducting the lowest value from the highest value in each column. The minimum and maximum values for each new column are 0 and 1, respectively and by using quantile transform technique, the features are changed to have a uniform or normal distribution. It is the technique of data normalization used in our study. The result after applying transformation is shown in the below figure.

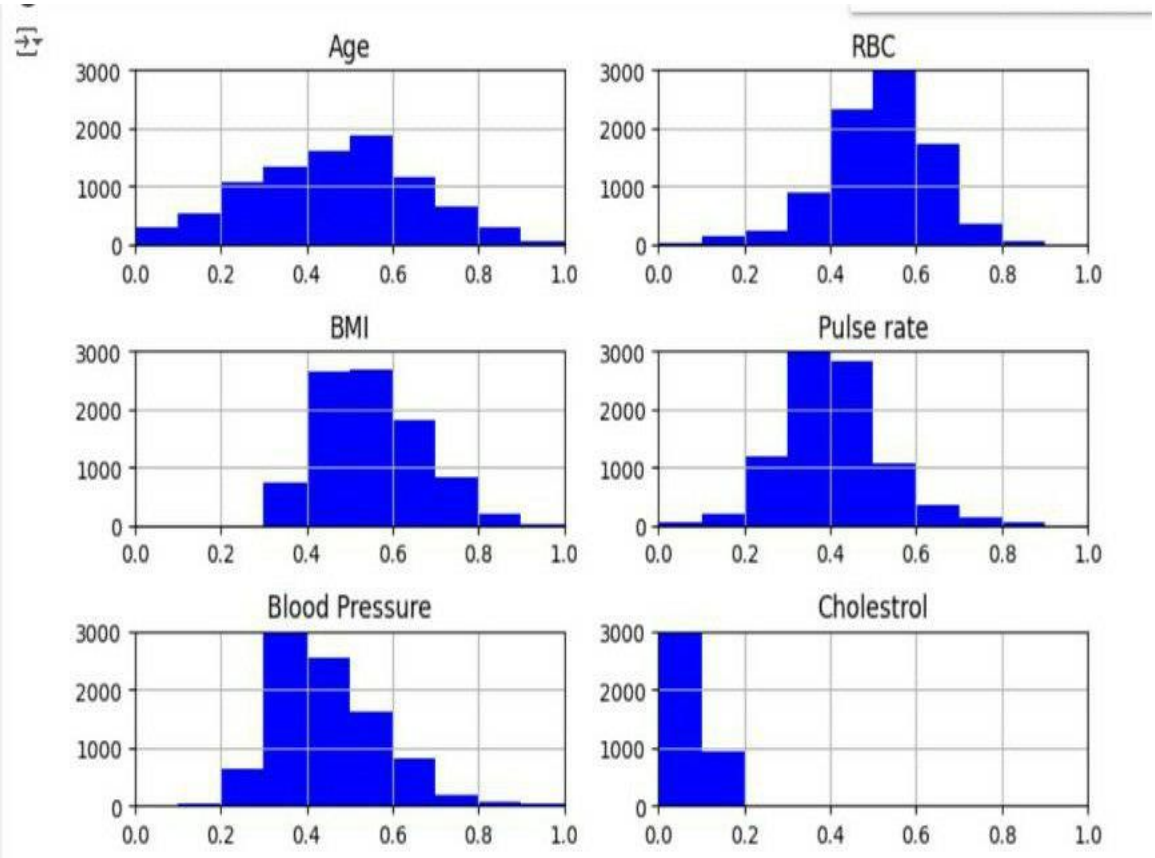



**Figure 4.3:** Normal data after applying Min-Max Scaling and quantile transform technique

### 4.4.4 Data balancing

Predictive modeling has difficulties when dealing with imbalanced datasets, however, these issues are expected as there are many imbalanced cases in real life[41]. The dataset in our study is imbalanced so, it needed to take measure to balance the dataset. In this study SMOTE over-sampling technique is used for balancing the dataset. SMOTE over-sampling technique is a way to achieve balance dataset by raising the proportion of instances of the minority class[42].The python code and the result after applying SMOTE over-sampling technique is shown below.

```
from imblearn.over_sampling import SMOTE
smote = SMOTE()
X_balanced, y_balanced = smote.fit_resample(X_selected, y)
y_balanced_df = pd.DataFrame(y_balanced, columns=['Stroke'])

# Plot the distribution of the target variable after SMOTE
plt.figure(figsize= (6, 4))
sns.countplot(data=y_balanced_df, x='Stroke')

# Add title and labels
plt.title('Class Distribution After SMOTE')
plt.xlabel('Class')
plt.ylabel('Count')

# Show the plot
plt.show()
```

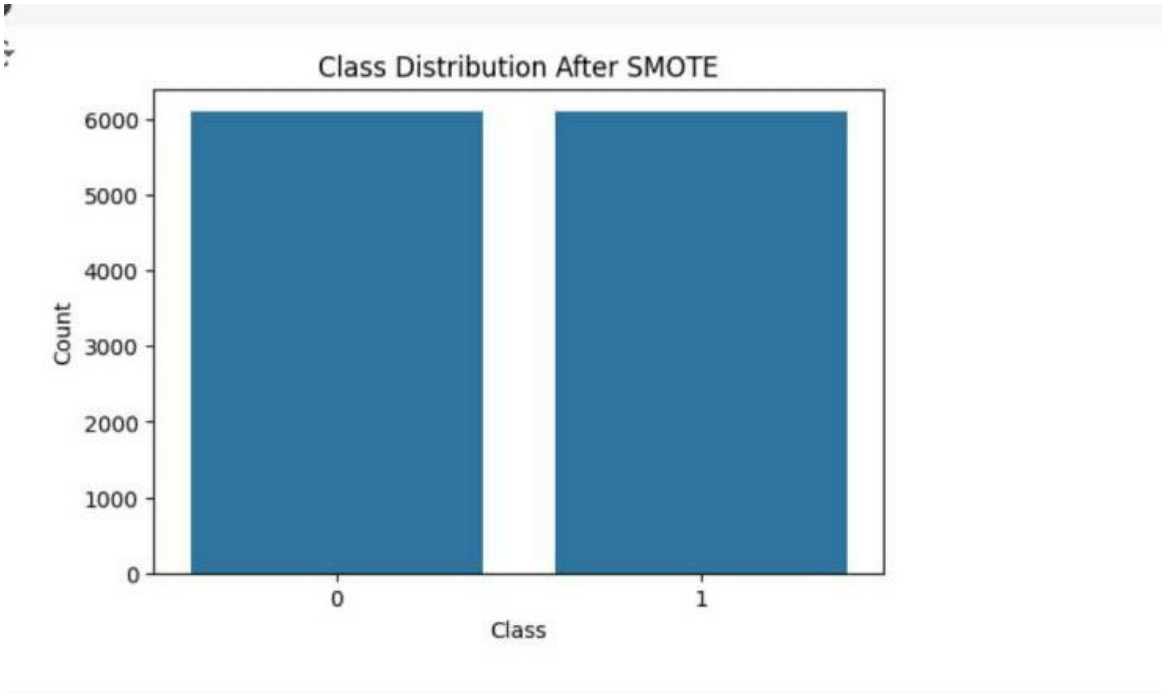



**Figure 4.4:** Class Distribution after SMOTE

### 4.5 Dataset Splitting

Methods for splitting data into training and testing subsets are called data splitting techniques. There are different techniques for splitting 60/40, 70/30 and 80/20. By comparing these techniques for this study 80/20 gives the better result for the model performance compared to the other techniques. This study uses 80% of the dataset for training purposes and 20% for testing. In this study the total number of dataset is 9203, after dropping the record with missing value it become 8930 as a record consists 273 datasets with missing value. This 8930 dataset split into training and testing dataset in 80/20 splitting techniques it means this study uses 7144 sample for training and 1786 sample for testing.

### 4.6 Ensemble Learning Model Selection

In this study, different types of ensemble learning algorithms are used but compared to others the best result was found by using Random forest, XGboost and LightGBM with stacking. Before evaluating these models, the study uses different preprocessing like Handling missing values, balancing the datasets, Normalization, scaling, and the likes. For Evaluation purpose, the study uses Accuracy, Recall, F1-Score, precision, confusion matrix, and correlation. For better performance of the models, the researcher used hyperparameter tuning.

### 4.7 Hyperparameter Tuning

The process of choosing the best collection of hyperparameters for a machine learning model is known as hyperparameter tuning, and it has a big influence on the model's functionality[43]. Optimizing the training process using a collection of hyperparameters that are just right can boost the model's performance. Well-tuned models using training data are more likely to perform well on test data that hasn't been seen before. This validates the accuracy of the model's predictions in practical settings. It can also shorten training periods and lower computational costs[43].In this study, we use the Grid search technique.

#### Grid Search

A conventional approach for hyperparameter adjustment in machine learning is called grid search. To identify the optimal model, it tries each combination of the supplied hyper-parameter values in detail. In essence, Grid Search trains a model for each possible combination of hyperparameters and then chooses the model that performs the best, so performing both model selection and hyperparameter tuning simultaneously[44].In our Study for hyperparameter tuning, we used Grid search.

**Table 4.1 Hyper-Parameter Tuning for all models**

| No. | Model | Parameters | Search space | Selected Parameter Values |
|---|---|---|---|---|
| 1 | Random Forest | n_estimator | 100,200,300 | 300 |
| | | min_samples_split | 2, 5, 10 | 2 |
| | | Max_depth | 10,20,30, None | 20 |
| | | min_samples_leaf | 1,2,4 | 1 |
| 2 | XGBOOST | learning_rate | 0.01,0.05,0.1,0.2 | 0.2 |
| | | max_depth | 3,6,9 | 9 |
| | | n_estimator | 100,200,300 | 200 |
| 3 | LightGBM | n_estimator | 100,200,300 | 200 |
| | | num_leaves | 31,62,100,150 | 150 |
| | | learning_rate | 0.01,0.05,0.1,0.2 | 0.2 |

### 4.8 Result and Analysis

In this section, we use Python with different ensemble learning tools which are analyzed in colab. For the prediction of stroke risk generally, 9203 is collected after applying all preprocessing techniques, the researcher split it into 80/20, 80% for training, and 20% for testing. The target group consists of two classes called non-stroke (0) and stroke (1) based on each prediction step starting from importing different libraries to colab. After doing all the preprocessing steps, the next step is analyzing each model's Random forest, XGboost, and LightGBM. For Evaluation, the study uses accuracy, precision, F1-score, Recall, Confusion matrix, and Correlation.

