Developing Thyroid Disorder Prediction Model Using Machine Learning Approach

Mihretu W/Yohanis Tsegaye

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Office of Graduate Studies

Adama Science and Technology University

Sep, 2023

Adama, Ethiopia

Developing Thyroid Disorder Prediction Model Using Machine Learning Approach

Mihretu W/Yohanis Tsegaye

Advisor: Tilahun Melak (Ph.D.)

A Thesis Submitted to the Department of Computer Science andEngineering

School of Electrical Engineering and Computing

Office of Graduate Studies

Adama Science and Technology University

Sep, 2023

Adama, Ethiopia

DECLARATION

I declare that this Master Thesis entitled "**Developing Thyroid Disorder Prediction Model using Machine Learning Approach**" is my original work and has not been submitted to any university for a similar purpose. The references used in this thesis are duly recognized by proper citations.

Mihretu W/Yohanis Tsegaye

_____	_____	_____
Name of student	Signature	Date

RECOMMENDATION

I/we, the advisor(s) of this Thesis, hereby certify that I/we have read the revised version of the thesis entitled "**Developing Thyroid Disorder Prediction Model using Machine Learning Approach**" prepared under my/our guidance by **Mihretu W/Yohanis Tsegaye**. Therefore, I recommend the submission of the Thesis to the department for further review and evaluation.

Tilahun Melak (PhD)

_____	_____	_____
Advisor/Supervisor	Signature	Date

APPROVAL SHEET

I/we, as the advisors of the thesis titled "**Developing Thyroid Disorder Prediction Model using Machine Learning Approach**" and developed by **Mihretu W/Yohanis Tsegaye** hereby confirm that the recommendations and suggestions provided by the examining board have been duly integrated into the final version of the thesis.

Tilahun Melak (PhD)

Major Advisor/Supervisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis by Mihretu W/Yohanis Tsegaye have read and evaluated the thesis entitled '**Developing Thyroid Disorder Prediction Model using Machine Learning Approach**" and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

Chairperson	Signature	Date
Internal Examiner	Signature	Date
Dr. Hussien Seid		Oct. 09, 2023
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis proposal are contingent upon the submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office Of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to Almighty God for guiding me throughout this research journey, from its inception to its successful completion in such a remarkable manner. I am deeply grateful to Mother Saint Mary and all the angels and saints for the numerous and exceptional blessings they have graciously granted me.

Then I sincerely thank my advisor, Dr. Tilahun Melak, for his invaluable guidance and unwavering support. His inspiration and motivation were instrumental in driving this research forward. Dr. Tilahun dedication to mentoring me contributed significantly to the successful completion of this study. I cannot find sufficient words to express my gratitude for his professionalism and understanding.

I am also profoundly grateful to the staff members of TASH and SPMMC for their cooperation during the data collection process. Special thanks are extended to Dr. Theodros Aberra (Internist, Consultant Endocrinologist) and Dr. Getahun Tarekegn (MD, Internist, Endocrinology Consultant) for their keen interest, guidance, and commitment to knowledge sharing. I take this opportunity to express my heartfelt appreciation to my family members for their unwavering encouragement and support throughout the journey of this thesis work.

Lastly, I wish to extend my sincere thanks to all my friends who provided valuable support and assistance during the entirety of my academic pursuit.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

A

ADA: - Adaptive Boosting

ANN: - Artificial Neural Network

B

BFE: - Backward Feature Elimination

BIDFE: - Bi-directional Feature Elimination

C

CSV: - Common Separated Value

CV: - Cross-Validation

D

DNN: - Deep Neural Network

DT: - Decision Tree

F

FFS: - Forward Feature Selection

FS: - Feature Selection

FT3: - Free Thyroxine

FT4: - Free Triiodothyronine

L

LR: - Logistic Regression

M

ML: - Machine Learning

MLFS:- Machine learning Feature Selection

MLP: - Multi-Layer Perception

N

NB:- Naïve Bayes

P

PCA: - Principal Component Analysis

R

RF: - Random Forest

S

SMOTE: - Synthetic Minority Over-Sampling Technique

SPMMC: - Saint Paul's Millennium Medical College

SVM: - Support Vector Machine

T

TASH: - Tikur Anbessa Specialized Hospital

TDD: - Thyroid Disorder Dataset

TDP: - Thyroid Disorder Prediction

TSH: - Thyroid Stimulation Hormones

X

XGBoosting: - Extreme Gradient Boosting

U

UCI: - University of California

W

WHO: - World Health Orginazation

ABSTRACT

Thyroid disorder is a prevalent disorder affecting endocrine system on a globally, placing a significant burden on healthcare budgets, particularly in low-income countries. Early prediction and prevention of this disorder based on symptoms represent a critical challenge for healthcare sectors, especially in developing nations like Ethiopia. Machine learning has a vital role in the analysis of vast medical datasets and offers solutions to the intricate task of early disease classification. Recently, predicting thyroid disorder prediction has gained significant importance as a task. Despite the current methods for diagnosing this condition, it often involve binary classification, utilize limited datasets, imbalacnce dataset and lack proper validation. Current efforts mostly concentrate on refining models, neglecting feature engineering. this study introduces an approach that addresses these limitations by delving into feature selection for machine learning models. Various techniques are employed, such as FFS, BFE, bi-directional feature elimination, and feature selection through machine learning, involving the use of extra tree classifiers. The Proposed approach aims to predict different types of thyroid disorders: Negative, Hyperthyroidism and Hypothyroidism. The data for this study was gathered from Tikur Anbessa Specialized Hospital and Saint Paul's Millennium Medical College. the five algorithms employed in the study were Random Forest, Logistic Regression, Support Vector Machine, Adaptive Boosting and Extreme Gradient Boosting. The models were assessed using a stratified 10-fold cross-validation technique. and an analysis of their classification performance, enabling a comparison between the different models. the performance of the given classification techniques was evaluated using accuracy, precision, recall and F1- score. the results of the performance evaluation showed that the XGBoost with the feature selection with cross-validation method performs better than other models, which obtained an accuracy of 98.9 and F1-score of 99.1. the proposed thyroid predictive model classifiers a common thyroid disorder based on a multi-class prediction approach to help domain experts.

Keywords: Thyroid disorder, Machine Learning, Backward Feature Elimination, Global Burden

CHAPTER ONE

1. INTRODUCTION

1.1 Background of the Study

The thyroid gland is an essential endocrine organ responsible for producing and releasing thyroid hormones. These hormones have a crucial role in governing the development, growth, reproduction, metabolism, and functioning of nearly all organs within the human body (Garber et al., 2012). Thyroid disorders encompass a set of non-communicable disease conditions resulting from abnormalities in the function and structure of the thyroid gland. These disorders can manifest as disruptions in the secretion of thyroid hormones, enlargement of the thyroid gland, or discomfort (Rugge et al., 2015). The functional data of the thyroid gland is crucial for accurate interpretation and diagnosis of diseases related to the gland. They are prevalent globally, and they affect people of all ages, races, and genders (Chaganti et al., 2022). The thyroid gland produces Free thyroxine (FT4) and Free triiodothyronine (FT3). Thyrotropin is responsible for the regulation of these two thyroid hormones (TSH). Abnormalities in the thyroid gland can result to several disorders, with hypothyroidism and hyperthyroidism being two of the most prevalent conditions. Hyperthyroidism occurs when the thyroid gland produces too much thyroid hormone, which can lead to symptoms such as increased appetite, weight loss, anxiety, sweating, irritability, and an irregular or rapid heartbeat. There are several causes of hyperthyroidism, including, toxic multinodular goiter, Graves' disease and thyroiditis. Treatment options may include radioactive iodine therapy, medications, or surgery[1]. Hypothyroidism occurs when the thyroid gland is not producing enough thyroid hormones, which can result in symptoms such as fatigue, constipation, weight gain, hair loss, dry skin, and sensitivity to cold temperatures. There are several causes of hypothyroidism, including radiation therapy, autoimmune disease, and certain medications. Treatment options may include hormone dietary changes, replacement therapy, and lifestyle modifications[2].

One of the most pressing issues in human society is healthcare since it directly affects how well inhabitants live their lives. However, it is a fact that seemingly invisible diseases afflict our

[1] *https://www.thyroid.org/hyperthyroidism/.*
[2] *https://www.thyroid.org/hypothyroidism/.*

families and cause individuals to suffer. Healthcare today plays a wider and more significant part in people's lives than ever in all countries, rich and poor. These conditions include thyroid disorders.

Thyroid disorders represent a significant global health issue and are the second most prevalent category of endocrine disorders, just after diabetes mellitus. They contribute to a substantial burden of endocrine diseases, accounting for approximately 30% to 40% of the cases (Vanderpump, 2011). These disorders impact an estimated 300 million individuals worldwide, with the concerning fact that over half of them may be unaware of their condition. The primary thyroid disorders of concern are hyperthyroidism and hypothyroidism, and these conditions put approximately 1.6 billion people at risk across more than 110 countries globally (Yadav et al., 2013). Hyperthyroidism and hypothyroidism can result from thyroid gland issues, often secondary to hypothalamic or pituitary gland problems. In regions with iodine-deficient diets, goiter and active thyroid nodules can become widespread, affecting up to 15% of the population. Additionally, the thyroid gland can be susceptible to various tumor types and autoimmune attacks, such as those caused by endogenous antibodies (autoantibodies)[3]. The prevalence of thyroid disorders is highest in Africa, and the World Health Organization (WHO) has reported that the continent represents more than 25% of the global burden of the disease (Reta Demissie, 2019). For instance, in Zambia, the prevalence of thyroid disease is reported to be 7.3%, while in Uganda, it is 3.6%. Studies haves shown that thyroid disease is very prevalent in Ethiopia; many hospitals are burdened with the disease. As per a UNICEF report from 1993, a significant portion of Ethiopia's population, approximately 78%, was affected by iodine deficiency, with 62% experiencing iodine deficiency and 26% presenting goiter. In some areas, the prevalence of goiter was as high as 59%. (Study done in Jimma, Southwest Ethiopia) (Suga & Abebe, 2020). And, the according to (Mohammed, 2020) Between February 2020 and June 2020, a study conducted at the Endocrinology Referral clinic of TASH revealed that 39.5% of the patients exhibited clear signs of hyperthyroidism, and 56.8% were diagnosed with hypothyroidism. Additionally, subclinical hypothyroidism and subclinical hyperthyroidism were detected in 2.6% and 1.1% of the 271 total patients, respectively. Therefore, the thyroid disorders are more prevalent in women than in men, and the risk increases with age. They are also more common in regions with insufficient dietary

[3] *https://www.who.int/westernpacific/health-topics/thyroid-diseases.*

iodine, which is an essential nutrient for thyroid hormone production. the burden of thyroid disorders is not limited to their direct effects on health. Untreated thyroid disorders can lead to several complications, including heart disease, infertility, and birth defects. Machine learning approaches can aid in the prediction and diagnosis of thyroid disorders by analyzing large amounts of patient data and identifying patterns and trends. Machine learning algorithms are trained on vast amounts of patient data, including symptoms, lab test results, and medical history, to develop models that can accurately predict the likelihood of thyroid disorders. These models can assist healthcare professionals in making informed decisions regarding diagnosis and treatment.

In recent times, machine learning methods have demonstrated encouraging outcomes when it comes to predicting and diagnosing thyroid disorders. These approaches can aid in the development of personalized treatment plans, reducing the risk of misdiagnosis and ineffective treatment. Moreover, machine learning can help in the early detection of thyroid disorders, which can lead to better patient outcomes and reduced healthcare costs (Ha et al., 2021). This study aims focus on thyroid disorder prediction using machine learning model based on dataset obtained from public hospital of Ethiopian such Tikur anbessa Specialized and St. Paul's Hospital Millennium Medical College. Due to advancements in data processing and computational technologies, machine learning classifier algorithms are now utilized to predict thyroid diseases at their early stages and categorize them into types such as Negative, hypothyroidism, hyperthyroidism, and more.

1.2 Motivational of the Study

The value of excellent health cannot be overstated in our lives. Happiness, peace, and success cannot exist without good health. This study is motivated by the profound and wide-ranging impact of thyroid disorders on individuals' health and well-being. Both hypothyroidism and hyperthyroidism have been associated with serious health risks, including cardiovascular disease and osteoporosis. These conditions not only lead to increased morbidity but also contribute to mortality (Journy et al., 2017). Thyroid disorder occurs when the thyroid gland does not produce the usual levels of hormones, resulting in disruptions to bodily functions. Physicians can identify such disorders through physical examination and medical assessments, and subsequently, they initiate a treatment plan. The diagnostic procedure relies on a range of tests, including blood and

urine tests. As in many areas, data mining and machine learning methods can be used to diagnose thyroid disorders. Providing intelligent, low-cost solutions to a variety of issues, machine learning has become an essential component of daily life. With this method, it is possible to use time more efficiently while also reducing diagnoses that result from human error. According to a WHO survey and other statistical studies, thyroid illness is the leading cause of disease burden worldwide. and also, Ethiopia is one of the African countries where the burden of thyroid disorders is bothersome.

This dilemma serves as the motivation for the study to focus on predicting thyroid diseases using machine learning techniques. These techniques aim to analyze medical data effectively and make accurate predictions based on common symptoms.

1.3 Statement of the Problem

Thyroid dysfunction is a prevalent condition that can be easily identified and effectively treated. However, if left undiagnosed or untreated, it can lead to severe and far-reaching negative consequences. Recent decades have witnessed a substantial growth in our comprehension of the impact of thyroid hormones in both normal and pathological situations. It is now evident that evident thyroid dysfunction is linked to substantial levels of illness and mortality. Now a day the world is facing a shortage of 4.3 million healthcare institution workers (EPH, ICF, 2019). As a result, various chronic illnesses, including thyroid gland disorders, pose significant health challenges and contribute to a substantial number of global fatalities.

Detecting thyroid diseases can be a complex and demanding process, typically involving clinical assessments and numerous blood examinations. Within the realm of medicine, machine learning emerges as a vital tool for disease detection, offering a multitude of classification techniques to enhance diagnostic accuracy. The clinical decisions are usually based on the doctor's intuition. According to (Sousan, 2016), Ethiopia is below the WHO standard for health workers to population ratio the standard is health workers 2.18 per 1000 population. These include low health service, disparities in hospital setups or accessible equipment, diagnostic delay, and the similarity of the disease's symptoms, which leads to misdiagnosis. Furthermore, medical practitioners cannot easily predict the disease, which demands expertise and higher knowledge for predict (Desalew, 2020).

Early detection and prevention of thyroid disorders based on symptoms present a significant challenge to the healthcare sector. Machine learning plays a pivotal role in analyzing extensive medical datasets and offers solutions for the complex task of early disease classification. However, existing approaches to thyroid disorder diagnosis have several limitations, including a preference for binary classification, reliance on limited dataset sizes, and issues with imbalanced data. Moreover, current efforts predominantly focus on improving machine learning models, often overlooking the critical aspect of feature engineering.

This study aims to propose a machine learning approach to address the challenges faced by medical domain experts in identifying thyroid disease predictions based on symptoms and appropriate datasets. Furthermore, the research seeks to facilitate early disease diagnosis and provide expert recommendations for suitable treatments.

1.4 Research Questions

This study aims to investigate and address the following questions.Those are:

RQ1: What are the determinant attributes for the prediction of thyroid disorder?
RQ2: Which machine learning algorithm is the best to construct the classifier model from the selected algorithm?

1.5 Objectives of the Study

1.5.1 General Objective

The Main objective of this study is to build a model that can predict thyroid disorders using Machine learning Approach

1.5.2 Specific Objectives

The specific objectives of this study are listed below.

- To review the existing strategies and algorithms used for Prediction in Thyroid Disorder
- Collect and prepare patient symptom data for model training
- Select the pertinent features from the original feature set for predicting thyroid diseases.
- To use an appropriate model and hyper-parameters for thyroid disorders prediction

- To evaluate and analyze the performance of the model

1.6 Significance of the Study

The significance of this study is described from different perspectives. The ambitious goals of the health sector transformation plan, which are in line with our nation's second growth and transformation plan, include increasing equity, coverage, and utilization of essential health services, improving healthcare quality and strengthening the health sector's capacity for system-wide implementation. The research was contributed to providing a prediction result of Thyroid disorder for health workers in health sectors easily. The proposed model carries substantial significance by addressing critical healthcare challenges, leveraging machine learning for disease prediction, emphasizing feature engineering, and delivering practical insights for improved thyroid disorder diagnosis, especially in resource-limited settings. Specifically, the significance of this research: -

- Improve medical procedures for identifying and treating Thyroid Disorder.
- Reduce the Medical cost

1.7 Scope and Delimitation of the Study

The scope of this study was focus to develop a machine-learning technique, to predict a class of disease such as Negative, Hypothyroidism, or Hyperthyroidism based on different symptoms that enables the expert to give appropriate diagnosis and treatment recommendations based on data collected from TASH, and SPMMC dataset.

We are concentrating our attention on the TASH and SPMMC datasets because these hospitals have dedicated endocrinology departments. These departments specialize in addressing various endocrine-related issues and provide medical services to both rural and urban populations. These areas have a high prevalence of thyroid problems, which can result from factors such as iodine deficiency, thyroiditis, and excessive iodine consumption. This study encompasses the entire process, from collecting and preparing the dataset to preprocessing and ultimately making predictions. Various machine learning algorithms are applied to the dataset for classification purposes. These algorithms include Logistic Regression (LR), Random Forest (RF), Support

Vector Machine (SVM), ADA, and XGBoost.

Delimitation of the study

Nevertheless, in Ethiopia, some hospitals and healthcare facilities exist, and obtaining well-organized electronic data is a challenging task. The process of transitioning from manually recorded data to electronic formats is also labor-intensive and time-consuming. Data collected from TASH, and SPMMC and There is no established guideline for the selection of pertinent features to predict thyroid disorders. This study considers only three common disorders namely: - negative hypothyroidism and hyperthyroidism and severity of each class was not addressed.

1.8 Organization of the Thesis

This section, described the structure and organization of the whole study as follows.

Chapter 1: the introduction part of the study includes the background, motivation, statement of the problem, research questions, significance of the study, scope, and limitation of the study described.

Chapter 2: states the scientific literature related to thyroid disorder and machine learning concepts and related works

Chapter 3: this chapter defined the research methodology, the source of the data, how to prepare the dataset, feature selection, and model evaluation

Chapter 4: this chapter states the proposed solution for thyroid disorder classification, proposed feature selection, data understanding, the model selected, and evaluation methods used

Chapter 5: describe the implementation and experimentation of the proposed solution, implementation of features selection, data preprocessing, working environment data set description model implementation, and evaluation.

Chapter 6: Discusses the outcomes of the dataset's class distribution, the process of feature selection, and the evaluation of performance both with and without feature selection.

Chapter 7: it describes the conclusion, Recommendation and feature work for further investigation of this research work

CHAPTER TWO

2. LITERATURE REVIEW AND RELATED WORK

The study reviews relevant and important works of literature from various books, journal articles, related reports, conference papers, published and unpublished documents, and various websites. Recent advancements in data processing and computational capabilities have enabled the prediction of thyroid disease through machine learning and deep learning methods. Early identification and classification of this condition as Negative, Hypothyroidism, or Hyperthyroidism are crucial for prompt treatment and improved recovery rates. This chapter's review encompasses a range of topics, including thyroid disorders, machine-learning techniques, data preprocessing, feature selection methods, prediction models, and model evaluation approaches.

2.1 Thyroid Disorder and Its Global Burden

The thyroid gland is situated in the lower anterior part of the neck, precisely at the level of the second to third tracheal rings, positioned just below the larynx. This gland has a distinctive shape resembling a shield, and its name is derived from the Greek word "thyreos," which means "shield," and it is named after the thyroid cartilage of the trachea due to its shieldlike appearance (“Wolters Kluwer Lippincott Williams & Wilkins,” 2012). This gland is composed of two lobes, with one situated on each side (right and left) of the tracheal wall. They are connected with a thin strip of thyroid tissue that extends across the anterior surface of the trachea called the isthmus (Fig. 2.1). Each lobe measures approximately 3–4 cm in length, 2 cm in width, and typically a few millimeters in thickness. The isthmus, on the other hand, is usually just a few millimeters thick and can be up to 15 mm in height.