#### 4.8.1 Random Forest (RF) Result

This subsection started by importing the libraries which is helpful to use the Random forest model. To get optimal results by using the model in this study hyperparameter is

applied. From the different types of hyperparameters, Grid search is used in this research. To apply Grid search the important libraries are imported and the researcher uses different parameters by changing the values a lot of times to get a better result. The parameters used and the value used for each are shown in Table 4.1.

**Table 4.1. Hyperparameters used for Random forest**

| Parameters | n_estimators | max_depth | min_samples_split | min_samples_leaf |
|---|---|---|---|---|
| Values | [100, 200, 300] | [10,20,30, None] | [2, 5, 10] | [1, 2, 4] |

The parameters used above indicate different things. n_estimators indicates number of trees used for the Random forest, max_depth indicates the maximum depth of each tree in the random forest, min_samples_split indicates the minimum number of samples required to split an internal node, min_samples_leaf indicates the minimum number of samples required to be at a leaf node.

#### 4.8.1.1 Evaluation Matrix Performance for Random Forest

This subsection shows the results of different evaluation matrices for the prediction of stroke risk using Random forest. For evaluation, the study uses accuracy, precision, recall, f1-score, and confusion matrix, and classification report.

**Accuracy:** the performance of random forest accuracy is 97.6%.

**Precision:** the performance of random forest precision is 98.0%.

**F1-score:** the performance of random forest f1-score is 96.1%.

**Recall:** the performance of random forest recall is 94.4%.

**Table 4.2 Result and Evaluation Matrix for Random Forest**

| Evaluation Matrix | Value in percentage |
|---|---|
| Accuracy | 97.6% |
| Precision | 98.0% |
| F1-score | 96.1% |
| Recall | 94.4% |

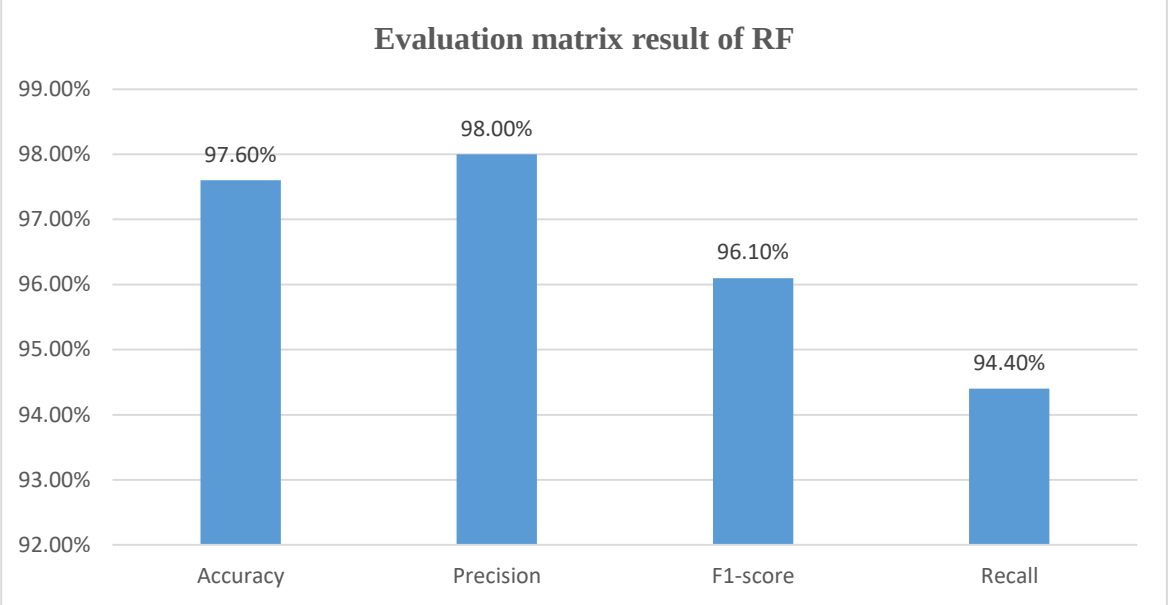



**Figure 4.5:** Evaluation metric result of RF

**Confusion matrix**

The confusion matrix provides a tabular estimate of the model's prediction accuracy. It includes the total number of accurate and inaccurate forecasts by class. The confusion matrix was created using the confusion_matrix () function, which takes as parameters the actual and expected outcomes. The following Python confusion matrix function can create a confusion matrix.

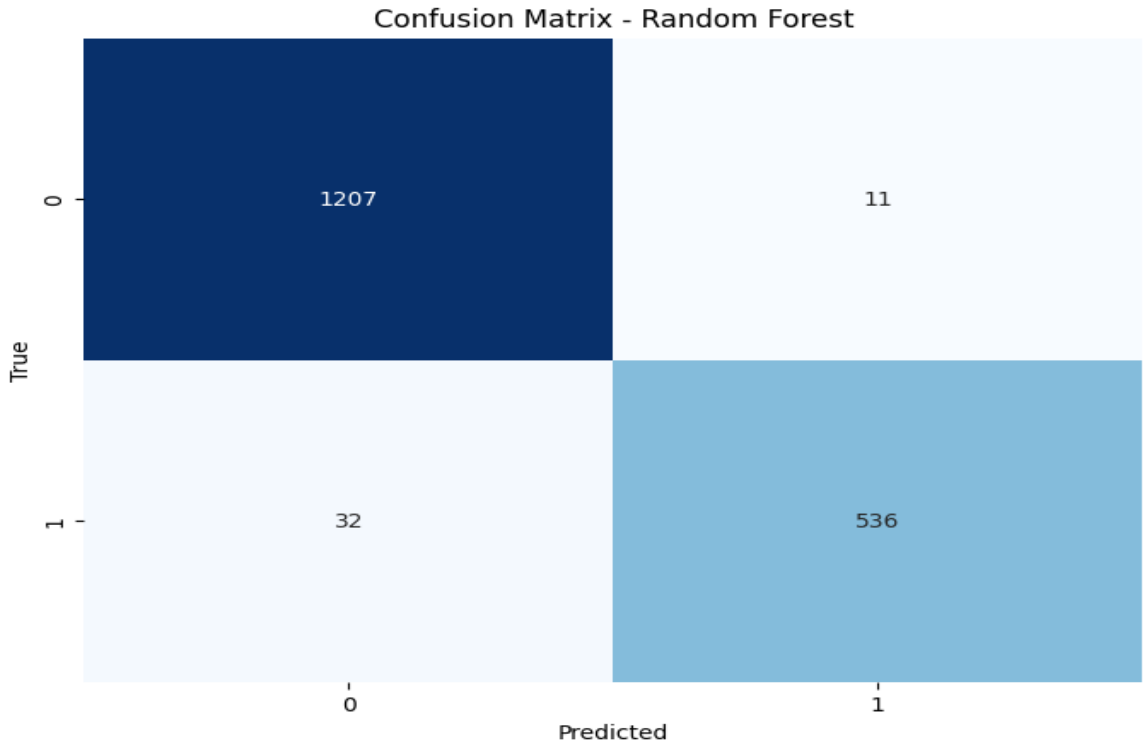



**Figure4.6:** Confusion matrix for classifiers using Random Forest

A confusion matrix contrasts the true and anticipated labels to assess how well a classification system performs.

**The matrix can be interpreted as follows:**

The positive class is represented by 1, which is the Stroke class, and the negative class by 0 is the non-stroke class. The columns correspond to the expected labels (what the model anticipated), while the rows relate to the true labels (actual).

**Interpretation of the Graph**

**True Negative - TN:** 1207 cases were classified as class 0 (negative) with accuracy.

**False Positive - FP:** 11 cases belong to class 0 (negative) but were mispredicted as class 1 (positive). The false positives (type I mistake) are these.

**False Negative - FN:** Although 32 cases are class 1 (positive), they were mistakenly forecasted to be class 0 (negative).

**True Positive, or TP:** 536 cases were classified as class 1 (positive) with accuracy.

**Classification report**

RandomForest classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.99 | 0.98 | 1218 |
| 1 | 0.98 | 0.94 | 0.96 | 568 |
| accuracy | | | 0.97 | 1786 |
| macro avg | 0.98 | 0.96 | 0.97 | 1786 |
| weighted avg | 0.97 | 0.97 | 0.97 | 1786 |

**Figure 4.7:** The Random Forest classification report

The Random Forest classification report shows a model's performance over two classes (labeled 0 and 1) in terms of multiple metrics. What each section of the report means is as follows:

**Metrics for each class:**

**Class 0(Non-stroke) -Support = 1218**

**Precision (0.97):** 97% of the cases accurately identified as class 0, out of all the predictions.

**Recall (0.99):** 99% of the real class 0 instances were identified properly.

**F1-Score (0.98):** shows a balance between the two.

**Class 1(Stroke) -Support = 568**

**Precision (0.98):** 98% of the cases accurately identified as class 1 out of all the predictions.

**Recall (0.94):** Ninety-four percent of the real class 1 cases were correctly identified.

**F1-Score (0.96):** strikes a compromise between recall and precision.

**Total Measures:**

Accuracy (0.97): 97% of all occurrences (combined from both classes) were properly classified by the model.

**Standard Measures:**

Macro avg (Precision = 0.98, Recall = 0.96, F1-Score = 0.97): This represents the F1-score, precision, and recall as an unweighted average for both classes, treated equally.

Weighted average (F1-score = 0.97, Precision = 0.97, and Recall = 0.97): This average gives the larger class more weight by accounting for the support (number of instances) for each class.

**Interpretation**

For both classes, the model exhibits strong performance, with good recall and precision.

The model's overall robustness is demonstrated by its 97.6% accuracy rate.

### 4.8.2 Extreme Gradient Boosting (XGBoost) Result

In this subsection, the study discusses how XGboost is used for the prediction of stroke risk. XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning framework. It is the top machine learning package for regression, classification, and ranking issues and offers parallel tree boosting[45].To work with this model importing the libraries which is helpful to use the XGBoost model will be the first step. To get optimal results by using the model in this study hyperparameter is applied. From the different types of hyperparameters, Grid search is used in this research. To apply Grid search the important libraries are imported and the researcher uses different parameters by changing the values many times to get a better result. The parameters used and the value used for each are shown in Table 4.3.

**Table 4.3 Hyperparameters used for XGBoost**

| Parameters | max_depth | learning_rate | n_estimators |
|---|---|---|---|
| Values | [3, 6, 9] | [0.01, 0.05, 0.1,0.2] | [100, 200, 300] |

For the XGBoost algorithm, the researcher uses three parameters called max_depth which means the maximum depth of each tree in the XGBoost, learning_rate decides how much each boosting iteration of the process updates the model weights, n_estimator speaking of the quantity of boosting rounds (or trees) that must be constructed throughout the training phase.

#### 4.8.2.1 Evaluation Matrix Performance for XGBoost

This subsection discusses the evaluation matrix used for the XGBoost model and the result of a model after evaluation. The evaluation matrix used here is Accuracy, precision, F1-score, Recall, Confusion matrix, and Classification report. The result of each matrix is shown below.

**Accuracy:** the performance of XGBoost accuracy is 96.1%.
**Precision:** the performance of XGBoost precision is 95.4%.
**F1-score:** the performance of XGBoost f1-score is 93.8%.
**Recall:** the performance of XGBoost recall is 92.3%.

Table 4.4 Result and Evaluation Matrix for XGBoost

| Evaluation Matrix | Value in percentage (%) |
|---|---|
| Accuracy | 96.1% |
| Precision | 95.4% |
| F1-score | 93.8% |
| Recall | 92.3% |

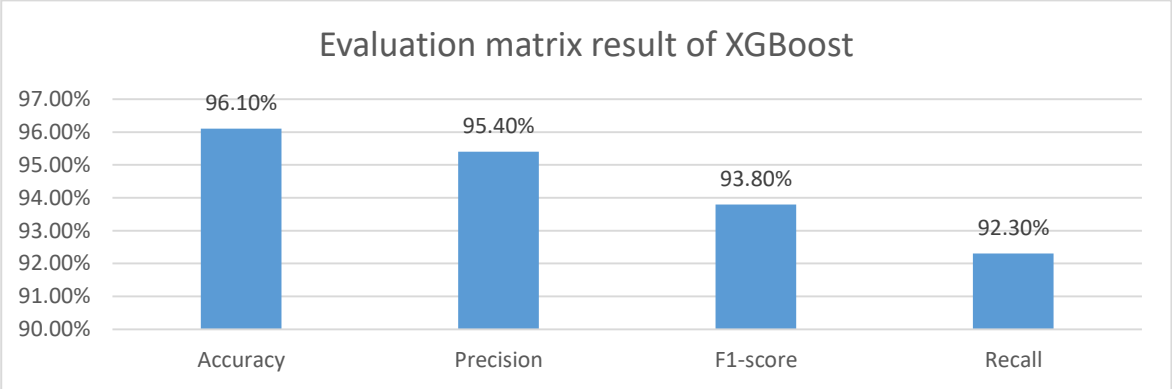



**Figure 4.8:** Evaluation matric result of XGBoost

**Confusion matrix**

Random sampling was used to separate the 1786 total test data instances into 1218 non-stroke (0) and 568 stroke (1) cases. As can be seen from the confusion matrix result above, 1193 instances of the first-row non-stroke class are accurately predicted to belong to the non-stroke class. 25 instances of non-stroke (0) classes, however, are mispredicted as stroke (1) classes. In the second row, 44 occurrences are erroneously assigned to the non-stroke (0) class, while 524 instances of stroke are accurately predicted to the stroke class (1). The confusion matrix for XGBoost is shown below.

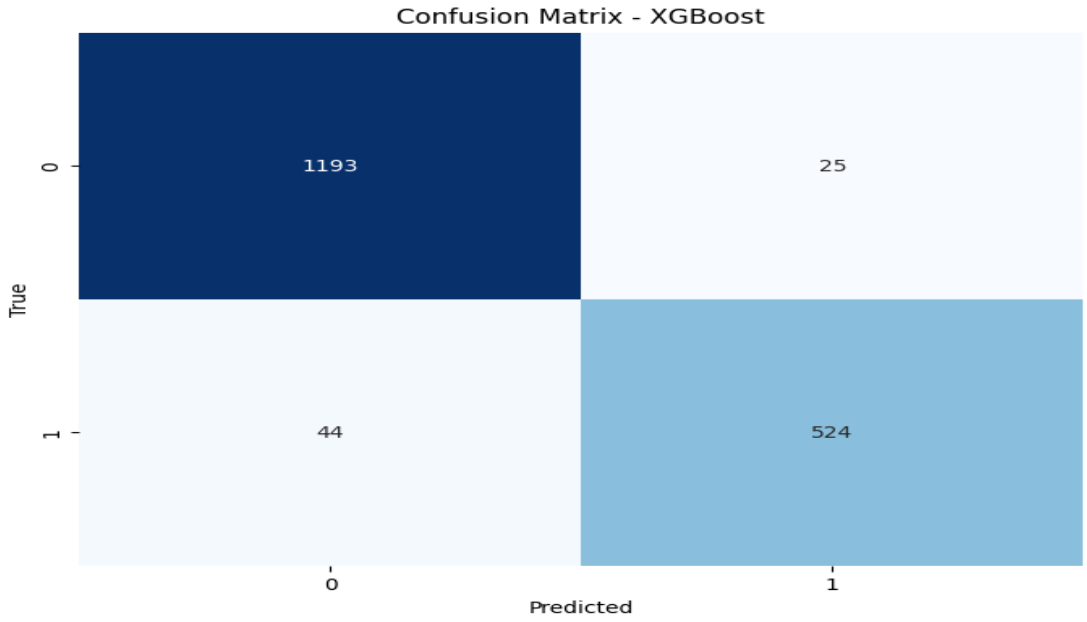



**Figure 4.9:** Confusion matrix for classifiers using

**Classification report**

XGB classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.98 | 0.97 | 1218 |
| 1 | 0.95 | 0.92 | 0.94 | 568 |
| accuracy | | | 0.96 | 1786 |
| macro avg | 0.96 | 0.95 | 0.96 | 1786 |
| weighted avg | 0.96 | 0.96 | 0.96 | 1786 |

**Figure 4.10:** The XGBoost classification report

The precision, recall, and f1-score are included in the following result. 95% precision, 92% recall, and 94% recall were attained in this experiment for the stroke group of the Target group, and 96% precision, 98% recall, and 97% recall for the Non-stroke group. The model's 97.6% accuracy rate shows how solid it is overall.

### 4.8.3 Light Gradient Boosting Machine (LightGBM)

In LightGBM, evaluation evaluates the performance of the trained model using metrics like accuracy for classification tasks or mean squared error for regression tasks. To assess model performance on unseen data and avoid overfitting, cross-validation techniques might be used[33].In this study, besides using cross-validation the hyperparameter grid search is used by using different parameters and different values by trying a lot of times. Grid search is one of the types of hyperparameters that is helpful to increase the performance of the model, to increase the optimal result of the models grid-search is used in this study also. To apply Grid search the important libraries are imported.

**Table 4.5. Hyperparameters used for LightGBM**

| parameters | n_estimators | num_leaves | Learning_rate |
|---|---|---|---|
| values | [100,200,300] | [31,62,100,150] | [0.01,0.05,0.1,0.2] |

For LightGBM three parameters are used these are num_leaves directly regulates the ensemble's decision trees' complexity, Max_depth which means the maximum depth of each tree in the LightGBM, Learning_rate decides how much each boosting iteration of the process updates the model weights.

#### 4.8.3.1 Evaluation Matrix Performance of LightGBM

For this model also different evaluation matrix evaluated and the researcher got different results after the evaluation. The evaluation matrix used is accuracy, precision, recall, f1-score, confusion matrix, and classification matrix. The results of each review are listed below.

**Accuracy:** the performance of LightGBM accuracy is 92.9%.
**Precision:** the performance of LightGBM precision is 94.9%.