The main role of the thyroid gland is to produce thyroid hormones, which are then released into the bloodstream. These hormones are essential for controlling metabolism and supporting growth and development in both adults and children. Understanding the functional aspects of the thyroid gland is vital for accurately diagnosing and interpreting diseases related to this organ. Essentially, the thyroid gland's primary function is to regulate the body's metabolism, and the levels of hormone secretion from the thyroid can have a profound impact on human growth and development. Thyroid disorders are widespread worldwide, affecting individuals of all age, races and gender.

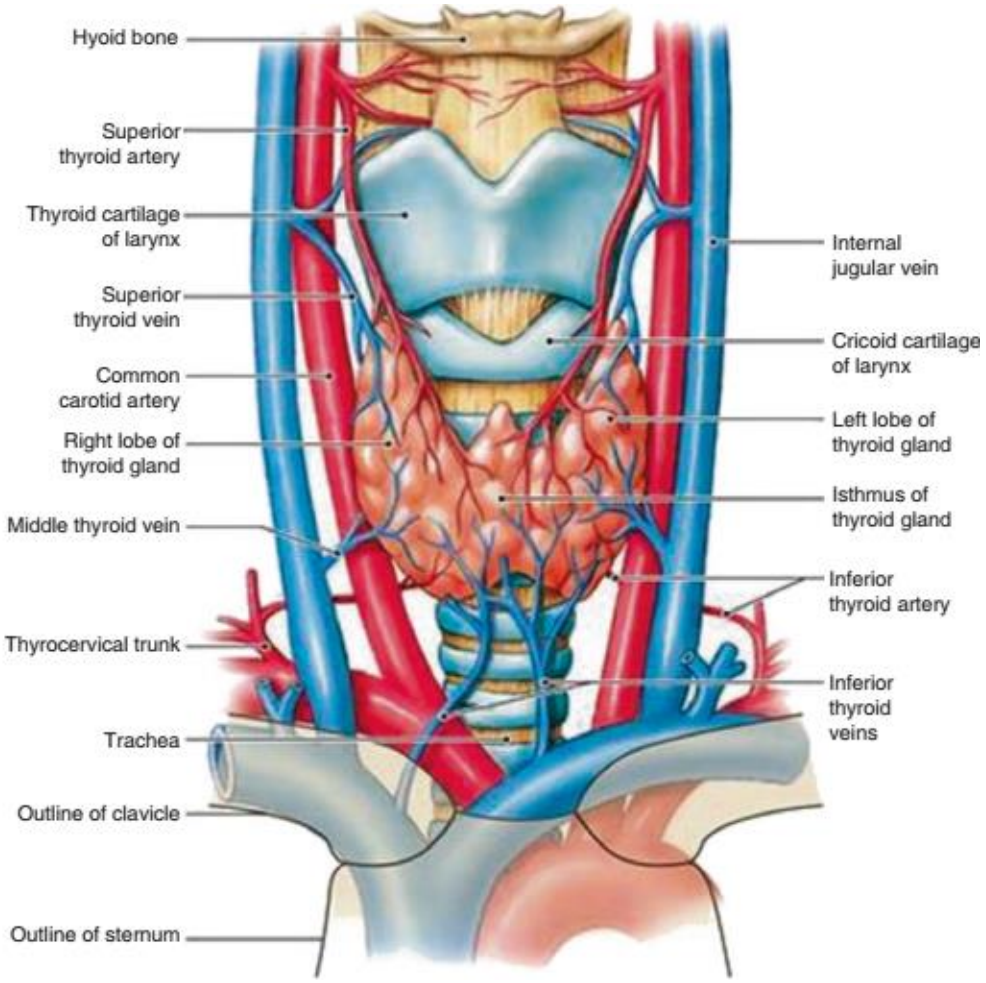


Figure 2-1 Anatomy of the thyroid gland[4]

The worldwide burden of thyroid disorders is significant, with an estimated 300 million people worldwide suffering from thyroid disorders. According to the WHO, about 30% of the global population may have some form of thyroid disorder. The prevalence of thyroid disease in Ethiopia is not well documented. However, a study conducted in 2017 found that the prevalence of thyroid disorders among adults in Ethiopia was 8.3%. This was higher than the global average of 4.6%. The most common type of thyroid disorder found in the study was hypothyroidism, which accounted for 6.2% of cases. Other types of thyroid disorders included hyperthyroidism (1.7%), goiter (0.4%), and nodular goiter (0.1%). The prevalence of thyroid disease was higher among women than men, with 9.2% and 7.3%, respectively. And, the according to (Mohammed, 2020) Between February 2020 and June 2020, a study conducted at the Endocrinology Referral clinic of TASH revealed that 39.5% of the patients exhibited clear signs of hyperthyroidism, and 56.8%

[4] *https://anatomytool.org/content/rcsi-drawing-thyroid-and-parathyroid-glands-and-vasculature-english-labels*

were diagnosed with hypothyroidism. Additionally, subclinical hypothyroidism and subclinical hyperthyroidism were detected in 2.6% and 1.1% of the 271 total patients, respectively. Therefore, the thyroid disorders are more prevalent in women than in men, and the risk increases with age. They are also more common in regions with insufficient dietary iodine, which is an essential nutrient for thyroid hormone production. the burden of thyroid disorders is not limited to their direct effects on health. Untreated thyroid disorders can lead to several complications, including heart disease, infertility, and birth defects.

In conclusion, thyroid disorders are a significant global health issue that affects millions of people worldwide. Early diagnosis and treatment are crucial to managing the condition and reducing the burden on individuals and healthcare systems.

2.2 Common Thyroid disorder

Some common thyroid disorders include hypothyroidism, hyperthyroidism, Graves' disease, Hashimoto's thyroiditis, goiter (enlarged thyroid), and thyroid nodules[5].

2.2.1 Hypothyroidism

Hypothyroidism is a medical condition that arises when the thyroid gland fails to produce an adequate amount of thyroid hormones necessary for the body's functions. The thyroid gland, resembling a small butterfly-shaped organ situated in the neck's base, plays a crucial role in regulating metabolism, heart rate, and body temperature. Dysfunction of the thyroid gland can result in various symptoms and potential health issues.

There are several possible causes of hypothyroidism, including autoimmune disorders (such as Hashimoto's thyroiditis), radiation therapy to the neck, surgical removal of the thyroid gland, and certain medications. In rare cases, congenital hypothyroidism (present at birth) may occur.

The symptoms of hypothyroidism may differ based on the seriousness of the condition and the individual, but they can include increased body weight, fatigue, constipation, sensitivity to cold, hair loss, dry skin, depression, and memory problems. In some cases, hypothyroidism can also cause infertility, joint pain, and heart disease.

[5] *https://www.healthline.com/health/common-thyroid-disorders*

Diagnosis of hypothyroidism typically involves a blood test to measure the levels of thyroid hormones and thyroid-stimulating hormone (TSH) in the body. Treatment options for hypothyroidism typically involve taking daily thyroid hormone replacement medication, which can help to restore hormone levels and alleviate symptoms.

It's important to note that hypothyroidism is a common condition that affects millions of people worldwide. While it can be a serious health concern, it is also highly treatable with appropriate medical care[6].

2.2.2 Hyperthyroidism

Hyperthyroidism is a medical condition characterized by the thyroid gland's excessive production of thyroid hormones, resulting in an overly active metabolism. This can lead to various symptoms such as weight loss, nervousness, intolerance to heat, elevated heart rate, and fatigue. Several factors can trigger hyperthyroidism, including Graves' disease, toxic adenomas, and thyroiditis.

Graves' disease stands as the most prevalent cause of hyperthyroidism and is an autoimmune disorder. It involves the immune system producing antibodies that stimulate the thyroid gland to produce excessive thyroid hormones. Although the exact cause of Graves' disease remains unclear, it is thought to result from a combination of genetic and environmental factors.

Toxic adenomas are growths or nodules on the thyroid gland that produce thyroid hormones independently, without the usual regulation by the pituitary gland. Thyroiditis, on the other hand, refers to inflammation of the thyroid gland, which can temporarily lead to hyperthyroidism, followed by a shift to hypothyroidism.

The diagnosis of hyperthyroidism typically involves conducting blood tests to assess thyroid hormone and thyroid-stimulating hormone (TSH) levels. Additionally, imaging studies like ultrasound or radioactive iodine uptake tests may be employed. Treatment strategies for hyperthyroidism vary depending on the underlying cause and the severity of the condition but may encompass medications, radioactive iodine therapy, or surgical interventions. If left untreated,

[6] *https://www.thyroid.org/hypothyroidism/.*

hyperthyroidism can lead to complications such as osteoporosis, heart problems, and eye problems such as bulging eyes and vision changes[7].

2.3 Overview of Machine Learning

Several advancements have been made as a result of the Internet, computing, and other technologies, including in the domains of law, medical, agriculture, and e-commerce. to satisfy their customers, conserve time and other resources, and increase profits, many businesses use computer systems and artificial intelligence today. A branch of computer science and artificial intelligence called "machine learning" uses a ton of different organizational data to do its work. It has developed into a fascinating and crucial research field for resolving extremely difficult issues with a vast amount of data and is used for the identification of target patterns in the data. Huge amounts of data and algorithms are used in machine learning to simulate how people learn from experience and develop their performance. Machine learning entails instructing the machine on what actions to take and enabling it to perform these actions automatically through learned experiences, rather than relying on explicit programming. There are four main categories of machine learning methodologies: supervised learning, reinforcement learning, unsupervised learning, and semi-supervised learning approaches

2.3.1 Supervised Machine Learning

Supervised learning is a category of machine learning algorithms that involves creating a predictive model based on known labels or outcomes. This type of algorithm relies on a target or dependent attribute that needs to be predicted using a set of predictor or independent attributes. In other words, it learns patterns and relationships in the data to make predictions or classifications based on historical or labeled data. The supervised learning method uses labeled or categorized data. The unobserved data with unknown outputs are then categorized using the newly learned rule. The two most frequent supervised machine learning problems are classification and regression. Support Vector Machine, Decision Tree, Naive Bayes, Random Forest, Logistic Regression, K-Nearest Neighbor, etc. are some examples of supervised learning[8].

[7] *https://www.thyroid.org/hyperthyroidism/*
[8] *https://www.simplilearn.com/tutorials/machine-learning-tutorial/introduction-to-machine-learning*

2.3.2 Unsupervised Learning

The algorithm used to work with unlabeled data is called unsupervised learning. Without knowing the label to which the instance belongs, this algorithm is a very potent tool for data analysis and pattern and trend identification. It is the learning algorithm that can be used to divide a given data set into various groups. Unsupervised learning is frequently used, with K-means as an example (Elghazel & Aussem, 2013).

2.3.3 Semi-Supervised machine learning

Semi-supervised machine learning is a technique used in machine learning that makes use of both labeled and unlabeled data when training a model. It falls between unsupervised and supervised learning approaches, utilizing a small set of labeled data to train the model while incorporating a larger amount of unlabeled data to improve its accuracy. Semi-supervised learning methods find applications in various domains, including NLP, image recognition, and speech recognition. (Chapelle et al., 2009).

2.3.4 Reinforcement Learning

A form of machine learning called reinforcement learning enables an agent to learn from its surroundings by acting and getting rewarded for it. This method is built on trial and error, where the agent learns from its errors and progressively enhances its performance. Maximizing the overall reward received by the agent over time is the main objective of reinforcement learning (Perner, 2015).

2.3.5 Ensemble Learning

One of the most effective predictive models is created using ensemble methods, which are machine learning approaches. Several ensemble learning models exist, including boosting, bagging, stacking, and random forest. In general, meta techniques such as ensemble methods are used to improve forecast accuracy. utilized for jobs involving both classification and regression. The choice of which ensemble method to use depends on various factors, including the nature of the problem, the quality of data, and the trade-offs between computational complexity and model performance.

Boosting ensemble learning

This is an ensemble learning method that assembles several weaker learners to form a more robust learner. The approach involves repeatedly training weak models on the same dataset, with each successive model concentrating on the data points that previous models misclassified. The ultimate prediction is generated by consolidating the predictions of all the weak models. Prominent algorithms in this category include Adaboost, gradient boosting, and stochastic gradient boosting, such as XGBoost.

2.4 Machine Learning in Healthcare Sector

The healthcare sector is one of the most important sectors in the world. It is responsible for providing medical care to people and ensuring their health and wellbeing. Good quality healthcare helps to prevent disease and improve the quality of life in society. In a developing nation like Ethiopia, the healthcare sector must contend with a number of issues, including the rising burden of disease, morbidity, mortality, and disability, increased demand for health services, delayed diagnosis, higher social expectations, a lack of manpower, and issues with inefficiency and low worker productivity (Atun, 2015). A fundamental transformation of the healthcare system is the key factor to overcoming these challenges and achieving Sustainable development goals (SDG) by 2030. In recent years, the healthcare sector has seen a rapid growth in technology, with the introduction of machine learning (ML) being one of the most significant developments.

Machine learning plays a significant role in preventive healthcare by enabling early disease prediction. This early prediction simplifies the diagnosis process and allows healthcare practitioners to initiate early treatment for patients. Once a disease has been identified through screening, healthcare providers and experts can use these predictive models to identify individuals at high risk for the disease. They can then recommend appropriate tests or interventions to mitigate the risk and promote early intervention and prevention (Garg & Bansal, 2023). Machine learning can be used in the healthcare industry for a variety of activities, including disease identification and diagnosis, medication development and production, medical imaging diagnosis, personalized medicine, epidemic prediction, etc. In order to effectively determine the presence or absence of the disease based on the patient history data, machine learning techniques are crucial in the health sector. This aids in resolving issues with diagnosis and therapy. This study uses supervised machine learning algorithms to analyze labeled data to forecast thyroid problems.

2.4.1 Machine Learning Approach in Disease Prediction

People today deal with a variety of ailments because of their way of life, the environment, and other things. So, earlier disease prediction becomes a crucial responsibility to preserve the population's life. Yet, based just on symptoms, reliable disease prediction becomes too challenging for medical professionals (Dahiwade et al., 2019). Using the latest technologyin the health system was significantly make these difficulties easy to handle. Nowadays, machine learning is successfully used in the healthcare system to identify and forecast a number of critical diseases using patient history information. In healthcare systems, machinelearning allows the health expert to process complex and huge amounts of datasets and then analyze the dataset with understandable clinical knowledge (Dahiwade et al., 2019). As a result, machine learning used in the healthcare system has a big influence on disease prediction to offer therapy and improve disease diagnostic accuracy.

Different researchers used different machine learning algorithms, as discussed below.

Random Forest: - It is a supervised machine learning technique for classification and regression known as random decision forest. It works by creating several trees with randomly subsampled features. Because of its comparatively high accuracy, resilience, and simplicity of use, many researchers chose the random forest machine-learning method for disease prediction (Mishra et al., 2020).

Support Vector Machine: - It is a supervised machine learning technique used for regression and prediction issues, and as a result, for the prediction of thyroid disorders. With the help of the data points, SVM creates a line-shaped hyperplane that divides the best tags, often referred to as a decision boundary for labels. For this reason, most of the studies done in machine learning disease prediction used SVM due to its accuracy, support multiple classes, and don’t suffer the condition of overfitting (Farooqui & Ahmad, 2020). The main SVM properties and ideas are listed below:

- Kernel trick: The kernel function implicitly maps the data into a higher-dimensional space, allowing for non-linear decision boundaries to be created in the original feature space.
- Margin: find the hyperplane with the maximum margin, which is the distance between the hyperplane and the nearest data points of each class. The larger the margin, the better the classifier's generalization ability.
- Soft-margin SVM: In cases where the data is not perfectly separable, SVM allows for a

soft-margin formulation. This allows some instances to be misclassified, trading off between maximizing the margin and minimizing classification errors.

- Regularization parameter
- Support vectors
- Multiclass classification: it can be expanded to handle multiclass problems utilizing a variety of strategies, such as one-vs-one or one-vs-rest.

Logistic Regression: Logistic regression is a kind of linear classifier used to solve binary classification problems, where the goal is to figure out the chance of a given input belonging to one of two classes. Similar to linear regression, logistic regression takes the input features and combines them linearly. Instead of predicting the output directly, it applies a non-linear function (like the logistic or sigmoid function) to the linear combination of the features, transforming the output into a probability value between 0 and 1[9].

Extreme Gradient Boosting: It is a machine learning technique used for both regression and classification problems. It is an ensemble method that combines multiple weak models to create a strong model. It is an iterative algorithm that works by fitting a model to the residuals of the previous model, and combining the models into a final ensemble[10].

XGB starts by fitting a model to the data and calculating the residuals. The algorithm then fits another model to the residuals and combines it with the previous model to improve the predictions. This process is repeated until a stopping criterion is met, or until a maximum number of iterations is reached.

XGBoost's key attributes are:

- Regularization: XGBoost offers various regularization penalties to avoid overfitting, promoting effective generalization in model training.
- Parallelization: It utilizes multiple CPU cores during model training.
- Scalability: It is designed to scale effectively, making it suitable for handling large datasets and complex tasks.
- Cross-validation: XGBoost includes cross-validation as a built-in feature.

[9] https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/

[10] https://neptune.ai/blog/xgboost-everything-you-need-to-know

- Non-linearity: It has the capability to identify and learn from non-linear patterns in data.

XGBoost often performs better on large datasets. It is optimized for both computational speed and predictive accuracy on extensive datasets. If you have a substantial amount of data, XGBoost may be more efficient.

Adaptive Boosting (ADA): - ADA is a technique in Machine Learning used as an Ensemble Method. It is specifically designed for binary classification tasks, but can also be extended to multi-class problems. Several "weak" classifiers are combined to form a stronger classifier as part of the AdaBoost method. A weak classifier refers to a basic model that achieves only slightly better performance than random guessing, and an example of this is a decision stump, which is essentially a decision tree with just a single split. On various subsets of the training data, the algorithm iteratively trains weak classifiers, increasing the weight of the misclassified cases with each iteration[11].

2.5 Feature Selection

It is an important step in machine learning, as it can improve the performance of the model, reduce overfitting, and increase interpretability. The specific feature selection method used will depend on the characteristics of the dataset, the number of features, and the specific machine learning algorithm being used. This helps to reduce the computational cost, better model interpretability, and leads to better performance (Dhal & Azad, 2021). Common feature selection methods include filter, wrapper, and embedded methods (Fu et al., 2022).

Figure 2-2 Feature Selection[12]

[11] analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/

[12] *https://scikit-learn.org/stable/modules/feature_selection.html*

Feature Extraction

It is a process in machine learning and computer vision where relevant features are automatically extracted from raw data, such as images, text, or audio, to represent the data in a more compact and meaningful way. The goal of feature extraction is to transform the input data into a set of features that can be used by a machine learning algorithm to make accurate predictions or classifications and also, independent component analysis, Principal component analysis (PCA), and linear discriminant analysis algorithms are some of the commonly used feature extractions used in the health dataset.

2.5.1 Feature Selection Methods

Filter, wrapper, and embedding techniques are the three categories into which it is divided (Fu et al., 2022).

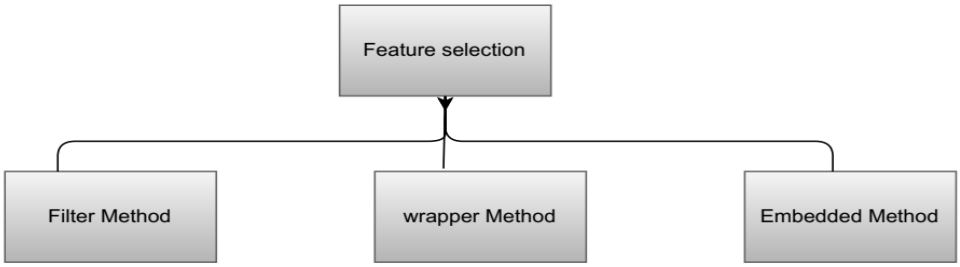


Figure 2-3 Filter Method Categories[13]

2.5.1.1 Filter Methods

These methods are a class of feature selection methods that rank the importance of features based on their intrinsic characteristics, without considering the machine learning algorithm used. Chi-squared feature selection, mutual information-based feature selection, and correlation-based feature selection are some common filtering techniques.