**F1-score:** the performance of LightGBM f1-score is 88.1%.

**Recall:** the performance of LightGBM recall is 82.2%.

**Table 4.6 Result and Evaluation Matrix for LightGBM**

| Evaluation Matrix | Value in percentage (%) |
|---|---|
| Accuracy | 92.9% |
| Precision | 94.9% |
| F1-score | 88.1% |
| Recall | 82.2% |

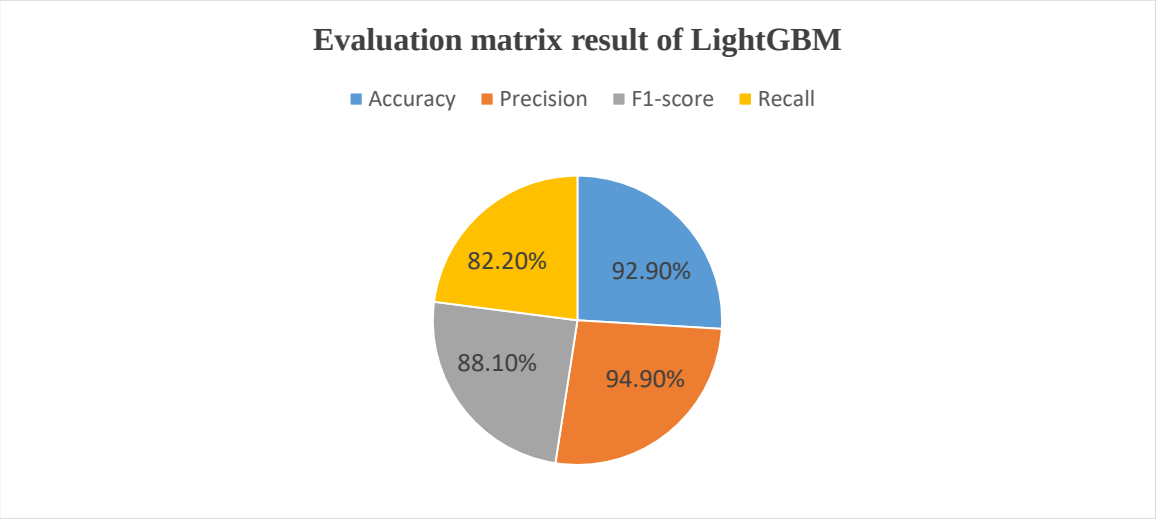



**Figure 4.11:** Evaluation matrix result of LightGBM

**Confusion matrix**

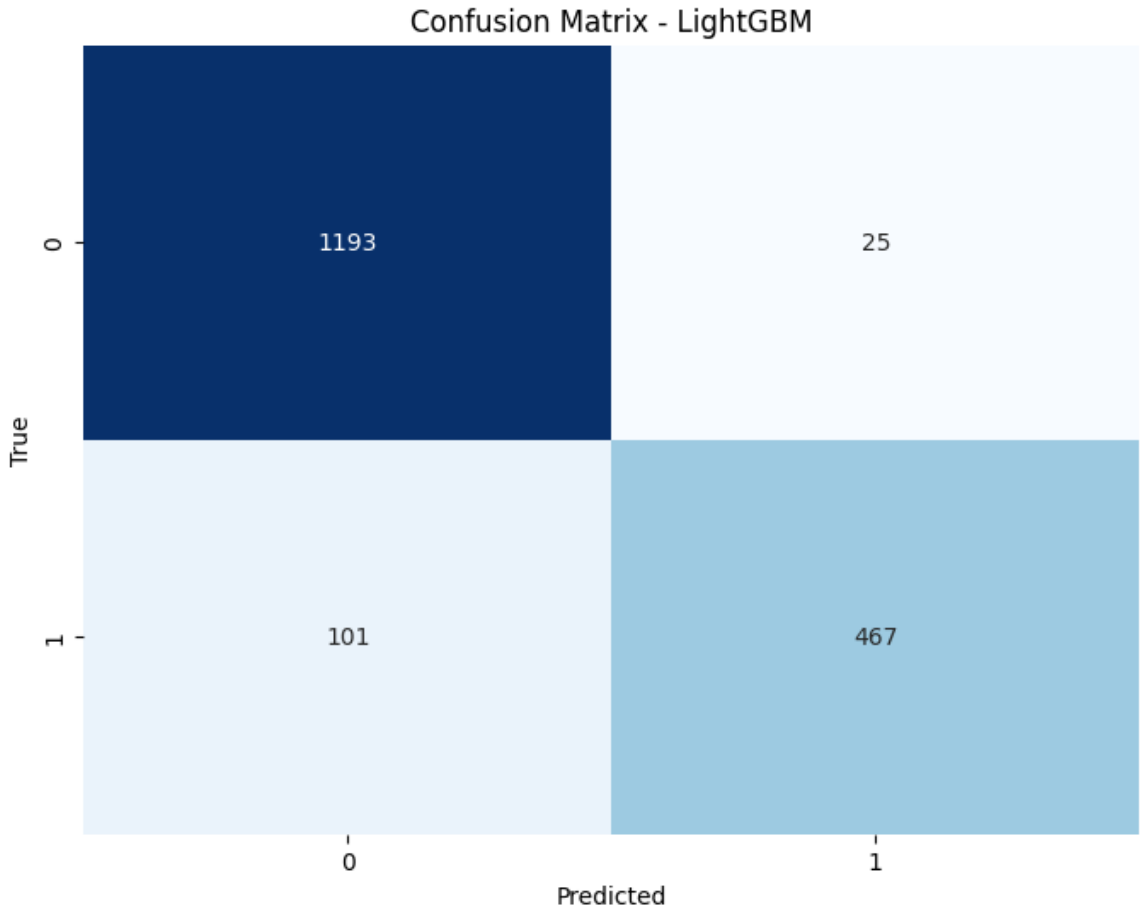



**Figure4.12:** Confusion matrix for classifiers using LightGBM

A total of 1786 test data instances were randomly separated into 1218 non-stroke (0) and 568 stroke (1) cases. As can be seen from the confusion matrix result above, 1193 instances of the first-row non-stroke(0) class are accurately predicted to belong to the non-stroke(0) class. Only 25 instances of non-stroke (0) classes, however, are mispredicted as stroke (1) classes. In the second row, 101 occurrences are wrongly assigned to the non-stroke (0) class, whereas 467 cases of stroke (1) are correctly projected to the stroke class.

**Classification report**

LGMB classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.92 | 0.98 | 0.95 | 1218 |
| 1 | 0.95 | 0.82 | 0.88 | 568 |
| accuracy | | | 0.93 | 1786 |
| macro avg | 0.94 | 0.90 | 0.92 | 1786 |
| weighted avg | 0.93 | 0.93 | 0.93 | 1786 |

**Figure4.13:** The LGMB classification report

According to the classification report mentioned above, precision describes how accurate our model is or how well it produces predictions. 95% test set of patients are expected to be at risk for stroke, according to the LGMB model. Therefore, when this prediction model indicates that a sample has a non-stroke, it is 92% correct. Metrics for recall evaluation assess how well the model can identify positive samples. 82% of the stroke patients and 98% of non-stroke in the test set are identified by the LGMB model. The model's overall performance is measured by the F1-score. 88 percent of stroke cases and 95% of non-stroke were successfully predicted by our LGMB prediction model.

### 4.8.3 Stacking Result

Using Random Forest, XGBoost, and LightGBM as base models, stacking is an ensemble learning strategy that enhances a predictive model's overall performance by combining the predictions of several base models. By merging their outputs through a meta-model, it maximizes each model's strengths while minimizing its own shortcomings.

#### 4.8.3.1 Evaluation Matrix Performance of Stacking

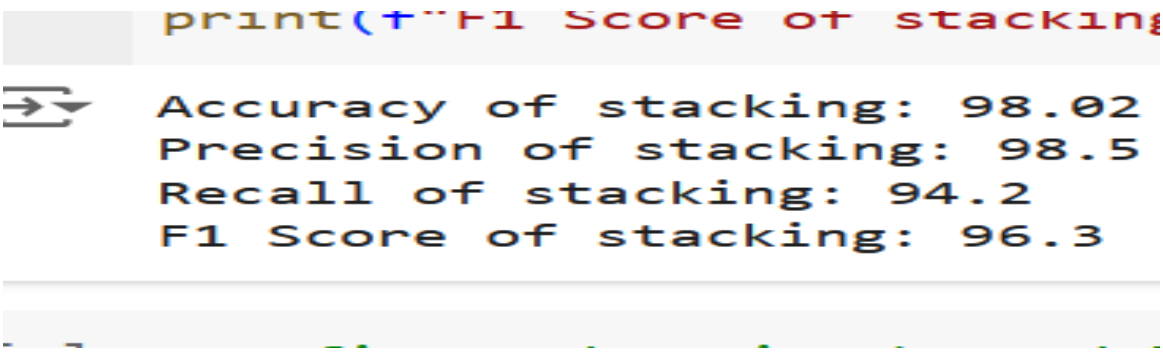
Additionally, multiple assessment matrices were evaluated for this model, and the evaluation yielded diverse outcomes for the researcher. Accuracy, precision, recall, f1-score, confusion matrix, and classification matrix are the evaluation matrices utilized. Below is a list of each review's findings.

**Accuracy:** the performance of Stacking accuracy is 98.02%.

**Precision:** the performance of Stacking precision is 98.50%.

**F1-score:** the performance of Stacking f1-score is 94.20%.

**Recall:** the performance of Stacking recall is 96.30%.




**Correlation Between Features**

This correlation heatmap shows how the various variables in our dataset relate to one another. The values of a correlation matrix span from -1 to 1: Positive correlation (nearer 1): This indicates that there is a tendency for both variables to rise in tandem with an increase in one. A negative correlation (nearer to -1) means that the two variables tend to decline as one rises. There is no linear relationship between the variables when there is zero correlation, or very close to 0.

An explanation of the main relationships is provided below:

Blood Pressure and Hypertension (0.52): Given that high blood pressure is a key feature of hypertension, this moderately positive association indicates that the chance of developing hypertension rises as blood pressure does.

Hypertension and Discontinuation Anti-HTN Drug (0.35): This positive association indicates that stopping antihypertensive medications, or anti-HTN medications, is linked to increased rates of hypertension. This is understandable as quitting medication can result in poorly controlled blood pressure.

Heart Disease and Pulse Rate (0.24): This weakly positive association suggests that people with greater pulse rates may be somewhat more likely to suffer from heart disease.

**Additional noteworthy correlations:**

Blood Pressure and Age (0.17): Although there is little evidence of a relationship, blood pressure tends to rise with age.

Age and Anti-HTN Drug Discontinuation (0.15): There is a weak correlation between stopping antihypertensive medications and getting older, which may suggest that medication adherence varies with age.

**Lower or Non-Significant Correlations**

Certain pairs, such as diabetes and pulse rate or cholesterol and sex, exhibit very low or almost zero correlations, suggesting little to no linear association.

In conclusion, the heatmap illustrates the interdependence of age and heart disease risk factors, blood pressure, hypertension, and stopping antihypertensive medications. This can assist in determining the key regions to keep an eye on in order to prevent strokes.

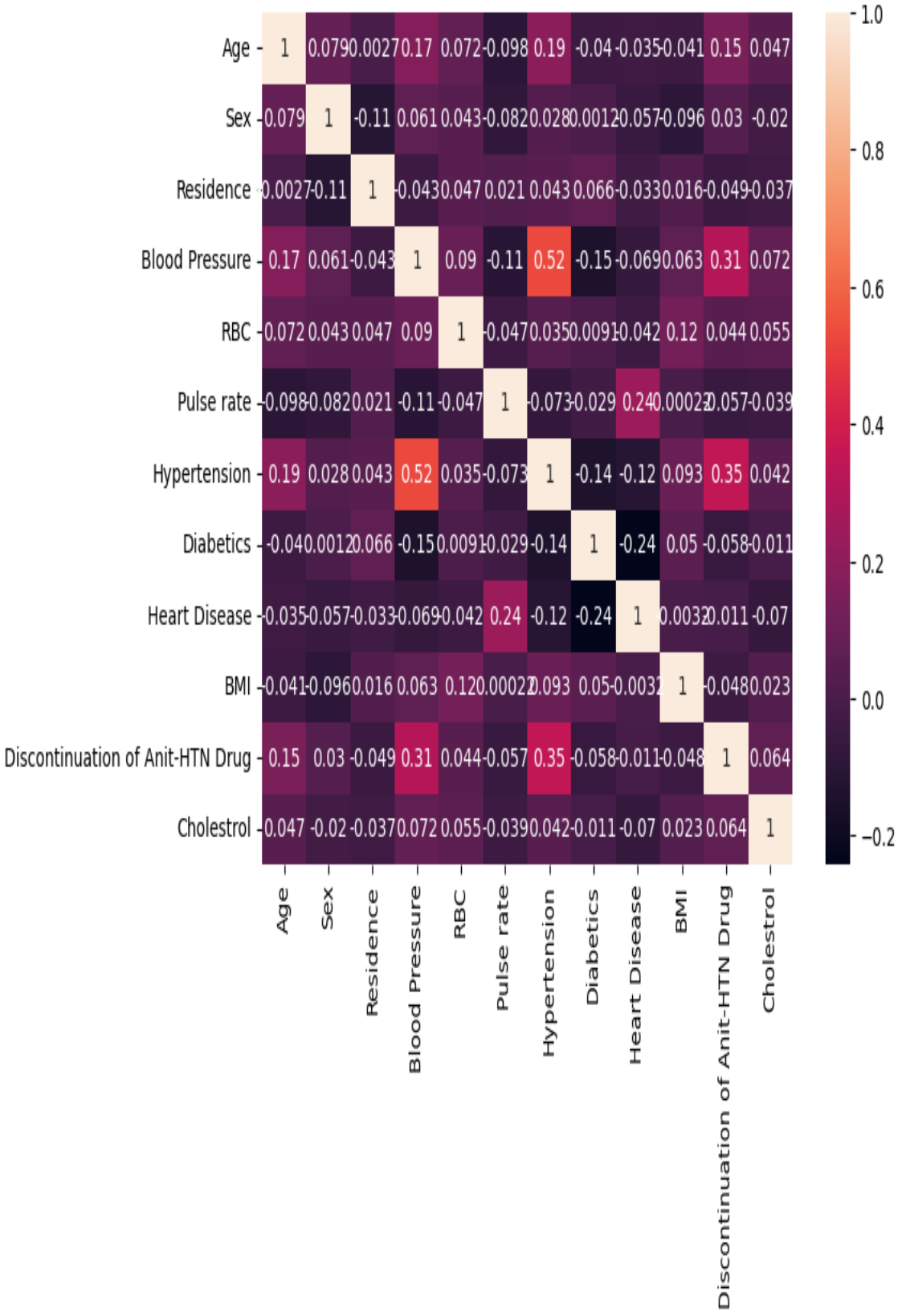



**Figure 4.14:** Correlation between features

### 4.9 Compare and Discussion Results

The main objective of this study is to find ensemble learning model which performs well prediction of stroke risk and finding the major risk factors which are more helpful for the prediction. Before going to the evaluation steps different preprocessing techniques takes place and the datasets splitted in two groups in 80/20 for training and testing purpose. Random Forest, XGBoost and LightGBM are the ensemble learning models used in the study.

The three ensemble learning algorithms were compared to one another in terms of performance. The evaluation measures' outcomes, including accuracy, precision, recall, and F1-score, form the basis of the findings. As can be seen here, the Random Forest prediction models outperformed the XGBoost and LightGBM models with stacking in terms of accuracy, precision, recall, and F1 score.

The accuracy of three distinct ensemble learning models—Random Forest (RF), XGBoost, and LightGBM with stacking—is graphically represented by the bar graph below in Figure 4.15.

The Stacking model exhibits the best accuracy compared to the other three alone, meaning that it makes more accurate predictions. While XGBoost performs admirably, it does not achieve the same level of accuracy as Random Forest. Out of the three, LightGBM has the lowest accuracy, indicating that it misclassifies more often than XGBoost and RF combined.

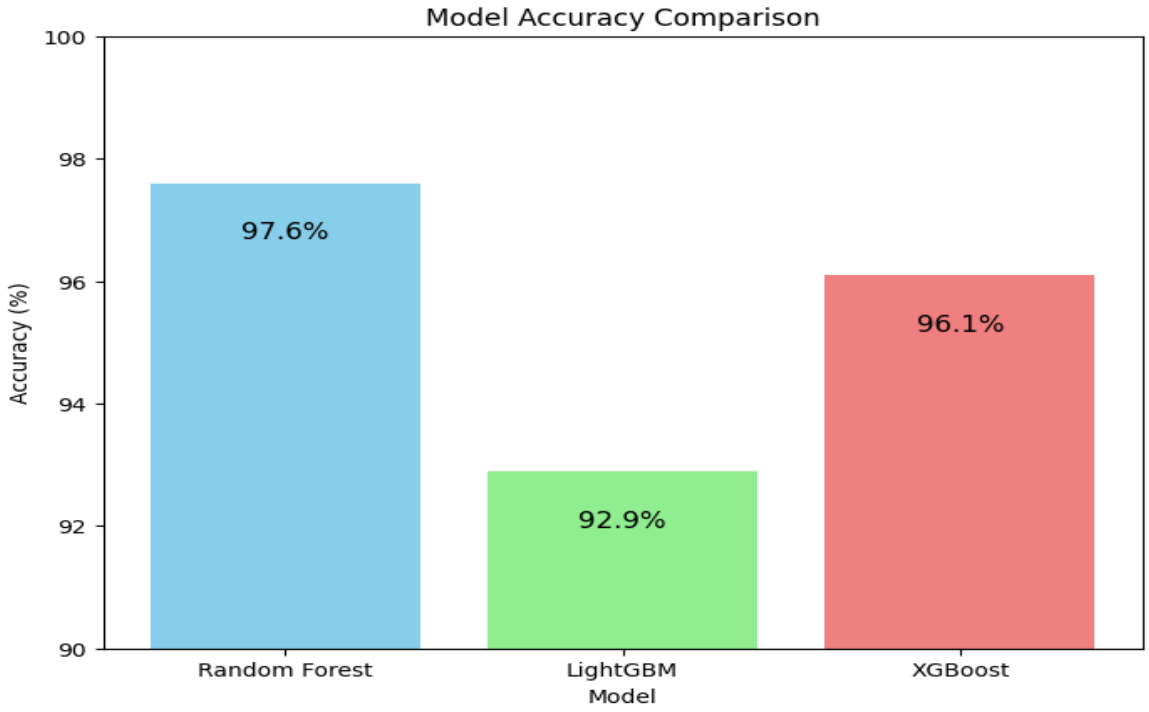



**Figure 4.15:** Comparison of Accuracy for the models

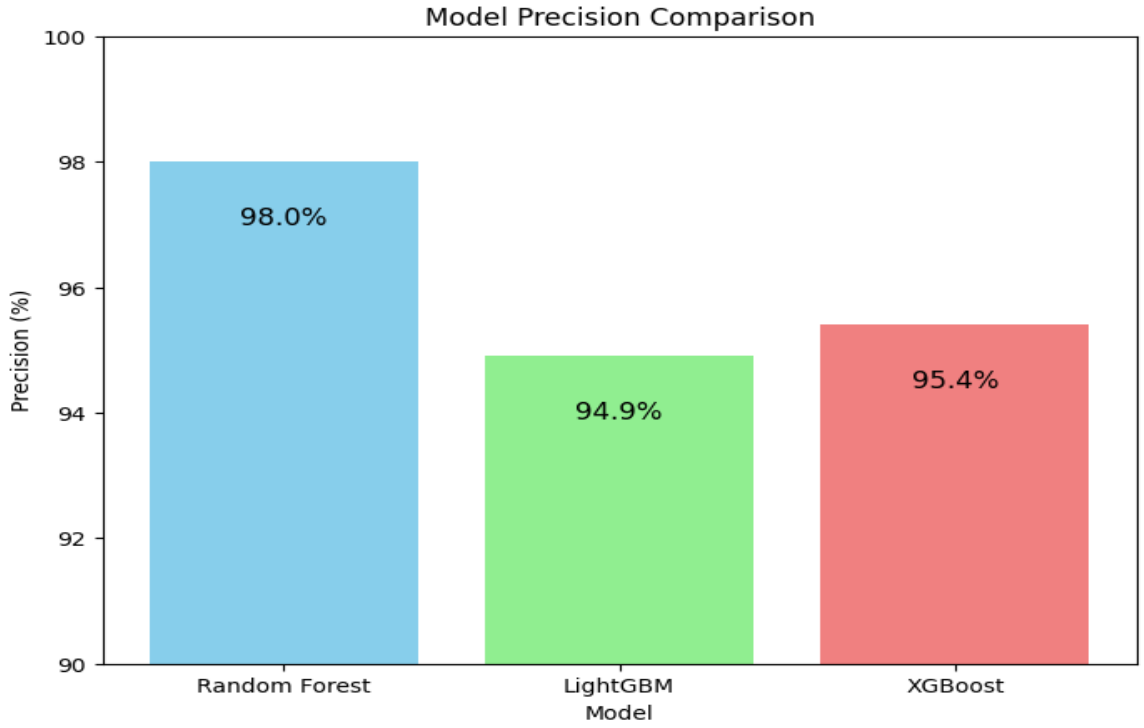



**Figure 4.16:** Comparison of Precision for the models

The precision of three ensemble learning models—Random Forest, LightGBM, and XGBoost—is compared in the above figure. Among the three models, Random Forest has the best precision, meaning it can identify positive cases more accurately than the other two. XGBoost and LightGBM display lower precision figures. This highlights how useful Random Forest is in this situation. The graph, take \as a whole, shows how different the models' degrees of precision are.