Filter methods are popular for their computational efficiency and ease of use. They can be applied to large datasets with a large number of features, and provide a quick way to identify potentially relevant features. However, filter methods may not identify complex interactions between features,

[13] https://www.datasciencesmachinelearning.com/2019/10/feature-selection-filter-method-wrapper.html

and may select redundant or irrelevant features. The techniques are discovered to be quick, simple, scalable, and independent of any machine learning classifier (Chen et al., 2020).

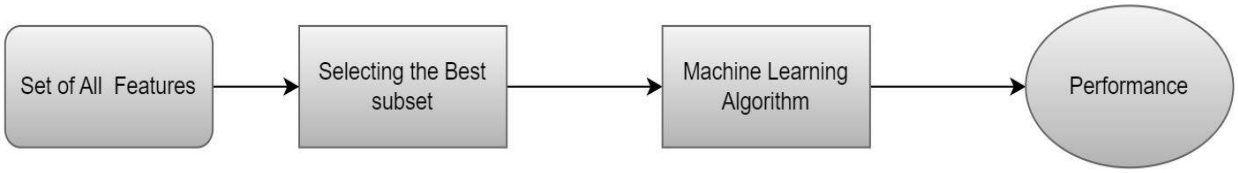


Figure 2-4 Flowchart for the Filter Method [14]

2.5.1.2 Wrapper Methods

These methods are a type of feature selection method in machine learning, where a subset of features is selected based on the performance of a specific machine learning algorithm. In wrapper approaches, a subset of features is chosen by assessing how well a machine learning algorithm performs while using various subsets of features. The algorithm is trained on different subsets of features, and the performance of the algorithm is evaluated based on a metric such as accuracy, precision, or recall.

Wrapper methods can be computationally expensive, but they often result in better performance compared to other feature selection methods. However, they are not always guaranteed to find the optimal subset of features (Chen et al., 2020).

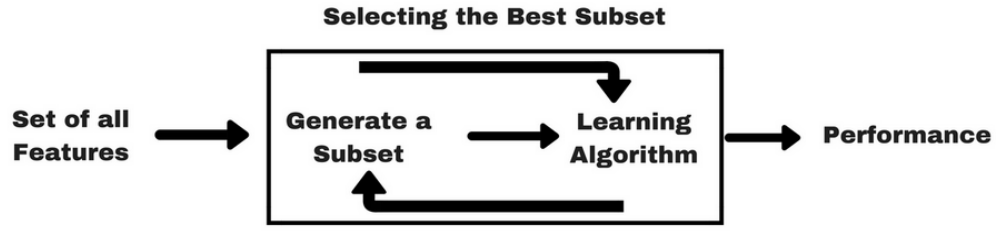


Figure 2-5 Flowchart for the Wrapper Method [15]

[14] https://analyticsindiamag.com/what-are-feature-selection-techniques-in-machine-learning/
[15] https://www.analyticsvidhya.com/blog/2021/03/7-popular-feature-selection-routines-in-machine-learning/

2.5.1.3 Embedded Methods

these methods are a type of feature selection technique in machine learning where feature selection is performed as part of the model training process. Unlike filter methods, which select features based on their statistical properties, or wrapper methods, which evaluate subsets of features using a specific model, embedded methods perform feature selection by incorporating feature selection directly into the model building process.

Embedded methods are particularly useful in situations where there are many features available, and the number of features is much larger than the number of samples. In such cases, embedded methods can help in reducing the risk of overfitting, as they are able to identify the most important features for the model during the training process (Chen et al., 2020).

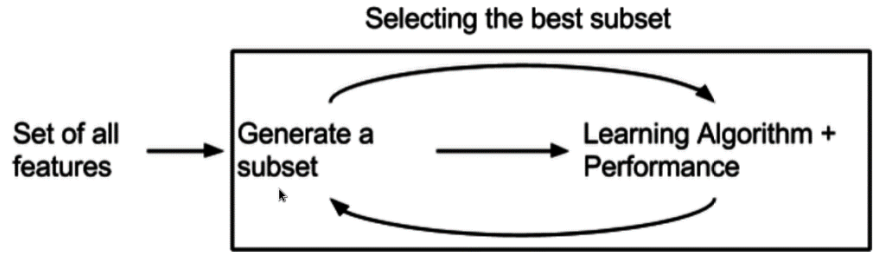


Figure 2-6 Flowchart for the Embedded Method [16]

Table 2-1 Feature Selection methods

Feature Selection Method	Description	Pros	Cons
Filter methods	These methods utilize statistical measures to assess and rank the features within a dataset. Subsequently, they select the highest-ranked features based on a predetermined threshold.	- Fast and computationally efficient. - Simple to implement. - Can be applied to large datasets. - No need to train a model.	- May not choose the best collection of features for a particular model. - May not consider relationships between features.
Wrapper methods	These methods use a machine learning algorithm	- Can select the optimal set of features for a	- Computationally expensive and time-

[16] https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

	to train and evaluate a subset of features on a model, and then iteratively add or remove features based on their impact on the model's performance.	given model. - Takes into account the interactions between features.	consuming. - Can lead to overfitting if the training set is small.
Embedded methods	These methods integrate feature selection with the training of a machine learning model. The algorithm selects the most relevant features during the training process.	- Can select the optimal set of features for a given model. - Can handle complex feature interactions.	- Computationally expensive and time-consuming. - May require a large amount of data to train the model.

2.6 Related Works with Thyroid disorder

The literature has offered several methods for detecting and classifying thyroid disease. For example, (Garcia De Lomana et al., 2021) LR, RF, SVM, GBM, and DNN used machine learning techniques to predict the molecules that are likely to influence thyroid hormone balance, particularly in the early stages of thyroid disease, early prediction of these molecules is valuable. The ToxCast databases were used to gather molecular events for this study. The thyroid hormone receptor (TR) and thyroid peroxidase (TPO), which had F1 scores of 81% and 83%, respectively, were shown to have the best predictive performance. The authors introduced a multi-kernel Support Vector Machine (SVM) for the classification of thyroid diseases. According to their findings, this multi-kernel SVM model achieved an impressive performance accuracy rate of 97.49% when tested on UCI thyroid datasets. The authors also mentioned that they utilized an improved gray wolf optimization technique to conduct feature selection, which further enhanced the performance of their model (Shankar et al., 2020).

They focused on making early predictions for hypothyroidism using various machine learning algorithms. The study evaluated five machine learning algorithms, including Support Vector Machine (SVM), Decision Trees (DT), Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF), in combination with three different feature selection techniques: univariate feature selection (UFS), recursive feature selection (RFE), and principal component analysis (PCA). The results indicated that the combination of RFE with machine learning algorithms outperformed the other feature selection methods. Specifically, when RFE was combined with the five machine

learning algorithms mentioned, they achieved an impressive accuracy rate of 98.35%. While high accuracy is desirable, particularly in the context of medical diagnosis, it is crucial to take into account other assessment metrics including recall, precision, and F1-score. However, it's worth noting that the study had a relatively small dataset consisting of only 519 records, and the authors highlighted the need for a larger dataset to further evaluate the effectiveness of their approach (Riajuliislam et al., 2021).

A performance comparison of various machine learning algorithms for classifying Thyroid disorder into Negative, hyperthyroidism and Hypothyroidism classification. They used a dataset sourced from the UCI machine learning library, consisting of 6200 samples, with each sample having 21 features. In their findings, the authors reported that Decision Trees achieved the highest performance with an accuracy rate of 98.23%. However, they noted that the classification was limited to these three categories. Additionally, the study lacked detailed information on data preprocessing methods and doesn't provide detailed information about the feature selection process, However, it doesn't elaborate on why these specific models were chosen or how their hyperparameters were tuned (Razia et al., 2018a). The authors (Leng et al., 2017) used feature selection approaches and image processing techniques to extract the key features from the dataset and get the greatest performance for predicting thyroid disease. The classification of thyroid disease is another significant issue in the health industry that has to be addressed.

In a study conducted by (Salman & Sonuc, 2021) The researchers evaluated the efficiency of machine learning algorithms in categorizing thyroid diseases. They utilized various machine learning algorithms in their study, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Trees (DT), K Nearest Neighbor (KNN), Naive Bayes (NB), and Multi-Layer Perceptron (MLP) to forecast thyroid diseases. Their dataset consisted of 1250 samples collected from laboratories and hospitals in Iraq. The MLP achieved a thyroid classification accuracy of 0.964. However, the authors acknowledged that there is potential for further enhancement in performance. (Hosseinzadeh et al., 2021) they introduced a MMLP technique for thyroid disease classification. When MMLP was applied in conjunction with a set of six networks, they achieved a 0.7% improvement in accuracy compared to using a single MLP. The MMLP achieved an impressive 0.99 classification accuracy on a large dataset. However, it's important to note that training deep learning techniques like MMLP can be computationally intensive and may require significant computational resources for faster training.

They conducted experiments using the K-Nearest Neighbors (KNN) algorithm with different distance functions to assess its capability in detecting thyroid diseases (Abbad Ur Rehman et al., 2021). Prior applying KNN with Cosine and Euclidean distances, the authors selected the best features using L1-based and chi-square feature selection methods. According to their findings, KNN yielded promising results. However, it's worth noting that the evaluation was based on a relatively small sample size of only 690 samples and accuracy 0.98. (Alyas et al., 2022) did a comparison of the artificial neural network (ANN), DT, RF, KNN, and other machine learning algorithms to identify thyroid disease. The experiments were carried out using the medium-sized dataset, which included both the original unsampled data and the sampled data, in order to diagnose thyroid disease. Random Forest achieved the highest accuracy at 0.948. Nevertheless, the authors did not conduct tests to predict the particular type of thyroid disorders. Additionally, authors utilized deep learning models for predicting the classification of thyroid disease.

Table 2-2 According to the evaluations listed below, several studies on the use of machine learning to predict thyroid disorders have been carried out. However, the prior conducted research was the limited feature set and dataset size, imbalanced Data, lack proper validation and the limited number of target classes considered. Beside Thyroid disorder prediction using a machine learning approach was not conducted using Ethiopia. On the other hand, the deficiency of iodine is high in developing countries like Ethiopia when compared to developed countries. This indicates that there is a difference in the thyroid dataset.

Table 2-2 Related Works

Authors and year	**Title**	**Methods and Findings**	**Gaps**
(Garcia De Lomana et al., 2021)	In silico models to predict the perturbation of molecular initiating events related to thyroid hormone homeostasis	ANN, RF, XGB, SVM LR	❖ The ToxCast datasets may not represent the full range of molecular events associated with thyroid hormone homeostasis. ❖ It is only use F1 score metric ❖ Didn't predict the type of thyroid disorders.
(Razia et al., 2018b)	A Comparative study of machine learning algorithms on thyroid disease prediction.	MLR, SVM, DT and NB	❖ Imbalanced Data (not provide information on the class distribution of the data ❖ Limited information on data preprocessing ❖ Didn't specify the feature selection process.
(Riajuliislam et al., 2021)	Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques	DT, SVM, LR, RF, and NB. RFE, UFS and PCA	❖ limited Sample Size ❖ Imbalanced Data ❖ didn't mention the other assessment measures like F1 score, recall, and precision
(Abbad Ur Rehman et al., 2021)	Analysis of the Performance of K-Nearest Neighbor Algorithms in Detecting Thyroid Disease	KNN without FS, KNN using chi-square-based FS and KNN using L1-based FS	❖ limited Sample Size ❖ Imbalanced Data (not provide information on the class distribution of the data) ❖ Limited Distance Metrics (only tested the KNN algorithm with Euclidean and Cosine distances)
(Alyas et al., 2022)	An Empirical Approach for the Classification of Thyroid Disease using Machine Learning	DT, RF, KNN, and ANN	❖ Didn't predict the type of thyroid disorders. ❖ Limited information about the data ❖ Didn't specify the other evaluation metric except accuracy

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Chapter Overview

This chapter discusses the general methodology and steps involved in creating a machine-learning-based prediction model for thyroid disorders based on determinant attributes. Today, machine learning is used in the health sector to manage a large amount of data that is difficult for humans to easily analyze. Different methods, techniques, and procedures must be used in each stage of the building process to create these prediction machine-learning models. Moreover, in this chapter data collection methods, preprocessing techniques, and machine learning models used in designing a thyroid disorders prediction model are explained. Lastly, the research performance evaluation metrics are given.

3.2 Research Design

The approach the researcher uses to gather data and preprocess it in order to address the research questions is known as research design. Quantitative, qualitative, or hybrid research designs are all possible. In this study, an experimental research design approach was employed. This approach falls under the quantitative research design category, where the research objectives are identified, and machine learning models are created to establish relationships between two variables: the dependent variable and the independent variables, within the dataset samples. (Shukla, 2022). The collected data have 23 features 22 independent and 1 dependent feature with a 3-target class.

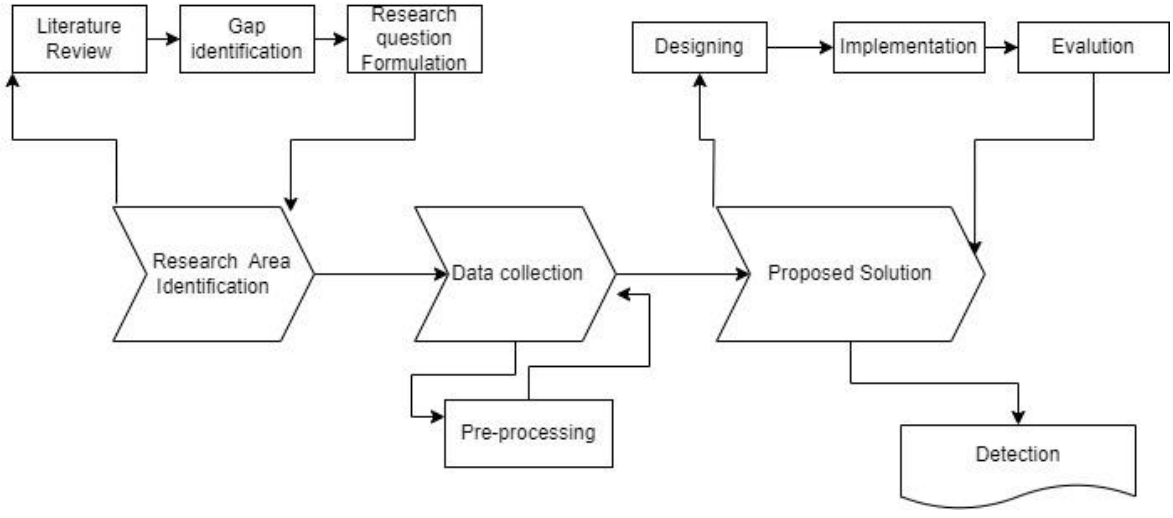


Figure 3-1 Methodology Flowchart for Thyroid Disorder Classification

3.2.1 Problem Identification

This study's crucial stage can help with comprehension of the subject matter or area of interest. We must be able to define the problem precisely and have a solid understanding of the dataset that will be utilized in the machine-learning activity before we can begin it. Due to the fact that the goal of machine learning techniques is decided by the data and the type of problem that needs to be solved. The review of several papers that concentrate on machine learning applications in healthcare was used as a source of support for this study. In addition, the discussion with domain experts was carried out; relevant documents were analyzed in depth to raise awareness about thyroid disorders. However, the investigator is unable to find published research on the prediction of thyroid disorders, namely Normal, Hypothyroidism, and Hyperthyroidism. Consequently, the researcher was motivated to develop an selected machine-learning model from SVM, RF, LR, ADA and XGBoost.

3.2.2 Data Source

Models that learn from examples and experience are created using machine learning algorithms. These examples are represented by datasets, which are collections of occurrences or observations. For the learning model to be trained, a high-quality dataset is necessary. These days, researchers either use public available datasets or local datasets that they have collected themselves. In this thesis work, the data was collected from TASH and SPMMC. TASH and SPHMC are two healthcare facilities that are situated in Ethiopia's capital city, Addis Ababa. TASH was founded in 1972, giving it a long history. It has changed through time and is intimately related to Addis Abeba University. It was named the primary teaching hospital by the federal ministry of health after 1998. TASH is now Ethiopia's biggest referral hospital, with 200 doctors on staff who work in a variety of clinics and divisions.

On the other hand, SPHMC, St. Paul Hospital and Millennium Medical College, has a rich history dating back to its founding in 1969. Emperor Haile Selassie collaborated with the German Evangelical Church in its establishment. Initially named St. Paul General Specialized Hospital, it underwent a significant transformation in 2007 with the opening of a medical school in the new millennium era of the Ethiopian calendar. Today, it stands as the largest specialized hospital in Ethiopia, contributing to healthcare and medical education in the region.

3.2.3 Data Collection

The main goal of this study is to use machine learning techniques to categorize thyroid diseases and to do so in the context of common diseases. TASH and SPMMC provided the data for this study, which had to be preprocessed because the raw data was unsuitable for the training and testing of machine learning models. Preprocessing data and dataset labeling for thyroid diseases prediction is a crucial step in the development of a dataset for Thyroid disorder model development.

3.3 Thyroid Disorder Prediction Modelling

3.3.1 Data Preprocessing

Preprocessing plays a crucial role in the development of our prediction model for this study. It serves as a means to transform large and noisy data into a clean and valuable dataset. Given that our collected data includes categorical variables and missing values, it's imperative to process it effectively to ensure it's in a suitable format for model utilization. The significance of preprocessing becomes evident when we consider that inconsistent data can significantly impact the accuracy of our machine-learning model. After gathering and preparing the dataset, the subsequent step is preprocessing, which involves cleansing the data before it is input into the classifiers. This preprocessing phase is indispensable for achieving accurate and reliable disease predictions, making it an integral component of our overall prediction model development process.

Cleaning Noise Data: - This process of cleaning noise data involves finding and fixing errors in datasets that can negatively affect the performance of the model. It helps machine-learning models get a quality dataset so they can be more accurate.

Handling Missing Values: There were instances of missing values in the collected data, primarily resulting from errors during the data collection process. Additionally, some patient history data remained incomplete. These gaps in patient data often included missing diagnostic test results, which are crucial for predicting the likelihood of diagnoses and assessing treatment effectiveness. The presence of missing values can significantly affect the precision of the prediction model. Dealing with these missing values in the dataset can be approached in various ways. One option is to remove instances with missing values, particularly if the impact on individual cases is minimal. Alternatively, missing values can be substituted with zeros, which won't alter the model's outcomes. Moreover, addressing missing data can involve data imputation techniques. This entails

replacing the absent values with statistical measures such as the mean, mode, or standard deviation. Another method is to predict the missing values by utilizing information from the available dataset.
Handling Categorical Data: in this step, the collected information is converted into a format that represents categories numerically. In this process, nominal data, which includes values like 'Yes' or 'No,' is transformed into numerical values, typically 0 and 1. For example, in the case of the "Vomiting" attribute, 'No' may be represented as 0, and 'Yes' may be represented as 1. This numerical representation allows machine learning models to work with categorical data effectively. In handling categorical data and after preprocessing the data should be in CSV file forms which hold the entire integer and float values data.

3.3.2 Feature Scaling

Feature scaling is a data preprocessing technique used to standardize or normalize the range of features or variables in a dataset. It is an important step in many machine learning algorithms, as features with different scales can have a disproportionate impact on the model's performance. Certain machine learning algorithms, including SVM and KNN, are sensitive to feature scaling because they use scaler products to exploit distance or similarity between data points. On the other hand, feature scaling has no effect on machine learning techniques like GBoost(GB) , random forest (RF), and Naive Bayes (Brownlee, 2020).
In this study from different feature scaling methods, the StandardizeScaller is a feature scaling technique used to transform the features of a dataset to have a mean of 0 and a standard deviation of 1. It is removing the mean and scaling to unit variance.

3.3.3 Feature Selection

It is essential to identify the most relevant characteristics in order to apply the machine-learning algorithm to provide the best results from the data. In feature selection, the most helpful features from the initial dataset are picked, and these attributes are then used as inputs for the models. Because certain initial dataset properties may actually be less significant for machine learning predictive modeling than others. Basically, feature selection is all about choosing the most relevant and important features in order to make the machine learning process more efficient and help the model to perform better. It's also connected to reducing the number of features in the data set, which is called dimension reduction. However, this is different from feature selection, as it

involves creating new attributes by combining existing ones, whereas feature selection only involves including or excluding features already present. The three main categories of feature selection methods, namely filter, wrapper, and embedded techniques. Wrapper method are applied in various clinical datasets, including those related to thyroid disorders (Phyu & Oo, 2016), (Sun et al., 2019), (Anggraeni et al., 2021).

3.3.3.1 Wrapper Methods

Wrapper methods are feature selection techniques that treat the selection of features as a search problem. Instead of evaluating the individual features, wrapper methods evaluate different subsets or combinations of features by training a model on each subset and assessing its performance (He et al., 2018). It is offering a comprehensive and tailored approach to feature selection, making them well-suited for disease prediction tasks where accuracy and model relevance are critical (Arun Kumar et al., 2022).

3.3.3.1.1 Forward Feature Selection

In the process of Forward Feature Selection (FFS), the goal is to build a model using one feature at a time, starting with an empty model (Solorio-Fernández et al., 2019). In each subsequent step, we choose the feature with the smallest p-value and then proceed to create models using two-feature combinations. At this stage, we prioritize using the combination of features that has the lowest p-value from the previous round. As we move on to building models with three-feature combinations, we continue to favor combinations with low p-values. This iterative process continues until the p-value of each feature in the feature set reaches a predetermined level of significance. At the end of this procedure, the final set of selected features comprises the most relevant and valuable ones for accurate detection and classification (Solorio-Fernández et al., 2019).

Step 1: Select the initial significance level (S) and initiate with an empty set.

$$Y\_0 = \{\emptyset\} \qquad (1)$$

Step 2: Utilize specific criteria to pick the initial feature. This may involve randomly selecting a feature from the list of available features. The selection process involves identifying the minimum value according to the equation provided. Features are chosen based on their p-values(probability value) from the entire set of features in use.

$$X^{+} = \underset{x \notin Y_k}{argmax}\, J(Y_k + x) \qquad (2)$$

Step 3: The iteration persists until the p-values of all the features are lower than or equal to the predefined significance level. Alternatively, the iteration halts if the minimum p-value identified falls below the significance level, and the total number of iterations equals the total number of features.

$$Y_k = Y_K + X^{+}; k = k + 1 \qquad (3)$$

3.3.3.1.2 Backward Feature Elimination

During the backward feature elimination process, we begin with a model that includes all the existing features. We then identify the feature with the highest p-value and remove it from the model, followed by model fitting. It's important to note that the significance level value should be greater than the p-value of the eliminated feature. We repeat this procedure while ensuring that every excluded feature has a p-value exceeding the significance level. This iterative process continues until all features with elevated p-values have been removed from the model. The final set of features remaining at the end of this process represents the most relevant and valuable ones that can be utilized for accurate detection and classification.

Step 1: In order to fit the model, start with all the features.

$$Y_0 = X \qquad (4)$$

Step 2: Choose the feature with a high p-value from the feature list. The significance level value (S) is contrasted with the high p-value feature. The condition x > S should be satisfied to consider the feature for elimination

$$X^{-} = \underset{x \in Y_k}{argmax}\, J(Y_k - x) \qquad (5)$$

Step 3: The high p-value item is eliminated from the list, and the process then returns to step 2 to carry out the subsequent iteration of feature removal (k + 1). The final list of features uses BFE to describe the chosen feature list when k = 0.

$$Y_k = Y_K - X^{-}; k = k + 1 \qquad (6)$$

3.3.3.1.3 Bi-Directional Elimination

The BiDFE method integrates both FFS and BFE techniques. It is similar to FFS in the sense that it selects new features one by one. However, it differs in that the backward elimination step comes into play after each new feature is chosen, where it assesses the newly selected feature in comparison to the features already chosen. Essentially, any previously selected features with p-values exceeding a specified significance level 'out' are excluded. In this method, two significance level values are computed, one using the 'in' range and another using the 'out' range In the feature selection process, a feature should have a p-value that is lower than the inner significance level and higher than the outer significance level to be considered for inclusion.

Step 1: We begin with a null set and initially select a feature based on predefined criteria. We then build the feature list using forward feature selection.

$$Y_F = \{\emptyset\}; Y_B = X \qquad (7)$$

Step 2: The p-value comparison is used to choose the next-best feature. The essential features are chosen using a conventional forward feature selection process.

$$X^+ = \underset{x \notin Y_{Fk}\, x \in Y_{Bk}}{argmax}\ J(Y_{Fk} + x) \qquad (8)$$

$$Y_{Fk} + 1 = Y_{Fk} + X^+ \qquad (9)$$

Step 3: The selection of the next optimal feature relies on comparing p-values. The subsequent feature is chosen using a conventional forward feature selection method, and any selected features that are deemed unimportant are later removed through a backward feature elimination process. When the value of 'k' gets closer to the total count of features, we can return to step 2 and reiterate this procedure.

$$X^- = \underset{x \in Y_{Bk} x \notin Y_{Fk+1}}{argmax}\ J(Y_{Bk} - x) \qquad (10)$$

$$Y_{Bk} + 1 = Y_{Bk} - X^-; k = k + 1 \qquad (11)$$

3.3.3.1.4 Machine Learning Feature Selection

To find the key features for a given task, one can use feature selection techniques based on machine learning, such as ensemble techniques. The extra tree classifier method has been taken into consideration. Methods used in this work for feature selection. With the help of the training dataset, the Extra Trees Classifier creates numerous decision trees at random (BABY et al., 2021). The nodes are recursively split based on a random subset of features to create each decision tree.

he Extra Trees Classifier makes decisions about splitting nodes in the decision tree by considering either entropy or the Gini index. These criteria help evaluate the quality of the split and determine the most suitable attribute for dividing the data.

The entropy (E) is calculated using the formula:

$$Entroyp(E) = -\sum_{i=1}^{C}(pi * log2(pi)) \qquad (12)$$

In the provided equation, Σ_i represents the summation across all potential classes (C), p_i signifies the probability associated with class i, and $\log_2$ denotes the logarithm with a base of 2.

Entropy serves as a metric quantifying the average information required to ascertain the class label of a randomly chosen item from a node. It gauges the level of disorder or unpredictability within the node by considering the distribution of class labels.

3.3.4 Machine Learning Model

To enhance the accuracy of predictive models, this research introduces a foundational learning model for prediction thyroid disorders. SVM, RF, LR, ADA, and XGB are integrated with and without feature selection techniques to formulate the machine learning models. These models undergo fine-tuning to maximize their effectiveness.

3.4 Predictive Model Evaluations

The crucial process of creating an accurate machine-learning model is performance evaluation. To check that the prediction model fits the dataset and performs effectively with fresh, unforeseen input data, model assessment techniques must be used. The goal of the model performance evaluation is to determine how well a model generalizes to unknown or out-of-sample data. Holdout and Cross-validation are the two subcategories of the performance evaluation approach. An objective assessment of learning performance is provided by holdout evaluation, which tries to evaluate a model on distinct data from that on which it was trained. Holdout divides mothed data into a validation, training, and test at random. Cross-validation is a useful method used to evaluate a model's performance using a test dataset that hasn't been used in its training process. This approach is especially beneficial in preventing overfitting in predictive models, particularly in situations where the available data is limited in quantity.

3.4.1 Cross-Validation

Hold-out is more straightforward, adaptable, and rapid than cross-validation. The issue with this

method's significant variability, though, is that performance variations may result from variations in the training and test datasets. Due to repeated recordings made during the training phase, the train-test-split technique may cause the model to overlook some data. Additionally, less data is used in the train test split technique to train and test the model.

Cross-Validation is one of the performance evaluation techniques for comparing and assessing models. It divides the data into two parts, one for training the model and the other for testing or validating it. The cross-validation technique is the most used evaluation technique because it prevents overfitting[17]. Therefore, we made sure that the models understood the pattern for the training dataset for this study by using the most used K-fold cross-validation assessment method. The original dataset is split into a training set and a test set using the cross-validation technique, which is used to train and test the model.

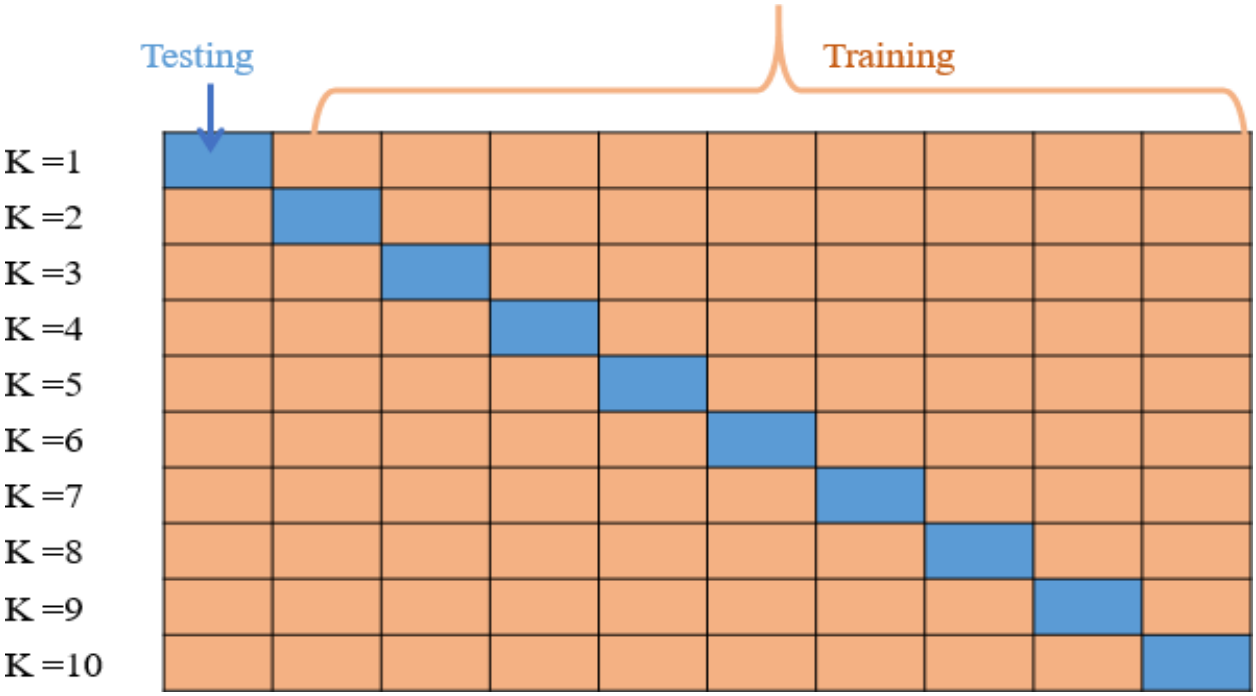
K-fold Cross-Validation is a technique used to assess the accuracy and practical performance of machine learning models. Additionally, this method is used to contrast various machine learning models, which enables us to select the one that will provide the highest accuracy score and be the most useful. Therefore, using the dataset, machine learning models can be created, and using this method, it can be determined how well the model would perform on untried data. With the following steps, it is a cross-validation method. Decide on a k-number of folds and divide the entire dataset into those folds first. K-1 folds of the training dataset should be used to build the model for each fold of the dataset. The model should then be put to the test to determine how well the kth fold performs. This process is repeated until all k-folds have been utilized as the test set, and the model's performance metric is determined by calculating the average accuracy across all k-folds, referred to as "cv accuracy."

Stratified K-Fold Cross Validation: K-fold cross-validation randomly partitions the dataset into folds, and this randomization can result in highly imbalanced folds, potentially causing bias during model training. When the data has a balanced class distribution, this approach can work effectively. However, in the case of an imbalanced class distribution, especially if the minority class is very small, it's possible for one or more folds to contain few or even no samples from the minority class. Because the model only needs to correctly predict the majority class, this leads to misleading model evaluation. Stratified k-fold cross-validation ensures that the class distribution within each data split closely resembles the distribution in the overall training dataset, thus preserving the balance

[17] https://www.geeksforgeeks.org/cross-validation-machine-learning/

of unbalanced classes in each fold (Widodo et al., 2022). The study utilizes the stratified 10-fold cross-validation method due to the imbalanced class distribution present in the collected data.

Table 3-1 Algorithm for Performing Stratified 10-Fold Cross-Validation



3.4.2 Prediction Model Performance Evaluation Metrics

The concept of using a machine-learning algorithm to create a predictive model is based on the idea of providing helpful feedbackCreate the model using a machine-learning algorithm, collect feedback on assessment metrics, make improvements, and keep continuing until the accuracy is what you desire. Metrics for performance assessment are able to assess the model's performance. The ability of evaluation metrics to distinguish between the outcomes of various models is a key feature. Accuracy, precision, recall, the f1-score, and confusion matrices are just a few of the metrics used in this study to assess the effectiveness of the chosen predictive modeling. The four types of outcomes that can result from classification are listed below.

- **True positives (TP)** happen when a prediction correctly determines that an instance belongs to a specific class. True positives are instances where the actual value is positive and the model correctly predicts it to be positive in the context of binary classification.

- When the model correctly predicts a negative example and the actual value is also negative, this is known as a **true negative** (TN). In other words, TN stands for situations in which the predicted value and the actual value are both negative.
- **False positives** (FP) happen when the model predicts an instance as belonging to a certain class when in fact it does not. In other words, FP stands for situations in which a positive value is predicted but a negative value actually occurs.
- **False negatives** (FN) occur when the model predicts incorrectly that an instance does not belong to a particular class when in fact the instance does. In other words, FN stands for situations where the predicted value is negative but the actual value is positive.

3.4.2.1 Accuracy

One of the most frequently employed performance evaluation metrics for classification tasks is accuracy (ACC). It calculates the model's accuracy in predicting the test dataset. It is determined by dividing the total number of predictions by the number of instances that were correctly classified.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (13)$$

3.4.2.2 Precision

For classification tasks, precision is a performance evaluation parameter that measures how well a classifier can recognize good examples. It calculates the percentage of accurate positive cases (true positive predictions) compared to all positive occurrences (positive predictions).

$$Precision = \frac{TP}{TP+FP} \qquad (14)$$

3.4.2.3 Recall

Recall, also referred to as sensitivity or the percentage of true positives, is a performance indicator for classification tasks that assesses a classifier's accuracy in identifying positive instances.

$$Recall = \frac{TP}{TP+FN} \qquad (15)$$

3.4.2.4 F1- Score

For classification tasks, the F1-score is a performance evaluation metric that combines recall and precision into a single value. It provides a balanced assessment of the classifier's performance by

calculating the harmonic mean of precision and recall.

$$F1\_Score = 2 * ((\frac{Precision * Recall}{Precision + Recall})) \quad (16)$$

3.4.2.5 Confusion Matrix

The method used to assess how well the classification model performs is called a confusion matrix. The confusion matrix includes data on the actual value and predicted value used to assess the model's performance based on the data in the matrix. In comparison to the actual classifications in the test data, the model's confusion matrix shows how many correct and incorrect predictions it made.

Table 3-2 Confusion matrix for three-class classification

		Predicted values		
Actual Values		Negative	Hyperthyroid	Hypothyroid
	Negative	True Negative	False Hyperthyroid	False Hypothyroid
	Hyperthyroid	False Negative	True Hyperthyroid	False Hypothyroid
	Hypothyroid	False Negative	False Hyperthyroid	True Hypothyroid

3.5 Design and Development Tools

Different tools were used to design and build the proposed work.

3.5.1 Design Tool

Draw.io: This tool serves as a versatile diagramming application designed to create a wide range of diagrams, such as flowcharts, process diagrams, network layouts, organizational charts, and more. It provides an easy-to-use interface along with a diverse array of shapes, symbols, and connectors, allowing users to visually convey information and relationships with precision and clarity.

3.5.2 Software Tools

Python[18]: For creating machine learning applications, Python is a well-liked and simple-to-learn programming language

TensorFlow[19]: This is an open-source platform that specializes in numerical computing and machine learning. It offers a range of libraries and resources for constructing and implementing machine learning models.

Anaconda Navigator[20]: - Help us to quickly and easily launch development applications. Manage Anaconda environments, channels, and packages without having to use command-line tools.

Jupiter Notebook[21]:- We can create and exchange documents with this free web application. It is helpful for machine learning, modeling, data visualization, and data cleaning and transformation.

Google Drive[22]: - A cloud storage service used to store the datasets and share them for preprocessing and model training on Google Colab.

Google Colab[23]: - It is a no-cost cloud-based Jupiter Notebook that doesn't require any configuration for use. Additionally, it provides free access to computational resources like GPUs.

3.5.3 Hardware Tools

The hardware components utilized in this research are given in the table below

Table 3-3 hardware tools

Tools	Used for
GPU	Accelerating training the ML models
RAM	Improving computer speed and performance
Hard Drive (HDD)	Storing the datasets and models

[18] python.org/doc/versions/
[19] https://www.tensorflow.org/
[20] https://docs.anaconda.com/free/navigator/index.html
[21] jupyter.org
[22] https://www.google.com/drive/
[23] colab.research.google.com

CHAPTER FOUR

4. PROPOSED THYROID DISORDER PREDICTION MODEL

In order to address the issues raised in chapter one and find an answer to the stated research questions that followed to close the gap discussed in the literature review, the proposed thyroid disorder prediction model using machine learning to classify thyroid diseases was discussed in detail in this chapter. Below is an illustration of the proposed solution's high-level architecture that explains each component's specific placement inside the suggested framework.

4.1 Proposed Thyroid Disorder Prediction Model Architecture

The goal of this study is to create a machine-learning model for thyroid disease prediction that substantially improves the classification of thyroid diseases. which classify the thyroid disease as disorder enter negative, hyperthyroid and hypothyroid based on the thyroid dataset collected from TASH and SPMMC.

The methodology makes an effort to investigate the potential of machine learning algorithms for thyroid disease prediction. Platform for preprocessing, platform for algorithms, and platform for training, as shown in the following figure. In the proposed architecture of this study, the initial dataset was in its raw form and underwent a series of preprocessing steps. These steps involved activities like data cleaning, addressing missing values, data transformation, and managing categorical variables. The ultimate aim was to prepare the dataset to be compatible with machine learning algorithms. Then Using a stratified K-fold Cross-Validation, preprocessed data is separated into training and testing sets. To train different machine learning algorithms, the training set is used. After the data has been preprocessed, feature selection techniques are used to find the features that are most important for the disease prediction. FFS, BFE, BiDFE, and MLFS using extra tree classifier with cross-validation is used in this step. these techniques help reduce the dimensionality of the dataset and improve the model’s performance.

Next, the preprocessed and selected feature dataset is split into training and testing sets using the stratified K-fold Cross-Validation. The training set is used to train various machine learning algorithms. For the platform of algorithms, there are numerous machine learning algorithms that can be explored for thyroid disease prediction, such as XGB, Logistic Regression, Random Forests, SVM and ADA. The training platform involves training the machine learning models

using the preprocessed dataset. The selected features from the feature selection step are used as inputs, and the corresponding targets (the thyroid disease classes) are used as the output. The models are trained using the training set, and their performance is evaluated using various metrics such as accuracy, precision, recall, and F1 score.