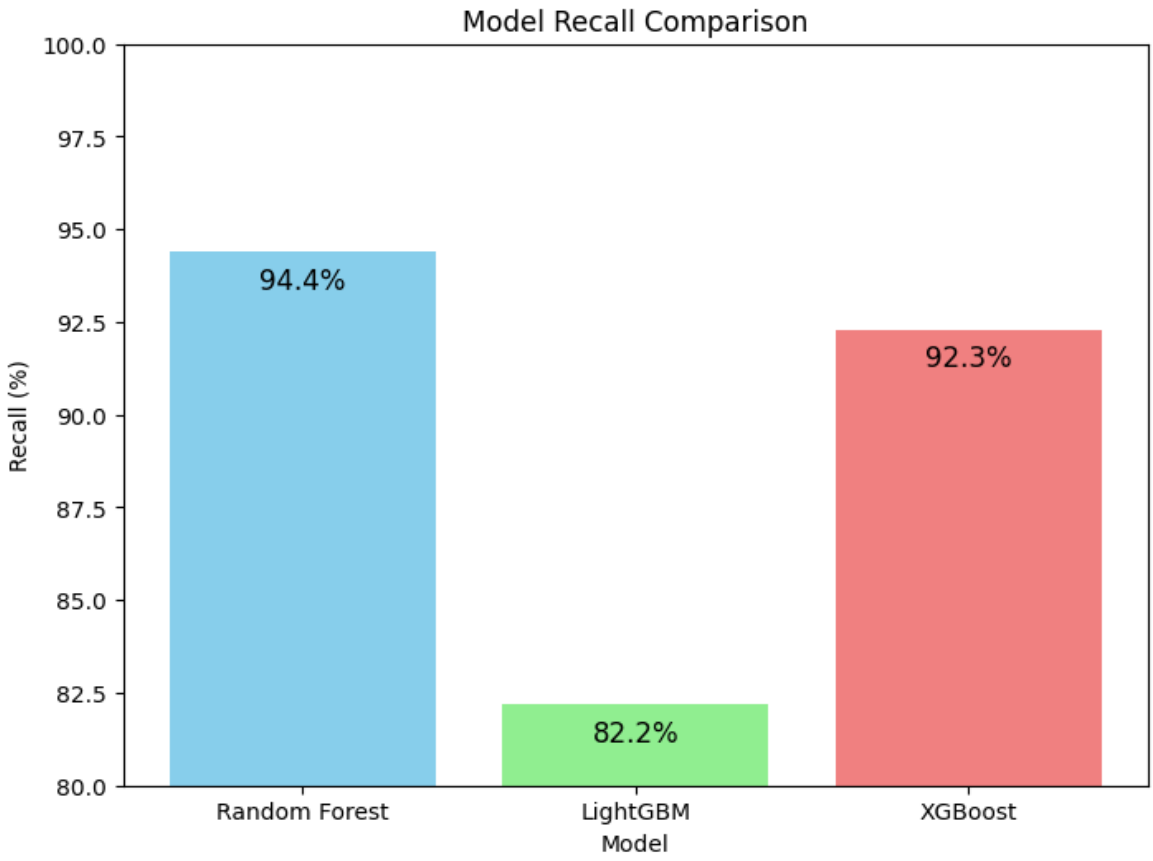



**Figure 4.17:** Comparison of Recall for the models

The recall comparison of the three models—Random Forest, LightGBM, and XGBoost—is displayed in Figure 4.12. Random Forest scores better than the other two models in terms of recall, demonstrating its increased capacity to recognize positive instances. In comparison to LightGBM and XGBoost, Random Forest appears to be more successful at catching true positives, based on its better recall. Conversely, the other models' lower

recall values suggest that they might be missing more positive instances. In terms of recall performance, this statistic demonstrates Random Forest's advantages overall. For this reason, Random Forest stands out as the most trustworthy choice out of the three models in this study.

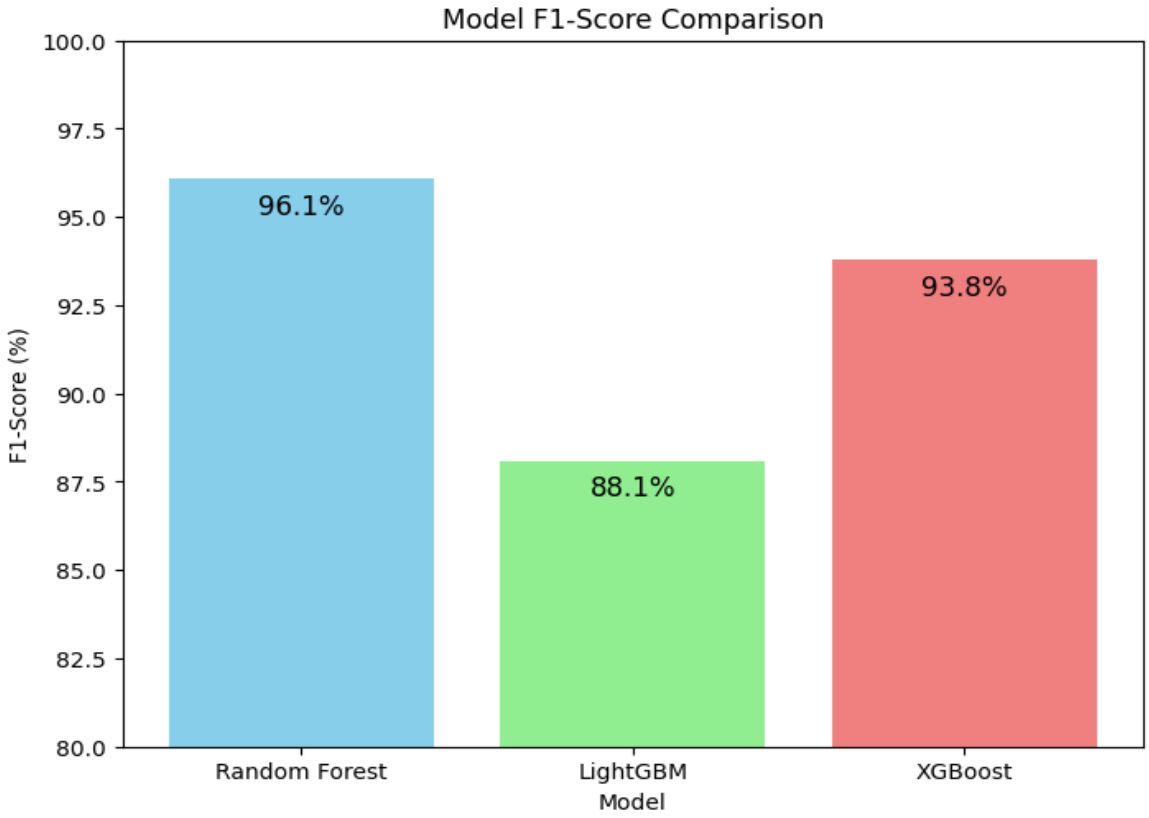



**Figure 4.18:** Comparison of F1-Score for the models

In the above comparison, the F1-Score performance of Random Forest is higher than that of the other two models. Therefore, we can conclude that Random Forest is the best ensemble learning model for our study.

### 4.10 Importance of Features Score

Feature importance rates input features according to how well they can predict a target variable[46].The findings of the feature importance analysis shed light on the variables that most strongly affect the model's predictions. Higher numbers denote greater

importance. The values represent each feature's contribution to the total predictive performance. The below Figure shows the importance feature scores of our study.

| | Feature | Importance |
|---|---|---|
| 11 | Discontinuation of Anit-HTN Drug | 0.337881 |
| 8 | Hypertension | 0.083107 |
| 4 | Blood Pressure | 0.072158 |
| 0 | Age | 0.064274 |
| 9 | Diabetics | 0.061460 |
| 6 | Sex | 0.059472 |
| 7 | Residence | 0.056758 |
| 5 | Cholestrol | 0.055875 |
| 3 | Pulse rate | 0.054602 |
| 2 | BMI | 0.053840 |
| 1 | RBC | 0.052646 |
| 10 | Heart Disease | 0.047927 |

**Figure 4.19:** Important Feature Scoring

The feature importance figure above illustrates that the most crucial aspect impacting the model's predictions is whether a patient has discontinuation of anti-HTN drug, demonstrating this has a strong impact on outcomes. The presence of hypertension and Blood pressure levels are also significant variables for the prediction of stroke risk. Less so do other characteristics including sex, age, and diabetes status, however they still have an impact. Even less of an impact is made by variables including residency, cholesterol, pulse rate, BMI, and RBC count. It's interesting to note that, despite its typical relevance in health-related models, heart disease has the least impact in this model, meaning it contributes the least to forecasts. In general, the model emphasizes factors discontinuation of anti-HTN drugs, Hypertension, and blood pressure are the major risk factors of stroke in the study area.