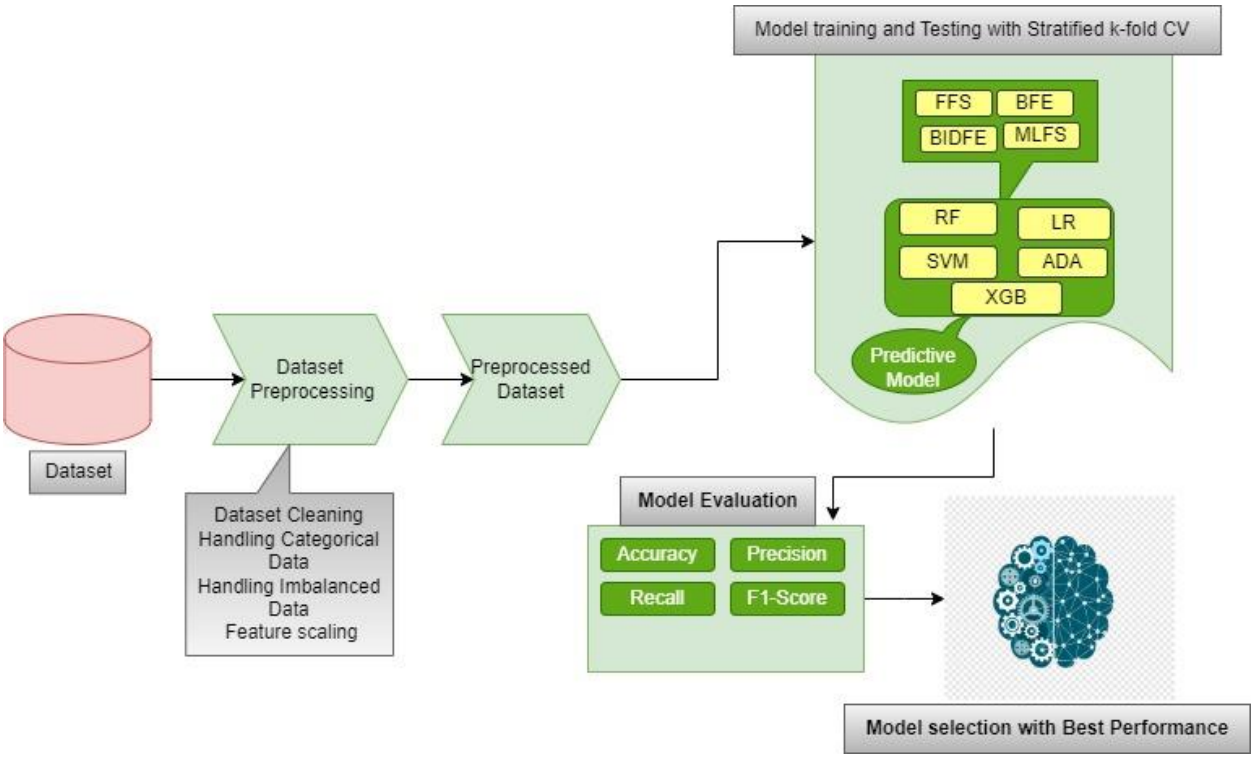


Figure 4-1 Proposed Thyroid Disorder Prediction Model Architecture

4.2 Understanding of the Data

After establishing a well-defined problem, the next step involves delving into the domain of data exploration. The objective here is to identify the datasets that are accessible. The success of machine learning and knowledge discovery relies significantly on the quality and quantity of these available datasets. This stage primarily revolves around gathering the desired dataset and determining its type, format, and quantity. During this process, the datasets are meticulously examined for completeness, missing values, redundancy, plausibility of feature values, and other relevant criteria.

Finally, the step includes the effectiveness of the dataset concerns machine learning technique (Eaton, 2018). The data from TASH and SPMMC hospitals were initially structured using an Excel file, but now there is a need to transform it into a format suitable for machine learning models. four year of datasets was collected from TASH and SPMMC. The total size of the initial datasets

in excel format is 511KB. The datasets used for the prediction of thyroid disorder was collected from TASH and SPMMC, Ethiopia, from 2019 up to 2023 with a total number of 5767 instances and 29 features.

Age: - The patient's age is essential in understanding the prevalence and management of thyroid diseases. Certain thyroid problems may be more likely to occur or be more of a risk depending on a person's age.

Sex: - The sex of the patient can influence the likelihood and characteristics of thyroid diseases. For example, autoimmune thyroid issues like Graves' disease and Hashimoto's thyroiditis are more common in women.

Query_on_thyroxine: - Is there anything uncertain about the patient taking their thyroxine medication?

On_thyroxine: - This feature specifies whether the patient is currently taking thyroxine medication, which is commonly used for thyroid hormone replacement therapy.

Pregnant: - Pregnancy can impact thyroid function, and thyroid disorders may arise or be affected during pregnancy. This feature indicates whether the patient is currently pregnant

Query_hypothyroid: - This feature suggests whether there is a suspicion or query regarding hypothyroidism, an underactive thyroid condition.

Query_hyperthyroid: - This feature indicates whether there is a suspicion or query regarding hyperthyroidism, an overactive thyroid condition.

TSH: - It is a hormone produced by the pituitary gland that regulates the thyroid gland's hormone production. TSH levels are commonly measured in thyroid function tests to evaluate thyroid function.

T3: - It is an active thyroid hormone. Measurement of T3 levels provides insights into thyroid function and potential abnormalities.

TT4: - represents the total amount of thyroxine hormone in the blood. TT4 measurement helps evaluate thyroid function and potential disorders.

T4U: - measures the binding capacity of thyroxine-binding globulin (TBG), a protein that transports thyroid hormones. It provides information about TBG availability and thyroid hormone binding.

FTI: - It is a calculated value that estimates the concentration of free thyroxine hormone in the blood. It combines measurements of TT4 and T4U to assess the biologically active form of T4.

4.3 Data Preprocessing

It is a method used to prepare data for machine learning models by undertaking tasks like addressing missing values, cleaning, managing categorical values, structuring raw data, and modifying data to make it suitable for use in machine learning models. The processed dataset is then converted to a CSV format, which can be easily parsed for further preprocessing in Python code. The collected data for this study requires cleaning, filling missing values, data transformation to improve quality, and handling categorical values.

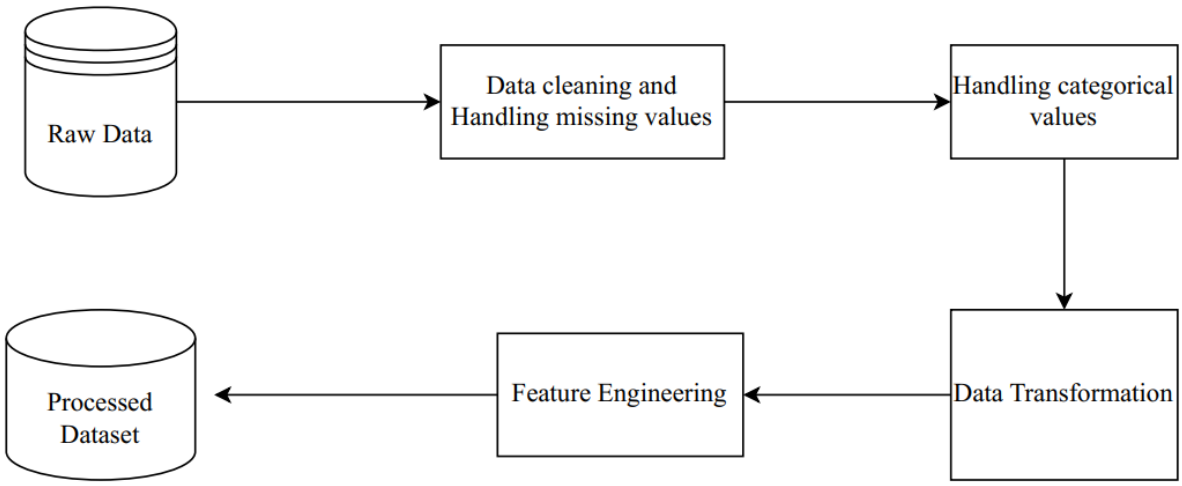


Figure 4-2 Dataset Preprocessing phase

4.3.1 Data Cleaning and Handling Missing Values

Data cleaning is a part of data preprocessing which is used to fix or get rid of incomplete or incorrectly formatted data in a dataset. In our study, the data we have gathered contains a limited amount of incorrectly formatted and unsuitable data that must be eliminated. Aside from the incorrectly formatted data, this dataset also has some missing values. Incomplete or missing data

is a frequent issue that has the potential to impact the process of data analysis. The dataset collected from TASH and SPMMC has total records of 5767 and 29 features including the class label. Table 4-1 show the features' names, the number of missing values, and the best method for imputed data. In order to fill in the missing values, the mean values of the numeric attribute were used, which contained missing data. The mode, which is the most frequent value, is used to fill in missing values when an attribute is categorical because it is the most common value.

Table 4-1 Missing value Imputation

Overall Missingness of TDP is: 6.22%

Attribute Name	Data type	Number of missing values	Missing values in %
TBG	numeric	5461	0.943093
T3	numeric	1492	0.255859
TSH	numeric	563	0.097321
T4U	numeric	420	0.072840
FTI	numeric	416	0.072127
TT4	numeric	198	0.034282
Sex	string	147	0.025194

4.3.2 Handling categorical Data

In this research, we convert categorical data into numerical data, making it suitable for creating machine-learning models. The label encoder employed assigns values ranging from 0 to n_classes-1, with n representing the count of distinct labels. If the label appears multiple times, it will retain the same value it was given before. The following are categorical data

Table 4-2 categorical data

Attribute name	Data type
sex	Category
on_antithyroid_meds	Category
on_thyroxine	Category
sick	Category
query_on_thyroxine	Category

pregnant	Category
thyroid_surgery	Category
Psych	Category
query_hypothyroid	Category
hypopituitary	Category
lithium	Category
Goiter	Category
Tumor	Category
query_hyperthyroid	Category
I131_treatment	Category
Target	Category

4.4 Feature Selection

After cleaning the data, addressing missing values, and handling categorical data, the subsequent step involves evaluating the most relevant features that enhance model performance while eliminating those features that introduce noise into the models. The goal of feature selection is to reduce the number of features to the smallest set believed to be most beneficial for improving the predictive accuracy of machine learning models, thereby decreasing overfitting and the time required for model training[24]. Various advanced feature selection methodologies were implemented, including FFS, BFE, BiDFE, and MLFS. These methods aid in extracting crucial features from the dataset for training the machine learning model.

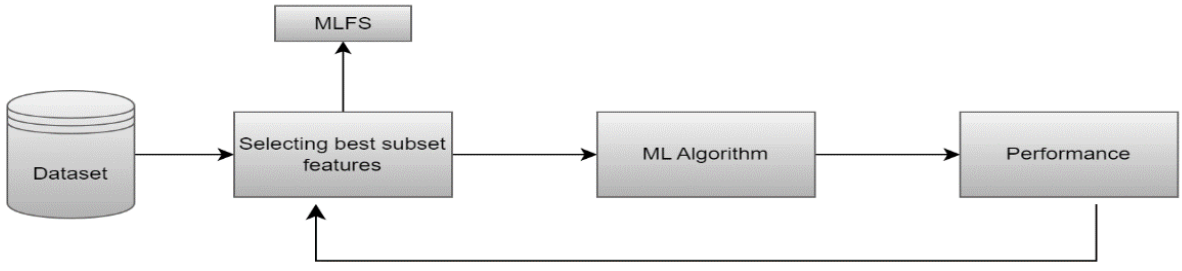


Figure 4-3 Diagram Illustrating the Workflow of Feature Selection

[24] https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

4.5 Machine Learning Model Building

The goal of this research is to use machine learning algorithms to build an automated that can predict which of the thyroid disorder: - Negative, hypothyroid and hyperthyroid based on the given features of both clinical and instrumental diagnosis parameters using machine learning algorithms. Hence, this study utilizes multiple machine learning models for predicting thyroid disorders. ADA , RF, SVM, LR, and XGB are employed to address the issue, and their performance is enhanced through the process of fine-tuning.

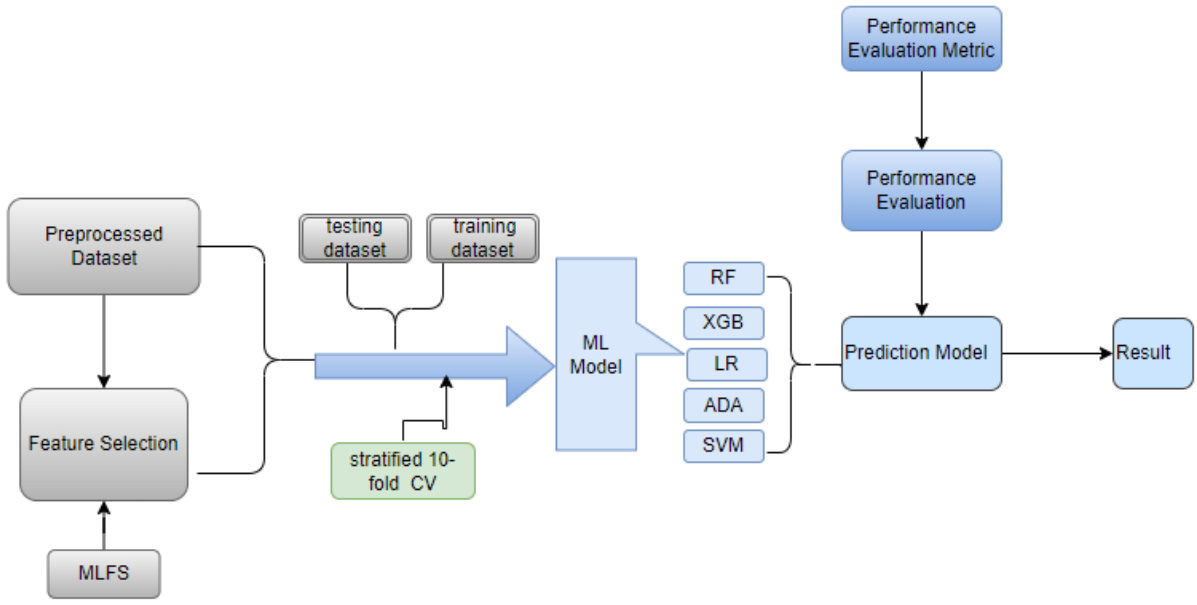


Figure 4-4 Diagram Illustrating the Process of Model Building

4.5.1 Hyperparameter Tuning

Hyperparameter tuning is the procedure of selecting the optimal hyperparameters for a machine learning model. Hyperparameters are variables that are chosen before the learning process and cannot be discovered through direct data learning. They control various aspects of the model's learning process and can have a significant impact on the model's performance. Hyperparameter configuration can be done manually or automatically. The first approach involves manually setting and experimenting with various groupings of hyperparameters. This is tedious and may not be practical in cases where there are many hyperparameters to try and a large search space. Automatic

hyperparameter tuning, on the other hand, uses algorithms and techniques to automate the process of finding the best hyperparameter configuration (Liashchynskyi, 2019). the second method used for hyperparameter optimization is random search. Random search and Grid search are two commonly used algorithms for this purpose. Grid search is a traditional method where a complete search is performed on a predetermined subset of the hyperparameter space. It systematically explores the parameter values by generating candidates from a specific grid. However, grid search can be inefficient and impractical in high-dimensional spaces. On the positive side, it can be easily parallelized since the algorithm works with independent hyperparameter values. On the other hand, random search overrides the complete selection of parameter combinations by randomly selecting them. It has been found to outperform grid search, especially when only a few hyperparameters significantly impact the performance of the machine learning algorithm.

This suggests that they expect only a subset of hyperparameters to have a significant impact on the performance of the machine learning algorithm under investigation. The random search algorithm provides flexibility and may lead to more efficient exploration of the hyperparameter space in such cases (Liashchynskyi, 2019). Random search for hyperparameter tuning involves randomly sampling hyperparameters from predefined ranges or distributions.

4.6 Model Evaluation and Testing

In this study, in order to evaluate and estimate the performance of machine learning models trained on the thyroid disease dataset, model evaluation and testing are essential. As multiple algorithms were used, it becomes important to compare and select the best model for further prediction. Using a stratified 10-fold cross-validation method, the prediction models' efficacy is assessed. Cross-validation helps in estimating how well the models generalize to unseen data. The dataset is divided into 10 folds of equal size, the models are trained on 9 folds, and evaluating their performance on the remaining fold. This process is repeated 10 times, each time using a different fold as the test set. This allows for a robust evaluation of the models' performance. Different performance evaluation metrics are used in this study to assess the models appropriately. These metrics include precision, recall, F1_score, accuracy, and the confusion matrix.

CHAPTER FIVE

5. IMPLEMENTATION OF THE PROPOSED SOLUTION

The proposed thyroid disorder prediction's architecture, design, and guiding principle were covered in the previous chapter. This chapter's actual code implementation of the suggested fix is resonant. Preparation, model construction, testing, and evaluation are all parts of the implementation process before prototyping. Each implementation component's specifics are explained in the paragraphs that follow.

5.1 Implementation and Experimentation Environment

Following the completion of the aforementioned tasks, the next step is to start the experiment with the chosen algorithm. In the development of this research, Python version 3.10.12 is utilized for algorithm training and designing a model. Python is a simple programming language to learn. It also stands out because it is effective with high-level data structures. In this research, we utilized Google Colab, a free platform offering GPU and TPU computing resources provided by Google. Google Colab provides access to Nvidia GPU with 15 GB of RAM. Additionally, we employed a desktop computer system equipped with an Intel(R) Core(TM) i5-6500 CPU running at 3.20GHz, 8 GB RAM, and a 64-bit operating system.

5.2 Dataset Description

We obtained the patient history information from electronically recorded patient history in order to build the dataset for this study. First, the researcher analyzed various materials about the hospital that specializes in treating thyroid disease and evaluated various information regarding thyroid disease in Ethiopia. Then TASH and SPMMC were selected to collect the data to prepare the dataset used for this study. All patient history information was not gathered due to a lack of time and complete information. As a result of this procedure, we obtain a dataset comprising a total of 5767 patient records categorized into negative, hyperthyroid, and hypothyroid classes. Initially, the dataset contained missing values, but through the application of data preprocessing techniques to address these gaps, the final dataset now consists of 5767 rows and 29 columns. Among these columns, one represents the target class, which is labeled as negative, hyperthyroid, or hypothyroid. The detailed procedure for constructing the dataset was covered in chapters three and four. Furthermore, chapter six provides information about result labeling and the size of each class

within the dataset.