# CHAPTER FIVE

# CONCLUSIONS AND FUTURE WORKS

## 5.1. Conclusions

This study's primary goal is to use ensemble learning to develop a prediction model for stroke risk by identifying the critical variables that influence stroke and finding the major risk factors of stroke in the study area. Predicting the risk of stroke is the aim. To properly conduct this study, it was necessary to first understand how strokes impact individuals before gathering data according to risk variables. Yirgalem General Hospital, Yanet Internal Medicine Specialty Center, and Hawassa Referral Hospital were the three hospitals from which the data was gathered. After gathering the original data, from the collected data the manual datasets were converted to electronic format and cleaned up in preparation for analysis. Age, sex, residence, RBC, pulse rate, hypertension, diabetes, discontinuation of the anti-HTN drug, stroke (target class), etc. are some of its 12 inputs. From these Hospitals total of 9203 datasets were collected.

Using Ensemble learning techniques like Python and several libraries, data preprocessing was done in this work to clean up the data and make it appropriate for ensemble learning models. In this study, records containing missing data were removed, normalization and QuantileTransformer, and oversampling using SMOTE were also used for balancing. Grid search technique is also used to find better performance of the models. Additionally, making use of the label encoding method. For non_stroke and stroke, respectively, 0, 1 is used in place of Stroke in the target class dataset.

The models employed and trained are Random Forest, LightGBM, and XGBoost. 80% of the data was utilized for training and 20% was used for testing in the evaluation model, which was run using an 80/20 data splitting. A variety of relevant Evaluation methods,

including precision, recall, accuracy, F1-score, and confusion matrix, were employed to assess the performance of those models. In addition to building the model with relevant data and selecting the best-fitting model, the study examined the optimal ensemble learning strategies for predicting the risk of stroke.

The experiment's accuracy on hold-out (test) data reveals that the Random Forest classifier (97.6%), LightGBM (92.9%), and XGBoost (96.1) performed best.when ensemble them together with stacking it performes better with accuracy 98.02%. Because of its stability across training sets, ability to handle imbalanced data effectively, and resilience to overfitting, Random Forest outperforms XGBoost and LightGBM in stroke risk prediction. This results in increased accuracy and superior overall performance. and is deemed suitable for predicting the risk of stroke.

The stroke risk factors with the greatest significant impact on the study can be determined from the input features provided by using the feature importance scale. The results of this study indicate that Discontinuation of anti-HTN drugs, Hypertension, and blood pressure are the three main risk factors for stroke, followed by age and diabetes.

### 5.2. Future Works

Expanding the geographic scope of studies to encompass a larger range of places is crucial to improving future research on stroke risk factors and improving the generalizability of findings. To obtain a more thorough understanding of stroke risks, researchers should also expand the scope of their examination of risk variables by adding a wide range of demographic and lifestyle characteristics, such as smoking, genetics, and substance use. It is also advised that hospitals switch to electronic health records in order to improve data accuracy and organization, which will enable more thorough analysis. Predictive models will be more reliable if the sample size is increased by incorporating

data from more healthcare facilities, since this will guarantee a larger and more diversified dataset.

Combining Predictive Models into CDSS (Clinical Decision Support Systems). This may make it possible for medical professionals to make data-driven choices, recognizing high-risk patients early and putting preventative treatments or lifestyle modifications into place before a stroke happens. To assist people in tracking their risk of stroke over time, creating a way to utilize these ensemble learning models either in mobile applications or health websites.

Future researchers should integrate longitudinal data and real-time monitoring has the potential to improve model accuracy by accounting for temporal variations in risk factors. Finally, Future researchers should add more risk factors that will help in for prediction of Stroke risk.

### 5.3. Contributions

By creating an appropriate Ensemble learning model that can predict stroke in the case of Sidama regional State, the suggested study added to the knowledge and practice already in place. Furthermore, the study provided a risk factor Discontinuation of anti-HTN drugs that was absent from the other studies' datasets and guided future researchers in predicting the risk of stroke in other parts of the Country.

Additionally, by gathering and organizing the manually formatted secondary data on stroke risk and translating it into a machine-readable format, the researcher helped future researchers. Because gathering every proven case of stroke risk from the patient history card and preparing it for model building requires a lot of work. Lastly, the suggested study produced predictive stroke data that can aid medical professionals in diagnosing patients more accurately.

## References

[1] WorldHealthOrganization,“World Stroke Day 2022.” Accessed: Oct. 28, 2023. [Online]. Available: https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022

[2] JOHNHOPKINSMEDICINE,“Stroke.” Accessed: Apr. 01, 2024. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke

[3]E. Shenkutie Greffie, “Risk Factors, Clinical Pattern and Outcome of Stroke in a Referral Hospital, Northwest Ethiopia,” *CMR*, vol. 4, no. 6, p. 182, 2015, doi: 10.11648/j.cmr.20150406.13.

[4]“WHO EMRO | Stroke, Cerebrovascular accident | Health topics,” World Health Organization - Regional Office for the Eastern Mediterranean. Accessed: Feb. 29, 2024. [Online]. Available: http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

[5]R. O. Akinyemi *et al.*, “Stroke in Africa: profile, progress, prospects and priorities,” *Nat Rev Neurol*, vol. 17, no. 10, pp. 634–656, Oct. 2021, doi: 10.1038/s41582-021-00542-4.

[6]S. A. Gebremariam and H. S. Yang, “Types, risk profiles, and outcomes of stroke patients in a tertiary teaching hospital in northern Ethiopia,” *eNeurologicalSci*, vol. 3, pp. 41–47, Jun. 2016, doi: 10.1016/j.ensci.2016.02.010.

[7]T. W. Abate, B. Zeleke, A. Genanew, and B. W. Abate, “The burden of stroke and modifiable risk factors in Ethiopia: A systemic review and meta-analysis,” *PLoS One*, vol. 16, no. 11, p. e0259244, Nov. 2021, doi: 10.1371/journal.pone.0259244.

[8]J.Brownlee,“A Gentle Introduction to Ensemble Learning Algorithms,” MachineLearningMastery.com. Accessed: Feb. 29, 2024. [Online]. Available: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

[9]M. Agazhe *et al.*, “Incidence and pattern of stroke among patients admitted to medical ward at Yirgalem General Hospital, Sidama Regional State, Southern-Ethiopia,” *SAGE Open Medicine*, vol. 9, p. 205031212110011, Jan. 2021, doi: 10.1177/20503121211001154.

[10]B. Deresse and D. Shaweno, “Epidemiology and in-hospital outcome of stroke in South Ethiopia,” *J Neurol Sci*, vol. 355, no. 1–2, pp. 138–142, Aug. 2015, doi: 10.1016/j.jns.2015.06.001.

[11]JOHNHOPKINSMEDICINE,“Stroke.” Accessed: Mar. 25, 2024. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke

[12]H. J. Lee, S.-I. Jang, and E.-C. Park, “Effect of adherence to antihypertensive medication on stroke incidence in patients with hypertension: a population-based retrospective cohort study,” *BMJ Open*, vol. 7, no. 6, p. e014486, Jul. 2017, doi: 10.1136/bmjopen-2016-014486.

[13] Akkio,“7 Reasons Why Machine Learning Forecasting Is Better Than Traditional Methods,” Akkio. Accessed: Oct. 28, 2023. [Online]. Available: https://www.akkio.com/post/5-reasons-why-machine-learning-forecasting-is-better-than-traditional-methods

[14]D. Mwiti, “A Comprehensive Guide to Ensemble Learning: What Exactly Do You Need to Know,” neptune.ai. Accessed: Feb. 29, 2024. [Online]. Available: https://neptune.ai/blog/ensemble-learning-guide

[15]Abhishek, “A Comprehensive Guide to Google Colab: Features, Usage, and Best Practices,” Analytics Vidhya. Accessed: Oct. 14, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/03/google-colab-machine-learning-deep-learning/

[16]C. Avail, “Which are the Programming Languages Supported by Google Collaboratory?,” Medium. Accessed: Sep. 28, 2024. [Online]. Available: https://medium.com/@codeavail/which-are-the-programming-languages-supported-by-google-collaboratory-6737b4c1613d

[17] Saturn Cloud,“How to Install a Library Permanently in Colab | Saturn Cloud Blog.” Accessed: Sep. 28, 2024. [Online]. Available: https://saturncloud.io/blog/how-to-install-a-library-permanently-in-colab/

[18]H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *IJDKP*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[19]R. Kimmons and R. E. West, “Rapid Academic Writing,” *Rapid Academic Writing*, 2018, doi: 10.59668/27.

[20] King SaudUniversity,“A comprehensive review on ensemble deep learning: Opportunities and challenges,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.

[21]G. Boesch, “Ensemble Learning: A Combined Prediction Model (2024 Guide),” viso.ai. Accessed: Mar. 28, 2024. [Online]. Available: https://viso.ai/deep-learning/ensemble-learning/

[22]M. Kalirane, “Ensemble Learning in Machine Learning: Bagging, Boosting and Stacking,” Analytics Vidhya. Accessed: Mar. 28, 2024. [Online]. Available:

https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/

[23] Corporate Finance Institute,“Ensemble Methods,” Corporate Finance Institute. Accessed:Mar.28,2024.[Online].Available:https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/

[24]M. Alene, M. A. Assemie, L. Yismaw, and D. B. Ketema, “Magnitude of risk factors and in-hospital mortality of stroke in Ethiopia: a systematic review and meta-analysis,” *BMC Neurol*, vol. 20, no. 1, p. 309, Dec. 2020, doi: 10.1186/s12883-020-01870-6.

[25]H. Al-Zubaidi, M. Dweik, and A. Al-Mousa, “Stroke Prediction Using Machine Learning Classification Methods,” in *2022 International Arab Conference on Information Technology (ACIT)*, Nov. 2022, pp. 1–8. doi: 10.1109/ACIT57182.2022.10022050.

[26]R. Islam, S. Debnath, and T. I. Palash, “Predictive Analysis for Risk of Stroke Using Machine Learning Techniques,” in *2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, Dec. 2021, pp. 1–4. doi: 10.1109/IC4ME253898.2021.9768524.

[27]G. Sailasya and G. L. A. Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” in *International Journal of Advanced Computer Science and Applications*, 2021. doi: 10.14569/IJACSA.2021.0120662.

[28]Y. Wu and Y. Fang, “Stroke Prediction with Machine Learning Methods among Older Chinese,” *IJERPH*, vol. 17, no. 6, p. 1828, Mar. 2020, doi: 10.3390/ijerph17061828.

[29] Wikipedia,“Sidama Region,” *Wikipedia*. Dec. 08, 2023. Accessed: Apr. 11, 2024. [Online].Available:https://en.wikipedia.org/w/index.php?title=Sidama_Region&oldid=1188987395

[30]P. E. Burian, L. Rogerson, and F. R. Maffei, “The Research Roadmap: A Primer to the Approach and Process,” *Contemporary Issues in Education Research*, vol. 3, no. 8, pp. 43–58, Aug. 2010.

[31]M. Lenzerini, “Data Integration: A Theoretical Perspective,” presented at the Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Jun. 2002, pp. 233–246. doi: 10.1145/543613.543644.

[32] ResearchGate, “Schematic diagram of the random forest algorithm | Download ScientificDiagram.”Accessed:Sep.29,2024.[Online].Available:https://www.researchgate.net/figure/Schematic-diagram-of-the-random-forest-algorithm_fig3_355828449

[33] Geeksforgeeks,“LightGBM (Light Gradient Boosting Machine),” GeeksforGeeks. Accessed: May 03, 2024. [Online]. Available: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

[34]İ. Kılıç, “Light GBM: A Powerful Gradient Boosting Algorithm,” Medium. Accessed: Sep. 29, 2024. [Online]. Available: https://medium.com/@ilyurek/light-gbm-a-powerful-gradient-boosting-algorithm-fe145a1cd8a6

[35] Julia Nikulski,“The Ultimate Guide to AdaBoost, random forests and XGBoost | by Julia Nikulski | Towards Data Science.” Accessed: Apr. 25, 2024. [Online]. Available: https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f

[36]P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, “Ensemble Learning for Disease Prediction: A Review,” *Healthcare*, vol. 11, no. 12, Art. no. 12, Jan. 2023, doi: 10.3390/healthcare11121808.

[37]T. Srivastava, “12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2023),” Analytics Vidhya. Accessed: May 10, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

[38] Eugenia Anello, “Machine Learning Evaluation Metrics: Theory and Overview,” KDnuggets.Accessed:May,10,2024.[Online].Available:https://www.kdnuggets.com/machine-learning-evaluation-metrics-theory-and-overview

[39]“What is Accuracy, Precision, Recall and F1 Score?” Accessed: May 10, 2024. [Online]. Available: https://www.labelf.ai/blog/what-is-accuracy-precision-recall-and-f1-score

[40] Deepchecks,“What is Normalization in Machine Learning? Techniques & Uses,” Deepchecks.Accessed:Sep.30,2024.[Online].Available:https://www.deepchecks.com/glossary/normalization-in-machine-learning/

[41] Encord, “Introduction to Balanced and Imbalanced Datasets in Machine Learning.” Accessed: Sep. 30, 2024. [Online]. Available: https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/

[42]D. Elreedy and A. F. Atiya, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” *Information Sciences*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[43] Run,“Hyperparameter Tuning: Examples and Top 5 Techniques.” Accessed: Jun. 10, 2024. [Online]. Available: https://www.run.ai/guides/hyperparameter-tuning

[44] Dremio,“Grid Search,” Dremio. Accessed: Sep. 14, 2024. [Online]. Available: https://www.dremio.com/wiki/grid-search/

[45] NVIDIA Data Science Glossary, “What is XGBoost?,” NVIDIA Data Science Glossary.Accessed:May.03,2024.[Online].Available:https://www.nvidia.com/enus/glossary/xgboost/

[46] Kolena,“Feature Importance: Methods, Tools, and Best Practices Kolena.”Accessed: Sep. 30, 2024. [Online]. Available: https://www.kolena.com/guides/feature-importance-methods-tools-and-best-practices/

[47]Analytics vidhya,“Random Forest Algorithms-Comprehensive Guide with Examples”.AccessedOctober29,2023.https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[48]Engineering Education (EngEd) Program | Section,“Machine Learning with XGBoostandScikit-learn”.Accessed;October30,2023. https://www.section.io/engineering-education/machine-learning-with-xgboost-and-scikit-learn/

# APPENDICES

## Appendix A: Patient Dataset

| Age | Sex | Residence | Blood Press | RBC | Pulse rate | Hypertensi | Diabetics | Heart Disea | BMI | Smoking | Alcholic | Discontinue | Cholestrol | Stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 71 | 1 | 0 | 130 | 4.04 | 82 | 1 | 1 | 0 | 22.98 | unknown | unknown | 0 | 167 | 1 |
| 61 | 1 | 1 | 130 | 2.22 | 60 | 1 | 0 | 0 | 19.38 | unknown | unknown | 0 | 186 | 0 |
| 38 | 1 | 1 | 130 | 4.4 | 83 | 0 | 1 | 0 | 23.93 | unknown | unknown | 0 | 183 | 0 |
| 59 | 0 | 1 | 160 | 4.63 | 85 | 1 | 0 | 0 | 20.8 | unknown | unknown | 1 | 346 | 1 |
| 32 | 1 | 0 | 130 | 5.4 | 65 | 0 | 0 | 0 | 31.91 | unknown | unknown | 0 | 183 | 0 |
| 39 | 0 | 1 | 150 | 4.88 | 100 | 1 | 0 | 1 | 30.07 | unknown | unknown | 1 | 224 | 0 |
| 56 | 1 | 1 | 130 | 4.57 | 89 | 0 | 1 | 0 | 25.99 | unknown | unknown | 0 | 151 | 1 |
| 48 | 0 | 1 | 130 | 4.39 | 80 | 0 | 1 | 0 | 20.06 | unknown | unknown | 0 | N/A | 0 |
| 60 | 0 | 1 | 130 | 4.42 | 90 | 1 | 1 | 0 | 27.68 | unknown | unknown | 1 | 222 | 1 |
| 57 | 1 | 1 | 120 | 3.63 | 82 | 0 | 1 | 1 | 30.71 | unknown | unknown | 0 | 234 | 0 |
| 64 | 0 | 0 | 180 | 3.59 | 75 | 1 | 0 | 0 | 19.33 | unknown | unknown | 1 | 143 | 1 |
| 62 | 0 | 0 | 100 | 3.95 | 88 | 0 | 1 | 0 | 22.6 | unknown | unknown | 0 | 166 | 0 |
| 48 | 0 | 1 | 160 | 4.71 | 79 | 1 | 0 | 1 | 28.04 | unknown | unknown | 0 | 168 | 0 |
| 43 | 1 | 1 | 160 | 4.54 | 100 | 1 | 0 | 1 | 24.72 | unknown | unknown | 0 | N/A | 0 |
| 54 | 0 | 1 | 120 | 4.32 | 88 | 0 | 0 | 1 | 30.11 | unknown | unknown | 0 | 131 | 0 |
| 61 | 0 | 1 | 130 | 4.59 | 70 | 0 | 0 | 1 | 22.58 | unknown | unknown | 0 | 156 | 1 |
| 61 | 1 | 0 | 110 | 3.8 | 75 | 0 | 1 | 0 | 25.61 | unknown | unknown | 1 | 154 | 1 |
| 54 | 0 | 1 | 140 | 3.31 | 105 | 1 | 0 | 1 | 27.63 | unknown | unknown | 0 | 158 | 0 |
| 85 | 1 | 1 | 160 | 4.05 | 72 | 1 | 1 | 0 | 22.49 | unknown | unknown | 0 | 132 | 1 |

## Appendix B: Sample Python Libraries Used for Proposed Model Development

```
# Import Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from lightgbm import LGBMClassifier
```

```
from xgboost import XGBClassifier
from sklearn.metrics import
confusion_matrix,accuracy_score,classification_report
from sklearn.metrics import f1_score
from sklearn.metrics import precision_score,recall_score
```

### Appendix C: Remove duplicates and Handle missing values

```
df.drop_duplicates(inplace=True)
df.dropna(inplace=True)
```

Appendix D: Sample Code

```
param_grids = {
    'RandomForest': {
        'n_estimators': [100, 200, 300],
        'max_depth': [10, 20, 30, None],
        'min_samples_split': [2, 5, 10],
        'min_samples_leaf': [1, 2, 4]
    },
    'XGB': {
        'n_estimators': [100, 200, 300],
        'learning_rate': [0.01, 0.05, 0.1, 0.2],
        'max_depth': [3, 6, 9]
    },
    'LGBM': {
        'n_estimators': [100, 200, 300],
        'learning_rate': [0.01, 0.05, 0.1, 0.2],
        'num_leaves': [31, 62, 100, 150]
    }
}

# Define the classifiers
classifiers = {
    'RandomForest': RandomForestClassifier(random_state=42),
    'XGB': XGBClassifier(random_state=42, use_label_encoder=False,
eval_metric='logloss'),
    'LGBM': LGBMClassifier(random_state=42)
}
```