Table 5-1 Description of the thyroid disorder dataset

No	Attribute Name	Feature full name	Data type	Description
1	Age	Age	numeric	The Patient's Age
2	Sex	Sex	bool	Sex Patient Identifies
3	Query on thyroxine	Query on thyroxine	bool	Is the patient currently taking thyroxine?
4	On_thyroxine	On_thyroxine	bool	Is the patient currently taking thyroxine?
5	Pregnant	Pregnant	bool	Whether the Patient is Pregnant
6	Sick	Sick	bool	Whether the Patient is sick
7	Thyroid_surgery	Thyroid_surgery	bool	Whether thyroid surgery was performed on the patient
8	On antithyroid meds	On antithyroid meds	bool	Is the patient currently prescribed antithyroid medication?
9	I131_treatment	I131_treatment	bool	Whether thyroid I131_treatment was performed on the patient
10	Query_hypothyroid	Query_hypothyroid	bool	Whether the patient thinks they have hypothyroidism
11	Query_hyperthyroid	Query_hyperthyroid	bool	Whether the patient thinks they have hyperthyroid
12	Goiter	Goiter	bool	The Presence of a Goiter in the Patient
13	Lithium	Lithium	bool	The Presence of a lithium in the Patient

14	Hypopituitary	Hypopituitary	bool	The Presence of a Hypopituitary in the Patient
15	Tumor	Tumor	bool	The Presence of a Tumor in the Patient
16	Psych	Psychological	bool	Whether Patient has Psych
17	T3_measured	Triiodothyronine measured	bool	T3 levels in the blood were measured
18	T3	Triiodothyronine	numeric	T3 Level in blood from lab work
19	TSH_Measured	Thyroid Stimulating Hormone Measured	bool	TSH levels in the blood were measured
20	TSH	Thyroid Stimulating Hormone	numeric	TSH Level in blood from lab work
21	TT4_measured	Total thyroxine measured	bool	TT4 levels in the blood were measured
22	TT4	Total thyroxine	numeric	TT4 Level in blood from lab work
23	FTI_measured	Free thyroxine index measured	bool	FTI levels in the blood were measured
24	FTI	Free thyroxine index	numeric	FTI Level in blood from lab work
25	T4U_measured	Thyroxine utilization measured	bool	T4U levels in the blood were measured
26	T4U	Thyroxine utilization	numeric	T4U Level in blood from lab work
27	TBG_Measured	Thyroxine-binding globulin measured	bool	TBG levels in the blood were measured
28	TBG	Thyroxine-binding globulin	numeric	The measurement of TBG levels in a blood sample obtained during laboratory

				tests.
29	target	target	string	

5.3 Preprocessing Implementation

In this study, thyroid disease dataset preprocessing is carried out using the Python programming language. Handling categorical and missing values are two preprocessing activities that are implemented. All non-numerical information must be converted to a numerical value. The study used the Label Encoder and replace method to transform the nominal and non-numerical values in addition to the mean strategy to fill in the missing values. Initially, when the dataset is transformed into a pandas data frame, it substitutes all empty values with NaN.

```
# Import the necessary libraries and package used for modeling and evaluation TD prediction model
import pandas as pd
import numpy as np
TDP = pd.read_csv('/content/drive/MyDrive/TDP16.csv')
```

Figure 5-1 Using the Panda library to implement loading a dataset

5.3.1 Handling Missing Values Implementations

To handle missing values in the dataset, we employed the fillna() method. This method replaces the missing values by calculating either the mean or mode values. Specifically, we used the Python SimpleImputer module with the mean strategy for filling in these missing values.

```
imputer = SimpleImputer(strategy='mean')
imputer.fit(TDP[['TSH']])
TDP['TSH'] = imputer.transform(TDP[['TSH']])
imputer = SimpleImputer(strategy='mean')
imputer.fit(TDP[['TT4']])
TDP['TT4'] = imputer.transform(TDP[['TT4']])
imputer = SimpleImputer(strategy='mean')
imputer.fit(TDP[['FTI']])
TDP['FTI'] = imputer.transform(TDP[['FTI']])
```

Figure 5-2 Implementation of Handling missing values

5.3.2 Implementation of Handling Categorical Values

In order to construct a precise and fitting dataset, it is crucial to convert non-numeric data into numerical form, ensuring an accurate model. This is how we address and incorporate categorical

values, as detailed in chapter four. The following are the resulting implementations.

```
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
label_encoder = LabelEncoder()
for column in categorical_columns:
    TDP[column] = label_encoder.fit_transform(TDP[column])
```

Figure 5-3 Handling categorical data

Upon cleansing the dataset, the attributes divide into distinct independent and dependent variables. The part of the data that has not undergone any specific feature selection is referred to as the X training dataset, while the labels indicating the class or category of the dataset are designated as Y. Subsequently, the classifiers are trained on the complete dataset, utilizing both X and Y. To initiate the training process for each classifier, we utilized the fit() method with the corresponding parameter settings, instantiating the respective class..

```
# Define the number of folds
n_splits = 10

# Initialize a StratifiedKFold cross-validator
cv = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)
confusion_matrices = []

# Initialize lists to store evaluation metrics for each fold
accuracy_scores = []
precision_scores = []
recall_scores = []
f1_scores = []

# Create a SMOTE instance
smote = SMOTE(random_state=42)
X_resampled, y_resampled = smote.fit_resample(X, y)
# Iterate through each fold
for train_idx, test_idx in cv.split(X_resampled, y_resampled):
    X_train, X_test = X_resampled.iloc[train_idx], X_resampled.iloc[test_idx]
    y_train, y_test = y_resampled.iloc[train_idx], y_resampled.iloc[test_idx]
```

Figure 5-4 Splitting Dataset into Dependent and Independent Features

5.4 Implementation of Feature Selection

Feature Selection is a sophisticated approach that identifies the most optimal set of features from the initial ones, boosting the effectiveness of machine learning algorithms. This technique is

implemented following a comprehensive exploration of the feature selection in chapter four. In this study, Machine Learning-based Feature Selection (MLFS) is a feature selection method that utilizes machine learning models to evaluate and select the most important features from a dataset. When used in conjunction with cross-validation, it becomes a powerful tool for feature selection while ensuring the model's robustness and generalizability.

```
        param_dist = {
  'max_depth': np.arange(3, 11),
  'learning_rate': np.linspace(0.01, 0.9),
  'n_estimators': np.arange(100, 1001, 100),
  'gamma': np.linspace(0, 5, 100),
  'reg_lambda': np.linspace(0, 2, 100),
}
  # Forward feature selection
  forward_sfs = SequentialFeatureSelector(estimator=xgb.XGBClassifier(),  # Create a dummy XGBoost model for feature selection
                                          direction='forward',
                                          scoring='accuracy',
                                          cv=cv)
  forward_sfs.fit(X_train, y_train)
  selected_features_forward = forward_sfs.transform(X_train)

  # Backward feature selection
  backward_sfs = SequentialFeatureSelector(estimator=xgb.XGBClassifier(),  # Create a dummy XGBoost model for feature selectio
                                           direction='backward',
                                           scoring='accuracy',
                                           cv=cv)
  backward_sfs.fit(X_train, y_train)
  selected_features_backward = backward_sfs.transform(X_train)

  # Take the intersection of selected features from forward and backward
  selected_features = list(set(selected_features_forward.columns).intersection(selected_features_backward.columns))

  # Define the XGBoost model
  xgb_clf = xgb.XGBClassifier(objective='multi:softmax',
                              num_class=3,
                              missing=1,
                              eval_metric=['merror', 'mlogloss'],
                              seed=42)

  # Initialize RandomizedSearchCV object
  random_search = RandomizedSearchCV(
      xgb_clf,
      param_distributions=param_dist,
      n_iter=50,  # Adjust the number of iterations as needed
      scoring='accuracy',
      cv=cv,
      verbose=2,
      n_jobs=-1,
      random_state=42
  )
```

Figure 5-5 MLFS with cross-validation implementation

5.5 Machine Learning Models Implementation

The suggested approach was applied to build machine-learning models using the sci-kit learning library package in Python. We took the conventional course of importing the necessary modeling library package and metrics for model evaluation.

```
# Import the necessary libraries and package used for modeling and evaluation TD prediction model
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import xgboost as xgb
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay, classification_report
from sklearn.metrics import balanced_accuracy_score, accuracy_score, precision_score, recall_score, f1_score
from sklearn.utils.class_weight import compute_sample_weight
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder, LabelEncoder
import itertools
import joblib
import sys
sys.modules['sklearn.externals.joblib'] = joblib
from mlxtend.feature_selection import SequentialFeatureSelector
%matplotlib inline
```

Figure 5-6 Importing the libraries necessary for the Modeling

5.5.1 Hyperparameter Tuning Implementation

In Chapter 4, Section 4.5.1 of this study, the Randomized Search method is employed for the purpose of hyperparameter tuning on the land suitability dataset. This technique helps automatically select the best hyperparameters to optimize the performance of the machine learning models used in this study, including RF, LR, ADA, SVM, and XGB.

XGBoost is a highly advanced machine learning algorithm that excels at making accurate prediction for various tasks. Its effectiveness comes from its combination of boosting, regularization and the careful management of different hyperparameters and it is implemented by the xgbclassifier() method of the sklearn of the ensemble package which is utilized to train and fit the model with optimized parameters.

```
# Define the XGBoost model
xgb_clf = xgb.XGBClassifier(
    objective='multi:softmax',
    num_class=3,
    missing=1,
    eval_metric=['merror', 'mlogloss'],
    seed=42
)

# Initialize RandomizedSearchCV object
param_dist = {
    'max_depth': np.arange(3, 10),  # Maximum depth of trees
    'learning_rate': np.linspace(0.01, 0.9),  # Learning rate
    'n_estimators': np.arange(100, 1001, 50),  # Number of trees
    'gamma': np.linspace(0, 5, 100),  # Minimum loss reduction required to make a further partition on a leaf node
    'reg_lambda': np.linspace(0, 2, 100)  # L2 regularization term on weights
}

random_search_xgb = RandomizedSearchCV(
    xgb_clf,
    param_distributions=param_dist,
    n_iter=100,
    scoring='accuracy',
    cv=cv,
    verbose=2,
    n_jobs=-1,
    random_state=42
)
```

Figure 5-7 python code for constructing and fit XGBoost model

Another classifier is Random Forest. RF model is created using the RandomForestClassifier() function from the sklearn ensemble package, and it serves as the classifier for our training process. This classifier is responsible for fitting numerous decision tree classifiers, with the number of trees determined by the n_estimators parameter.

```
# Define the parameter grid for random search
param_dist = {
     'max_depth': np.arange(3, 11),
     'n_estimators': np.arange(100, 1001, 100),
     'min_samples_split': np.arange(2, 11),
     'min_samples_leaf': np.arange(1, 11),
     'max_features': ['auto', 'sqrt', 'log2', None]
}

# Define the Random Forest model
rf_clf = RandomForestClassifier(n_jobs=-1, random_state=42)

# Initialize RandomizedSearchCV object
random_search_rf = RandomizedSearchCV(
     rf_clf,
     param_distributions=param_dist,
     n_iter=50,
     scoring='accuracy',
     cv=cv,
     verbose=2,
     n_jobs=-1,
     random_state=42
)
```

Figure 5-8 Code for constructing and fit RF model

The third classifier is Support Vector Machine. The SVM model in the study was created using the SVC () classifier provided by the sklearn package and it initializes a process for hyperparameter tuning of the SVM classifier using RandomizedSearchCV. It aims to find the best values for ‘C’ , ‘gamma’, and ‘class weight’ to improve the models overall performance.

```
# Define the SVM model
svm_clf = SVC(kernel='rbf', random_state=42)

# Initialize RandomizedSearchCV object
param_dist = {
    'C': np.logspace(-3, 3, 100),  # Regularization parameter
    'gamma': np.logspace(-3, 3, 100),  # Kernel coefficient
    'class_weight': ['balanced', None]
}

random_search_svm = RandomizedSearchCV(
    svm_clf,
    param_distributions=param_dist,
    n_iter=50,
    scoring='accuracy',
    cv=cv,
    verbose=2,
    n_jobs=-1,
    random_state=42
)
```

Figure 5-9 Code for constructing and fit SVM

The fourth classifier is Logistic Regression. This study employed the LogisticRegression() classifier from the sklearn library to construct the logistic regression model. It then defines a dictionary called param_dist, which contains various hyperparameter settings for the Logistic Regression model, such as regularization strength (C), penalty types ('l1', 'l2', 'elasticnet', 'none'), and solver algorithms ('newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga').

```
# Define the Logistic Regression model
lr_clf = LogisticRegression(max_iter=100, random_state=42)

# Initialize RandomizedSearchCV object
param_dist = {
    'C': np.logspace(-3, 3, 100),  # Regularization parameter
    'penalty': ['l1', 'l2', 'elasticnet', 'none'],
    'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga']
}

random_search_lr = RandomizedSearchCV(
    lr_clf,
    param_distributions=param_dist,
    n_iter=50,
    scoring='accuracy',
    cv=cv,
    verbose=2,
    n_jobs=-1,
    random_state=42
)
```

Figure 5-10 code for constructing and fit LR

The finally classifier is ADA. The provided code initiates an AdaBoost classifier denoted as ada_clf and proceeds to perform hyperparameter optimization using the RandomizedSearchCV technique. AdaBoost is a type of ensemble learning method, and the specific hyperparameters under consideration are n_estimators (representing the count of base learners) and learning_rate (which indicates the influence of individual base learners).

```
# Define the AdaBoost model
ada_clf = AdaBoostClassifier(random_state=42)

# Initialize RandomizedSearchCV object
param_dist = {
    'n_estimators': np.arange(50, 1001, 50),  # Number of weak learners
    'learning_rate': np.linspace(0.01, 2, 100)  # Learning rate
}

random_search_ada = RandomizedSearchCV(
    ada_clf,
    param_distributions=param_dist,
    n_iter=50,
    scoring='accuracy',
    cv=cv,
    verbose=2,
    n_jobs=-1,
    random_state=42
)
```

Figure 5-11 Code for constructing and fit ADA

5.6 Model Testing and Evaluation

Assessing the effectiveness of machine learning algorithms is a crucial component of this study. We adopted a widely recognized approach known as k-fold cross-validation, with K set to 10. This

method evenly partitions the data into 10 folds, with 9 of them utilized for training and the remaining 1 for evaluation or testing. To perform this evaluation, we employed the cross_val_score() and mean_accuracy () functions from the sklearn library, both configured to use a K value of ten. These techniques employ both the models and the dataset to assess and validate the learning proficiency of each model.

```
# Calculate mean and standard deviation of evaluation metrics across folds
mean_accuracy = np.mean(accuracy_scores)
std_accuracy = np.std(accuracy_scores)

mean_precision = np.mean(precision_scores)
std_precision = np.std(precision_scores)

mean_recall = np.mean(recall_scores)
std_recall = np.std(recall_scores)

mean_f1 = np.mean(f1_scores)
std_f1 = np.std(f1_scores)

# Print the results
print("Cross-Validation Results:")
print("Mean Accuracy: {:.2f} +/- {:.2f}".format(mean_accuracy, std_accuracy))
print("Mean Precision: {:.2f} +/- {:.2f}".format(mean_precision, std_precision))
print("Mean Recall: {:.2f} +/- {:.2f}".format(mean_recall, std_recall))
print("Mean F1-score: {:.2f} +/- {:.2f}".format(mean_f1, std_f1))
```

Figure 5-12 code to Constructing for Model Evaluation

CHAPTER SIX

6 RESULTS, EVALUATION AND DISCUSSIONS

The primary objective of this chapter is to present comprehensive details about the experimental outcomes derived from this study. The chapter encompasses an overview of the utilized dataset and its distribution across different classes. It delves into the performance of the models, evaluation results, and examines how feature selection influenced their performance. Finally, we conclude by examining the outcomes of each experiment conducted in this research

6.1 Dataset and Class Distribution Results

In this study, we utilized a dataset consisting of 5767 rows and 29 columns, which includes the target class. This dataset was prepared after preprocessing the raw data collected from TASH and SPMMC hospitals. Among these, 1625 patients were classified as having hypothyroidism, 492 patients had hyperthyroidism, and 3650 patients were categorized as having no thyroid disease or being negative. Class distribution in three class datasets is shown in detail in figure 6.1 below

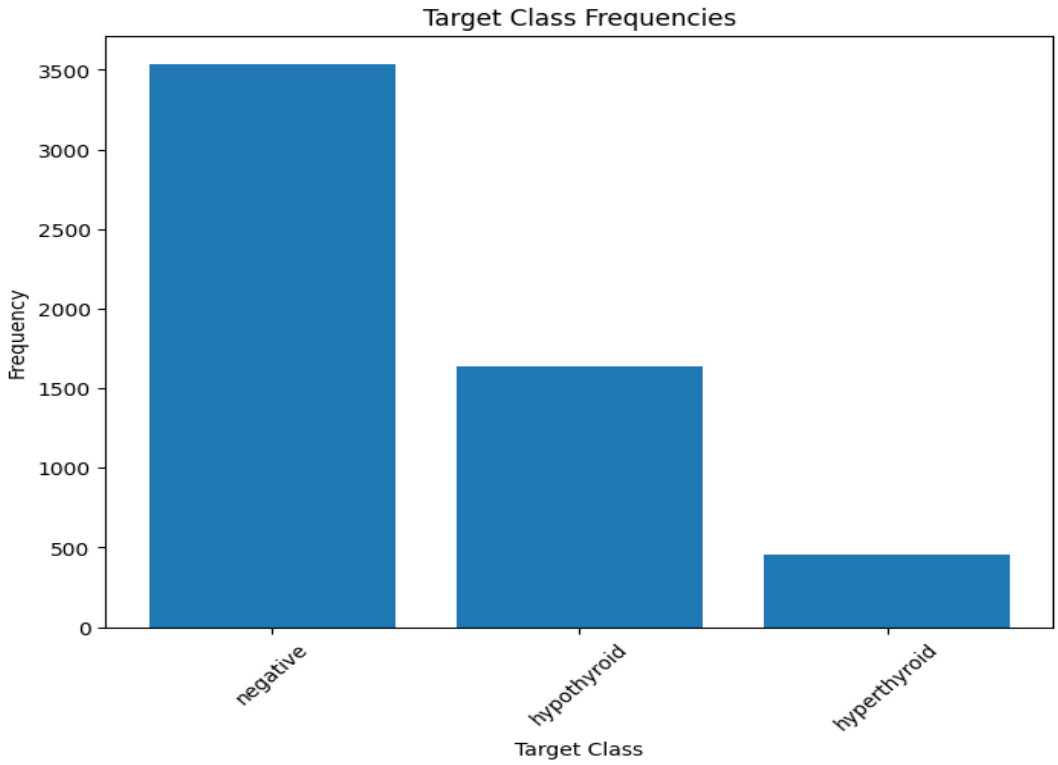


Figure 6-1 Class Distribution for three class

To address the imbalance in the multi-class dataset, we employed a data resampling technique.

we used the SMOTE data resampling method to balance the representation of the minority class with that of the majority class. After balanced the total dataset size is 10620.

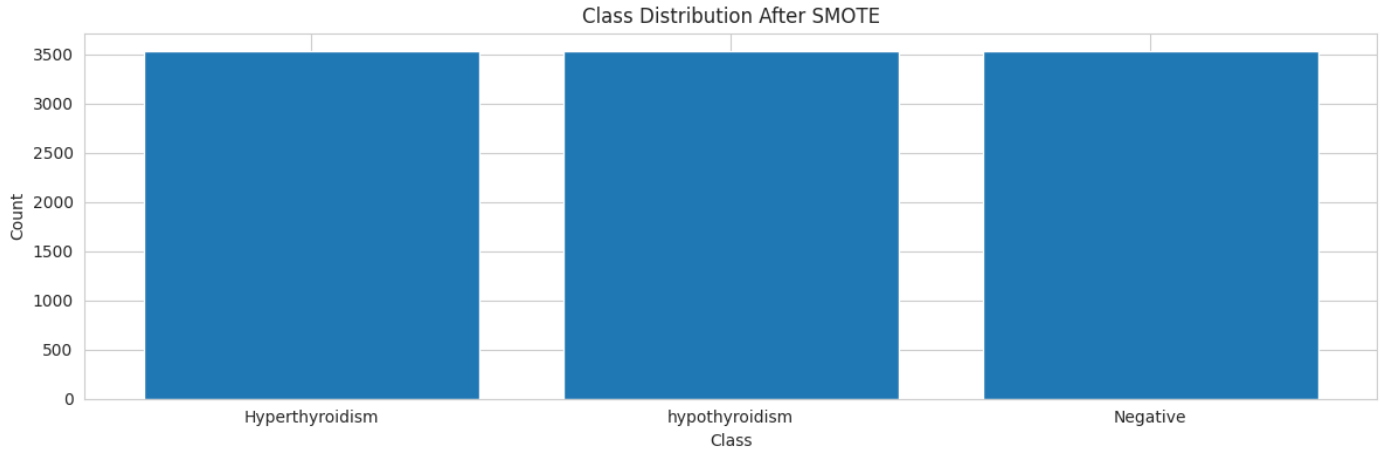


Figure 6-2 Class Distribution of three class dataset after balancing

6.2 Feature Importance and Selection Result

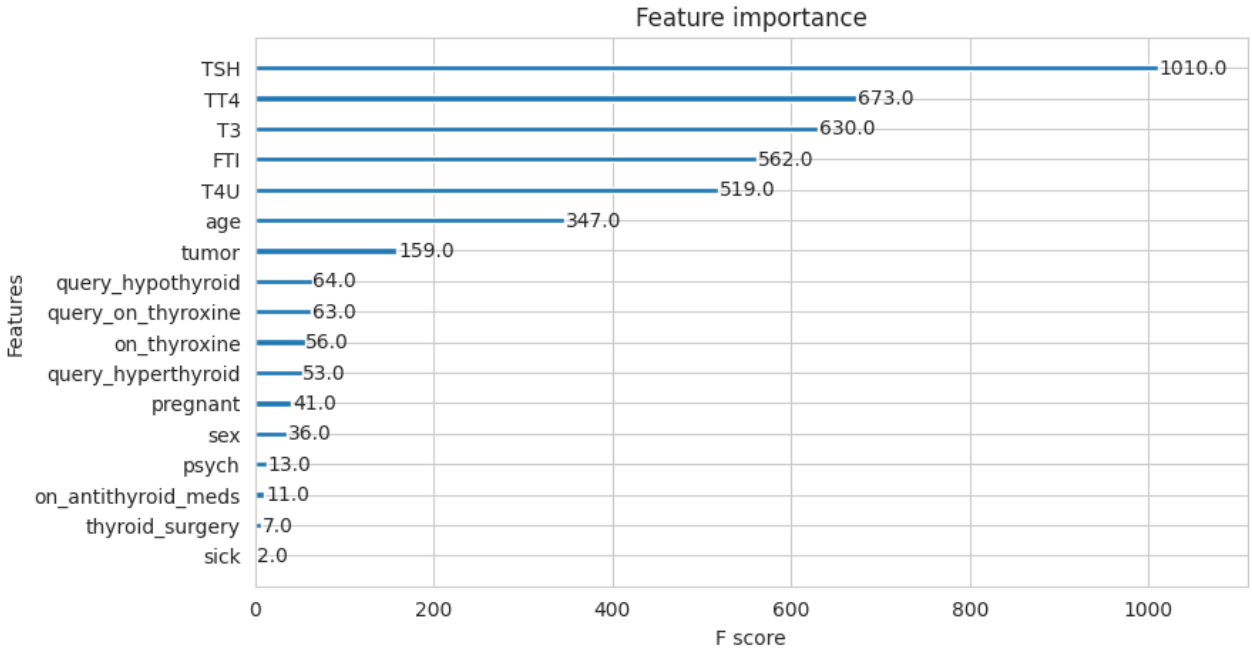


Figure 6-3 Feature Importance using XGBoost

Figure 6.3 indicates the ranking of feature based on their relevance score based on the xgboost classifier algorithm. Feature importance is a crucial factor in determining the impact of features on achieving the desired model accuracy. Features are ranked based on their significance, with

those having higher performance scores being more vital than those with lower scores. This assessment involves comparing the importance scores of different features within the dataset.

This study incorporates feature selection techniques to choose the most suitable set of features, enhancing the predictive capabilities of machine learning classifiers. Such as forward FS, backward FE, bi-directional FE and machine learning feature selection, where used in the training of the RF, LR, SVM, ADA and XGBoost used in the training dataset. The below table shows the feature selection result for each selected algorithm.

Table 6-1 The outcome of feature selection for each of the chosen algorithms

FS Methods	Selected ML Models	Selected Features	Original Features
FFS	RF	15	22
	LR	13	22
	SVM	12	22
	ADA	13	22
	XGBoost	16	22
BFE	RF	13	22
	LR	12	22
	SVM	11	22
	ADA	13	22
	XGBoost	14	22
BiDFE	RF	17	22
	LR	13	22
	SVM	16	22
	ADA	13	22
	XGBoost	18	22
MLFS	RF	17	22
	LR	14	22
	SVM	13	22
	ADA	16	22
	XGBoost	18	22

Table 6.1 shown above indicates the number of features selected wrapper methods with machine

learning models different from the original features. The section below delves into the comparison of model performance when utilizing the feature selection method on the chosen set of features, as well as when not using it.

6.3 Model Building

Five unique models- RF, LR, SVM, ADA and XGBoost are presented in this study. The dataset must first be divided into independent features and dependent features as the first stage. Independent features are characterized by values that are not influenced by other variables, while the dependent feature refers to the target or class, where its value is contingent upon the independent features. Then the models are built on multiclass datasets with and without using feature selection techniques.

6.4 Hyper-Parameters Tuning Results

Although the use of automatic hyper-parameter tuning is computationally expensive and requires huge memory resources, this study employs a random search approach for hyper-parameter tuning coupled with cross-validation, as explained in chapter four in section 4.3.1, instead of manual tuning of hyper-parameter, which is probably an inefficient strategy leading result in poor and very tedious performance. This study explores both grid search and cross-validation random search methods for hyperparameter tuning, and the result revealed that cv random search performs better than grid search, and the suggested parameters are shown in the table below.

Table 6-2 parameters used in Machine learning models as returned by randomized search CV

class	Hyper-Parameter tuning
RF	n_estimator=100, max_depth=5, max_features=’auto’
LR	solver=liblinear, C=1.0, Penalty=’I1’
SVM	Kernel=’rbf’, C=1.0, Class= ‘balanced’
ADA	n_estimators=100, learning_rat=0.5
XGB	max_depth=5, n_estimators=100, learning_rat=0.5, gamma=5

6.5 Models Performance and Evaluation Result

The experiment involved constructing a Machine Learning classifier algorithm that five distinct models: LR, ADA, SVM, RF, and XGBoost, integrated with MLFS with CV based on a XGBoost classifier. The assessment of these models' effectiveness employed a Stratified 10-fold cross-validation methodology, supported by performance evaluation metrics discussed in the preceding chapters (chapters three and four). In the context of Stratified 10-fold cross-validation, datasets were chosen in a manner that maintains proportional representation across different categories. The training of models encompassed utilizing nine out of ten folds, while the remaining fold was used for testing in an iterative fashion for each fold. This investigation deliberates on the performance of the chosen models, comparing outcomes with and without feature selection, with the goal of pinpointing the most proficient model based on the selected features.

6.5.1 Performance and Evaluation Results of Models with Original Features

The performance evaluation result of RF, LR, SVM, ADA and XGBoost models based on origianal features of the thyroid disorder dataset without applying any feature selection methods are discussed. After balancing the dataset, the total size used for training and testing the model was 10,620.

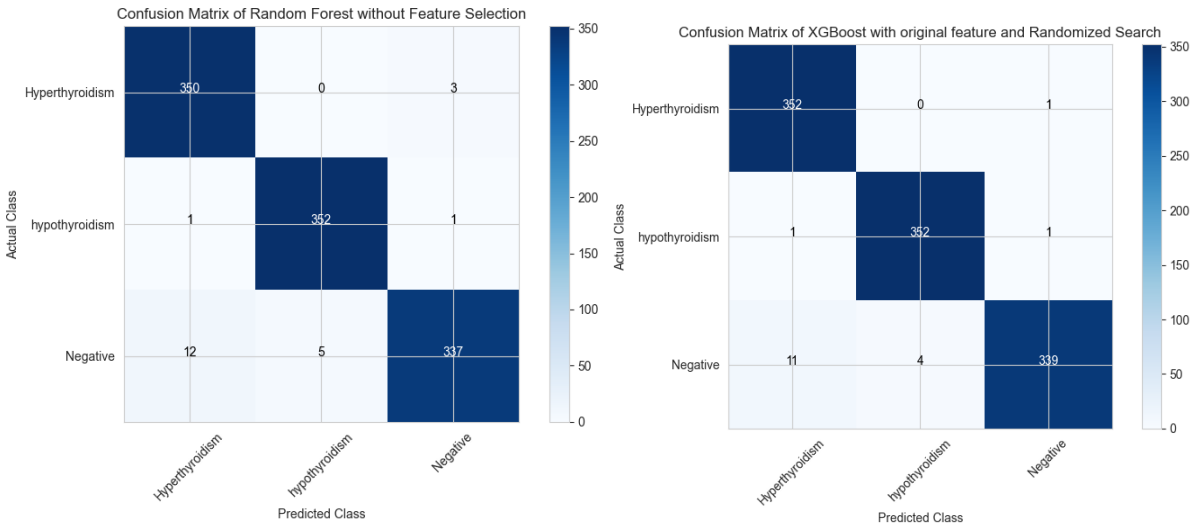


Figure 6-4 Confusion Matrix for RF (left) and XGBoost(right)

Figure 6.4 provides insights into the performance of the RF and XGBoost algorithms on the thyroid disorder dataset using the confusion matrix. Specifically, the results indicate that 350 out of 354, 352 out of 354, 337 out of 354 and 352 out of 354, 352 out of 354, 330 out of 354 dataset instances are correctly classified as hyperthyroidism, hypothyroidism and Negative respectively. Where only the algorithm misclassifies 3 instances of hyperthyroidism as negative, 1,1 instances of hypothyroidism as hyperthyroidism, negative, 12,5 instance of negative as hyperthyroidism, hypothyroidism and 1 instance of hyperthyroidism as negative, 1,1 instance of hypothyroidism as Hyperthyroidism, Negative, 11,4 instance of Negative as hyperthyroidism, hypothyroidism respectively.

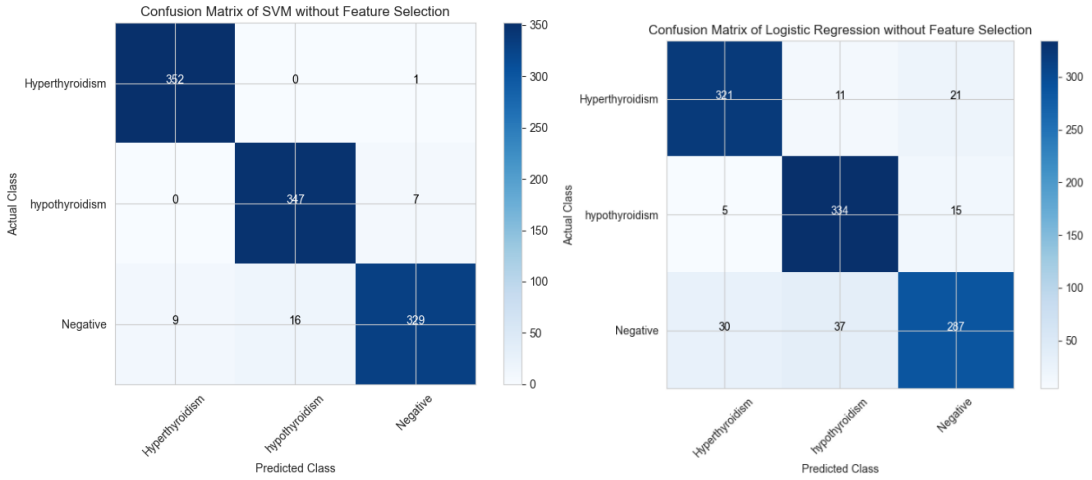


Figure 6-5 Confusion Matrix for SVM (left) and LR (right)

Figure 6.5 illustrates the performance of the SVM and LR algorithms on the thyroid disorder dataset, as evaluated using the confusion matrix. The results show that 352 out of 354, 347 out of 354, 329 out of 354 and 321 out of 354, 334 out of 354, 287 out of 354 dataset instances are correctly classified as hyperthyroidism, hypothyroidism and Negative respectively. Where only the algorithm misclassifies 1 instances of hyperthyroidism as negative, 7 instances of hypothyroidism as negative, 9,16 instance of negative as hyperthyroidism, hypothyroidism and 11,12 instance of hyperthyroidism as Hypothyroidism, negative, 5,15 instance of hypothyroidism as Hyperthyroidism, Negative, 30,37 instance of Negative as hyperthyroidism, hypothyroidism respectively.

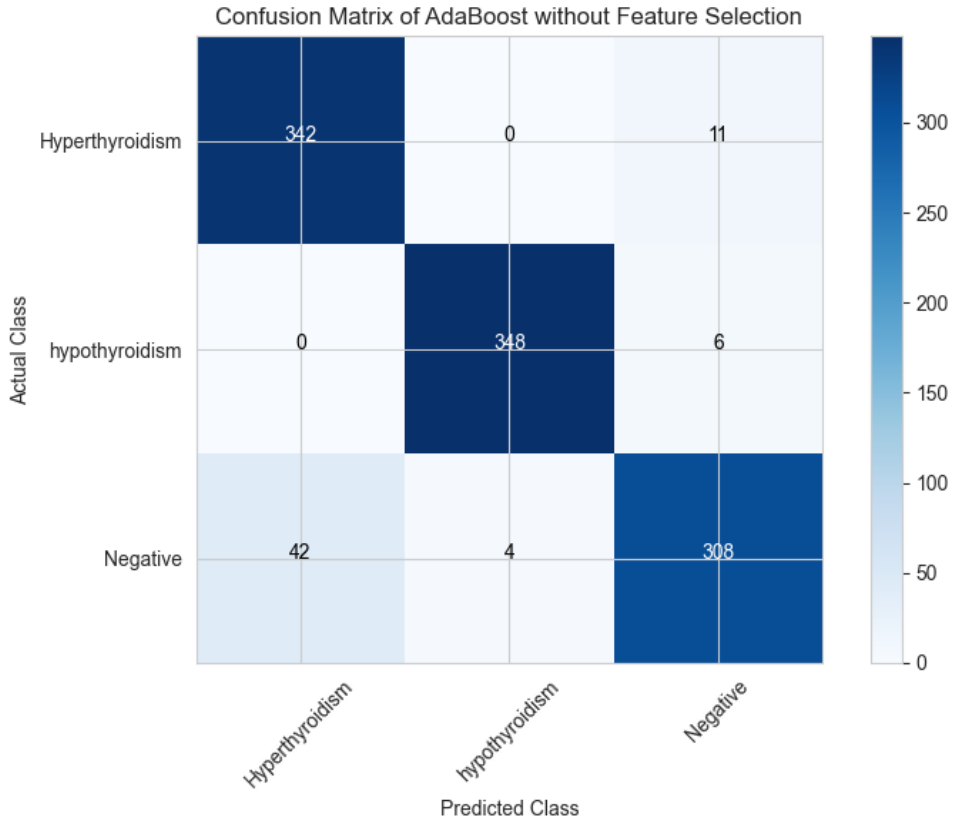


Figure 6-6 Confusion Matrix for ADA

Figure 6.6 illustrates the performance of the ADA (AdaBoost) algorithm on the thyroid disorder dataset, as evaluated using the confusion matrix. The results show that 342 out of 354, 348 out of 354, 308 out of 354 dataset instances are correctly classified as hyperthyroidism, hypothyroidism and Negative respectively. Where only the algorithm misclassifies 11 instances of hyperthyroidism as negative, 6 instances of hypothyroidism as negative, 42,4 instance of negative as hyperthyroidism, hypothyroidism respectively.

Table 6-3 Machine learning model results when using the original features.

Stratified 10-fold accuracy result (%) for RF, SVM, LR, ADA and XGBoost											
Model	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc (%)
RF	0.979	0.964	0.951	0.961	0.980	0.964	0.965	0.977	0.98	0.97	0.97
SVM	0.87	0.86	0.854	0.87	0.88	0.863	0.88	0.850	0.86	0.872	0.87

LR	0.856	0.86	0.854	0.87	0.88	0.863	0.88	0.850	0.86	0.872	0.87
ADA	0.925	0.915	0.93	0.915	0.9	0.94	0.9	0.91	0.921	0.905	0.92
XGBoost	0.979	0.964	0.981	0.991	0.980	0.974	0.982	0.984	0.981	0.983	0.98

The table displayed above illustrates the accuracy achieved in each of the ten folds during stratified 10-fold cross-validation for RF, LR, ADA, SVM, and XGBoost Models. It is evident that XGBoost outperformed the others with the highest accuracy score of 0.98, surpassing RF, LR, ADA, and SVM. Additionally, the table below presents performance metrics for these five algorithms.

Table 6-4 performance metrics for RF, SVM, LR, ADA and XGB mdels with originals features

Stratified 10-fold CV score of Performance Evaluation Metrics				
Model	Precision (%)	Recall (%)	F1-score (%)	Acc (%)
RF	0.97	0.971	0.974	0.97
SVM	0.87	0.86	0.864	0.87
LR	0.85	0.86	0.854	0.87
ADA	0.925	0.915	0.92	0.92
XGBoost	0.98	0.982	0.984	0.98

Table 6.4 illustrates the scores of performance metrics used in this study for each model. The Machine Learning models outshine the other models across all selected metrics in this study.

6.5.2 Performance and Evaluation Result of the Models with Feature selection

Based on the XGBoost model MLFS with cross-validation was used to identify the most effective feature subset for improving the predictive performance of machine learning models. Hence, the evaluation of RF, LR, SVM, ADA, and XGBoost models in relation to their confusion matrices and various performance metrics is conducted using stratified 10-fold cross-validation. This evaluation is performed with the feature selection methods described below, and the model's performance is compared to the one using the original features

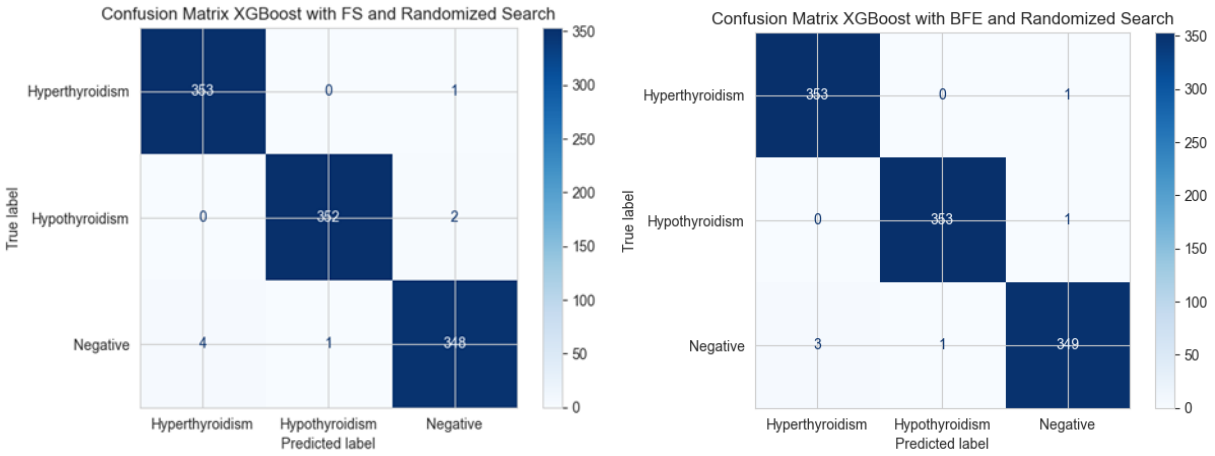


Figure 6-7 Confusion Matrix for XGBoost with FFS and BFE

As can be seen from figure 6.7 above, the performance of the XGBoost with FFS and BFE model with MLFS with cross-validation examined on thyroid disorder dataset using confusion matrix, as such 353 out of 354, 352 out of 354, 348 out of 354 and 353 out of 354, 353 out of 354, 349 out of 354 dataset instances are correctly classified as hyperthyroidism, hypothyroidism and Negative respectively. Where only the algorithm misclassifies 1 instances of hyperthyroidism as negative,2 instances of hypothyroidism as negative, 4,1 instance of negative as hyperthyroidism, hypothyroidism and 1 instance of hyperthyroidism as negative, 1 instance of hypothyroidism as Negative, 3,1 instance of Negative as hyperthyroidism, hypothyroidism respectively.

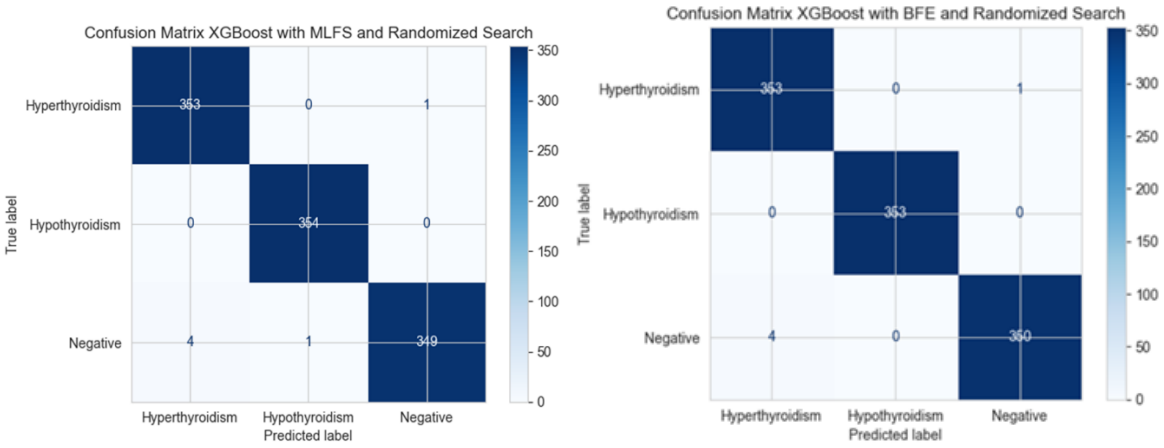


Figure 6-8 confusion matrix of XGBoosting with MLFS

As can be seen from figure 6.9 above, the performance of the XGBoost with MLFS and BFE model with MLFS with cross-validation examined on thyroid disorder dataset using confusion matrix, as such 353 out of 354, 354 out of 354, 349 out of 354 and 353 out of 354, 353 out of 354, 350 out of 354 dataset instances are correctly classified as hyperthyroidism, hypothyroidism and Negative respectively. Where only the algorithm misclassifies 1 instances of hyperthyroidism as negative, 4,1 instance of negative as hyperthyroidism, hypothyroidism and 1 instance of hyperthyroidism as negative, 4 instance of Negative as hyperthyroidism respectively.

The table below presents the outcomes of performance evaluation metrics for RF, SVM, LR, ADA and XGBoost models. These models were evaluated using the MLFS with cross-validation technique, and the evaluation was carried out through stratified 10-fold cross-validation.

Table 6-5 RF, SVM, LR, ADA and XGBoost models accuracy result with FS

Stratified 10-fold accuracy result (%) for RF, SVM, LR, ADA and XGBoost											
Model	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Acc (%)
RF	0.989	0.984	0.981	0.991	0.980	0.984	0.985	0.987	0.985	0.983	0.98
SVM	0.97	0.96	0.964	0.97	0.98	0.963	0.98	0.970	0.97	0.972	0.97
LR	0.88	0.86	0.91	0.87	0.88	0.893	0.88	0.90	0.91	0.882	0.89
ADA	0.945	0.965	0.95	0.965	0.94	0.937	0.94	0.96	0.951	0.945	0.95
XGB	0.989	0.984	0.991	0.991	0.99	0.994	0.985	0.987	0.989	0.989	0.989

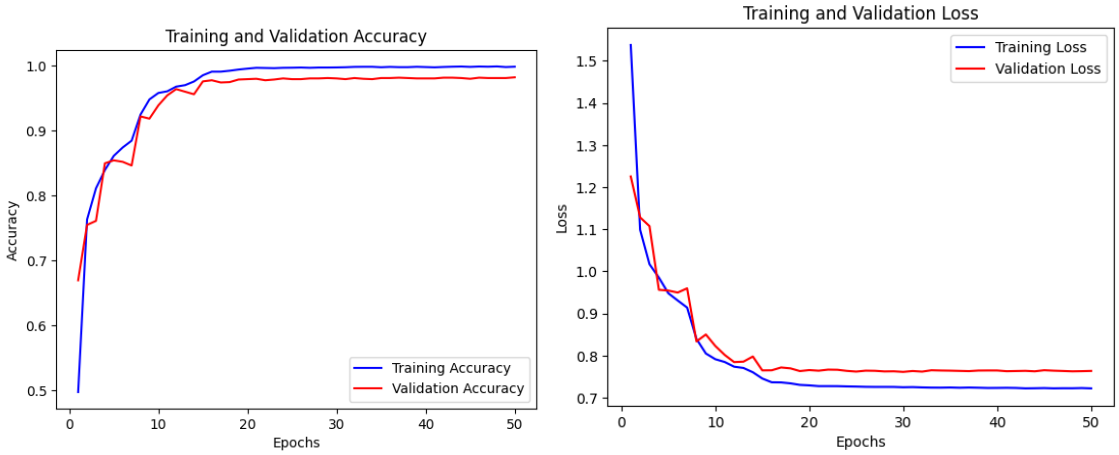


Figure 6-9 Training and Validation Curve of accuracy and loss for XGB model

The table below presents the average scores of various performance metrics, including precision, recall, and f1-score, for the RF, SVM, LR, ADA, and XGB models when utilizing MLFSCV.

Table 6-6 Performance Metrics for RF, SVM, LR, ADA and XGB with FSCV models

Stratified 10-fold CV score of Performance Evaluation Metrics				
Model	Precision (%)	Recall (%)	F1-score (%)	Acc (%)
RF	0.989	0.984	0.981	0.98
SVM	0.97	0.96	0.97	0.97
LR	0.887	0.889	0.892	0.89
ADA	0.96	0.95	0.95	0.95
XGBoost	0.987	0.984	0.991	0.989

6.6 Discussion

Predicting thyroid diseases has proven to be a complex task due to the difficult in identifying and assessing thyroid symptoms without the guidance of a medical professional. Consequently, developing solutions for classifying thyroid diseases becomes crucial, aiming to precisely determine the specific type of thyroid condition, whether it pertains to hypothyroidism or hyperthyroidism, provided that machine learning models are trained using an ample amount of

data samples and their performance is fine-tuned. This study's initial phase included two research topics. So, this part aims to provide answers to those queries.

- What are the determinant attributes for the Prediction of thyroid disorder?

In order to address this question, MLFS with cross-validation was employed to identify the key attributes that contribute significantly to predicting thyroid disorders. The results revealed that the following attributes play a crucial role in making accurate predictions for thyroid disorder. TSH, T3, TT4, T4U, FTI, age, query_hperthyroid, on_thyroxine, sex, tumor, query_on_thyroxine, on_antithyroid_meds, psych, query_hypothyroid, l131_treatment, lithium, pregnant and thyroid_surgery.

- Which machine learning algorithm is the best to construct the classifier model from the selected algorithm?

For this question, the selected machine learning algorithm are distinct and form a new model which is a ML for thyroid disorder classification. Furthermore, to address this question, the researcher opted for five suitable classification algorithms and various feature selection methods. The performance of these models was then compared both with and without feature selection using stratified 10-fold cross-validation.

Five selected models such as RF, SVM, LR, ADA and XGB compared with each other, and then the MLFS based on extra-tree classifier was developed. According to the experiment the XGB model outperforms with 99.1% F1-score and 98.9 accuracy as the best classifier models.

CHAPTER SEVEN

CONCLUSION, RECOMMENDATION AND FUTURE WORK

In this section, we provide an overview of the proposed approach for predicting thyroid disorders using machine learning techniques. We also present the conclusions drawn from this study and discuss recommendations and possible direction for future research.

7.1 Conclusion

Thyroid disorder is a significant global health issue, particularly affecting developing countries like Ethiopia. Early detection and prevention are crucial, and machine learning can play a pivotal role in improving diagnosis. However, current diagnostic methods often rely on binary classification, limited datasets, imbalanced data and lack proper validation. Data collection was followed by data preprocessing, a crucial step in ensuring that the dataset was clean and suitable for machine learning. We addressed issues such as noisy data, missing values, and categorical data transformation. Cleaning noisy data was essential to maintain data integrity, while handling missing values and converting categorical data into numerical format were key steps in preparing the data for modeling. This study aims to address these limitations by focusing on feature engineering and employing various machine learning algorithms to predict different types of thyroid disorders, including Negative, Hyperthyroidism, and Hypothyroidism. The study collects patient data from TASH and SPMMC hospitals in Ethiopia. It investigates feature selection techniques like FFS, BFE, BDFE, and FS using extra tree classifiers. The chosen algorithms for classification include Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), ADA, and XGBoost. Evaluation of these models is conducted using 10-fold cross-validation, assessing metrics like accuracy, precision, recall, and F1-score. Results suggest that the XGBoost algorithm with feature selection and cross-validation outperforms other models, achieving an accuracy of 98.9% and F1-score 99.1%.

Overall, the study aims to advance thyroid disorder prediction through machine learning, ultimately benefiting healthcare in Ethiopia.

7.2 Recommendation and Future work

The major goals of this work were achieved by using machine learning to the prediction of thyroid diseases. The research's findings have led to the following recommendations for further study by

researchers and healthcare facilities.

- Future research should involve extensive clinical testing and validation of the proposed machine learning models for thyroid disease prediction. Collaboration with healthcare institutions and practitioners is essential to ensure the reliability and clinical utility of the models.
- Efforts should be made to collect and utilize larger and more diverse datasets for thyroid disease prediction. This can enhance the models' generalizability and robustness, especially for rare thyroid disorders.
- Expand the classification to include a broader range of thyroid disorders, allowing for more precise diagnosis and treatment recommendations.
- Extend the model's capabilities to predict a broader range of thyroid disorder types, including subtypes and severity levels. This can provide more detailed diagnostic information to healthcare providers.

REFERENCES

Abbad Ur Rehman, H., Lin, C. Y., & Mushtaq, Z. (2021). Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. *Journal of the Chinese Institute of Engineers, Transactions of the Chinese Institute of Engineers,Series A*, *44*(1), 77–87. https://doi.org/10.1080/02533839.2020.1831967

Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., & Ahmad, A. (2022, June 7). Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. BioMed Research International, 2022, 1–10.

Anggraeni, A. N., Mustofa, K., & Priyanta, S. (2021, July 31). Comparison of Filter and Wrapper Based Feature Selection Methods on Spam Comment Classification. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 15(3), 245. https://doi.org/10.22146/ijccs.66965

Arun Kumar, R., Vijay Franklin, J., & Koppula, N. (2022). A Comprehensive Survey on Metaheuristic Algorithm for Feature Selection Techniques. Materials Today: Proceedings, 64, 435–441. https://doi.org/10.1016/j.matpr.2022.04.803

Atun, R. (2015, August). Transitioning health systems for multimorbidity. The Lancet, 386(9995), 721–722. https://doi.org/10.1016/s0140-6736(14)62254-6

BABY, D., DEVARAJ, S. J., HEMANTH, J., & M, A. R. M. (2021, October 4). Leukocyte classification based on feature selection using extra trees classifier: a transfer learning approach. TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES, 29(SI-1), 2742–2757. https://doi.org/10.3906/elk-2104-183

Brownlee, Jason. (2020). Classification Tasks in Machine Learning. pp. August 19, 2020, accessed date: Nov. 20, 2021, https://machinelearningmastery.com/types-of-classification-in-machine.

Chaganti, R., Rustam, F., de La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. *Cancers*, *14*(16). https://doi.org/10.3390/cancers14163914

Chapelle, O., Scholkopf, B., & Zien, Eds., A. (2009, March). Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. IEEE Transactions on Neural Networks, 20(3), 542–542. https://doi.org/10.1109/tnn.2009.2015974

Chen, C., Tsai, Y., Chang, F., & Lin, W. (2020, April 3). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems, 37(5). https://doi.org/10.1111/exsy.12553

Chen, D., Hu, J., Zhu, M., Tang, N., Yang, Y., & Feng, Y. (2020). Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest. *BioData Mining*, *13*(1). https://doi.org/10.1186/s13040-020-00223-w

Dahiwade, D., Patle, G., & Meshram, E. (2019). Designing Disease Prediction Model Using Machine Learning Approach. *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, 1211–1215. https://doi.org/10.1109/ICCMC.2019.8819782

Desalew, A. S. (2020). Cause and predictors of neonatal mortality among neonates admitted to neonatal intensive care units of public hospitals in eastern Ethiopia: a facility-based prospective follow-up study. BMC Pediatr.

Dhal, P., & Azad, C. (2021, July 23). A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence, 52(4), 4543–4581. https://doi.org/10.1007/s10489-021-02550-9

EPH,ICF. (2019). Ethiopia Mini Demographic and Health Survey 2019: Key Indicators. Rockville, Maryland, USA.

Elghazel, H., & Aussem, A. (2013, April 3). Unsupervised feature selection with ensemble learning. Machine Learning, 98(1–2), 157–180. https://doi.org/10.1007/s10994-013-5337-8

Farooqui, M. E., & Ahmad, D. J. (2020, July). DISEASE PREDICTION SYSTEM USING SUPPORT VECTOR MACHINE AND MULTILINEAR REGRESSION. International Journal of Innovative Research in Computer Science & Technology, 8(4). https://doi.org/10.21276/ijircst.2020.8.4.15

Fu, Y., Wu, Q., Liu, K., & Gao, H. (2022, August 30). Feature Selection Methods for Extreme Learning Machines. Axioms, 11(9), 444. https://doi.org/10.3390/axioms11090444

Garg, A., & Bansal, D. (2023, February 28). Application of Machine Learning in Disease Prediction. International Journal for Research in Applied Science and Engineering Technology, 11(2), 47–51. https://doi.org/10.22214/ijraset.2023.48954

Garber, J. R., Cobin, R. H., Gharib, H., Hennessey, J. V., Klein, I., Mechanick, J. I., et al. (2012). Clinical practice guidelines for hypothyroidism in adults: cosponsored by the American Association of Clinical Endocrinologists and the American Thyroid Association. Endocr Pract, 18(6), 988-1028.

Garcia De Lomana, M., Weber, A. G., Birk, B., Landsiedel, R., Achenbach, J., Schleifer, K. J., Mathea, M., & Kirchmair, J. (2021). In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis. *Chemical Research in Toxicology*, *34*(2), 396–411. https://doi.org/10.1021/acs.chemrestox.0c00304

Ha, J., Baek, H., Jeong, C., Yeo, M., Lee, S. H., Cho, J. H., Baek, K. H., Kang, M. I., & Lim, D. J. (2021, June 10). Heart Rate Variability in Postoperative Patients with Nonfunctioning Pituitary Adenoma. Endocrinology and Metabolism. https://doi.org/10.3803/enm.2021.978

He, Z., Li, L., Huang, Z., & Situ, H. (2018, May 15). Quantum-enhanced feature selection with forward selection and backward elimination. Quantum Information Processing, 17(7). https://doi.org/10.1007/s11128-018-1924-8

Heuck, C. C. (1993). The World Health Organization's role and future plans in laboratory standardization / C. C. Heuck. *Scandinavian Journal of Clinical Laboratory Investigation 1993 ; 212 Supplement : 3-7*. https://apps.who.int/iris/handle/10665/49724

Hosseinzadeh, M., Ahmed, O. H., Ghafour, M. Y., Safara, F., hama, H. kamaran, Ali, S., Vo, B., & Chiang, H. sen. (2021). A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. *Journal of Supercomputing*, *77*(4), 3616–3637. https://doi.org/10.1007/s11227-020-03404-w

Idarraga, A. J., Luong, G., Hsiao, V., & Schneider, D. F. (2021). False Negative Rates in Benign Thyroid Nodule Diagnosis: Machine Learning for Detecting Malignancy. *Journal of Surgical Research*, *268*, 562–569. https://doi.org/10.1016/j.jss.2021.06.076

Journy, N. M., Bernier, M. O., Doody, M. M., Alexander, B. H., Linet, M. S., & Kitahara, C. M. (2017, August). Hyperthyroidism, Hypothyroidism, and Cause-Specific Mortality in a Large Cohort of Women. Thyroid, 27(8), 1001–1010. https://doi.org/10.1089/thy.2017.0063

Kwon, M. R., Shin, J. H., Park, H., Cho, H., Hahn, S. Y., & Park, K. W. (2020). Radiomics study of thyroid ultrasound for predicting BRAF mutation in papillary thyroid carcinoma: Preliminary results. *American Journal of Neuroradiology*, *41*(4), 700–705. https://doi.org/10.3174/AJNR.A6505

Leng, L., Li, M., Kim, C., & Bi, X. (2017). Dual-source discrimination power analysis for multi-instance contactless palmprint recognition. *Multimedia Tools and Applications*, *76*(1), 333–354. https://doi.org/10.1007/s11042-015-3058-7

Liashchynskyi, O. (2019). Practical hyperparameter optimization: Grid search with random search. arXiv preprint arXiv:1902.06059.

Mishra, D., Buyya, R., Mohapatra, P., & Patnaik, S. (Eds.). (2020, August 29). Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2 (Vol. 153), 385-390. https://doi.org/10.1007/978-981-15-6202-0

Mohammed, A. (2020, December 1). Pattern and Clinical Profile of Thyroid Disorders among Patients Attending Endocrine Clinic of Tikur Anbessa Specialized Hospital. Pattern and Clinical Profile of Thyroid Disorders Among Patients Attending Endocrine Clinic of Tikur Anbessa Specialized Hospital. http://etd.aau.edu.et/handle/123456789/24885

Phyu, T. Z., & Oo, N. N. (2016). Performance Comparison of Feature Selection Methods. MATEC Web of Conferences, 42, 06002. https://doi.org/10.1051/matecconf/20164206002

Razia, S., Swathi Prathyusha, P., Krishna, N. V., & Sumana, N. S. (2018b). A Comparative study of machine learning algorithms on thyroid disease prediction. In *International Journal of Engineering & Technology* (Vol. 7, Issue 2).

Razia, S., Swathi Prathyusha, P., Krishna, N. V., & Sumana, N. S. (2018a). A Comparative study of machine learning algorithms on thyroid disease prediction. In International Journal of Engineering & Technology (Vol. 7, Issue 2).

Reta Demissie, W. (2019). Prevalence, Clinical Presentation and Patterns of Thyroid Disorders Among Anterior Neck Mass Patients Visiting Jimma Medical Center, Southwest Ethiopia. *Biomedical Journal of Scientific & Technical Research*, *18*(2). https://doi.org/10.26717/bjstr.2019.18.003126

Riajuliislam, M., Rahim, K. Z., & Mahmud, A. (2021). Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques. *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, 60–64. https://doi.org/10.1109/ICICT4SD50815.2021.9397052

Rugge JB, Bougatsos C, Chou R. Screening and treatment of thyroid dysfunction: an evidence review for the U.S. Preventive Services Task Force. Ann Intern Med. 2015;162(1):35–45

Salman, K., & Sonuc, E. (2021). Thyroid Disease Classification Using Machine Learning Algorithms. *Journal of Physics: Conference Series*, *1963*(1). https://doi.org/10.1088/1742-6596/1963/1/012140

Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & de Albuquerque, V. H. C. (2020). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. *Journal of Supercomputing*, *76*(2), 1128–1143. https://doi.org/10.1007/s11227-018-2469-4

Shukla, M. (2022, May 3). Research Methods in Machine Learning: A Content Analysis. INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 06(05). https://doi.org/10.55041/ijsrem12736

Singh, A. K., & Loscalzo, J. (2019). The Brigham Intensive Review of Internal Medicine (3rd edition)

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2019, January 29). A review of unsupervised feature selection methods. Artificial Intelligence Review, 53(2), 907–948. https://doi.org/10.1007/s10462-019-09682-y

Sousan, H. H. (2016). Nurses' Immigration: Causes and Problems,. International Journal of Medical Research & Health Sciences, vol. 5, no. 9S, pp. 486-491.

Specialty Imaging: Postoperative SpineRossJeffrey Philadelphia, PA: Wolters Kluwer Lippincott Williams & Wilkins, 2012. ISBN 978-1-931884-89-1. Hardcover, $249.00; pp 400. (2012, December). Radiology, 265(3), 694–694. https://doi.org/10.1148/radiol.12124043

Srivastava, Tavish. (2019). "Important Model Evaluation Metrics for Machine Learning Everyone should know. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-errormetrics/ accessed date:Nov 27,2021.

Suga, Y., & Abebe, E. (2020). Patterns of Surgically Treated Thyroid Disease: A Two Years Review at St. Paul Hospital Millennium medical Collage, Addis Ababa, Ethiopia. *Ethiopian Journal of Health Sciences*, *30*(1), 31–36. https://doi.org/10.4314/ejhs.v30i1.5

Vanderpump, M. P. J. (2011, September 1). The epidemiology of thyroid disease. British Medical Bulletin, 99(1), 39–51. https://doi.org/10.1093/bmb/ldr030

Widodo, S., Brawijaya, H., & Samudi, S. (2022, October 3). Stratified K-fold cross validation optimization on machine learning for prediction. Sinkron, 7(4), 2407–2414. https://doi.org/10.33395/sinkron.v7i4.11792

Yadav, N. K., Thanpari, C., Shrewastwa, M. K., Sathian, B., & Mittal, R. K. (2013, April). Socio demographic wise risk assessment of thyroid function abnormalities in far western region of Nepal: A hospital based descriptive study. Asian Pacific Journal of Tropical Disease, 3(2), 150–154. https://doi.org/10.1016/s2222-1808(13)60060-2

Witemeyer ---, S. (n.d.). *Thyroid Disorders: Hypothyroidism and Hyperthyroidism*.

APPENDIX

Appendix A: Dataset Description

Age: - The patient's age is essential in understanding the prevalence and management of thyroid diseases. Certain thyroid problems may be more likely to occur or be more of a risk depending on a person's age.

Sex: - The sex of the patient can influence the likelihood and characteristics of thyroid diseases. For example, autoimmune thyroid issues like Graves' disease and Hashimoto's thyroiditis are more common in women.

Query_on_thyroxine: - Is there anything uncertain about the patient taking their thyroxine medication?

On_thyroxine: - This feature specifies whether the patient is currently taking thyroxine medication, which is commonly used for thyroid hormone replacement therapy.

Pregnant: - Pregnancy can impact thyroid function, and thyroid disorders may arise or be affected during pregnancy. This feature indicates whether the patient is currently pregnant

Query_hypothyroid: - This feature suggests whether there is a suspicion or query regarding hypothyroidism, an underactive thyroid condition.

Query_hyperthyroid: - This feature indicates whether there is a suspicion or query regarding hyperthyroidism, an overactive thyroid condition.

TSH: - It is a hormone produced by the pituitary gland that regulates the thyroid gland's hormone production. TSH levels are commonly measured in thyroid function tests to evaluate thyroid function.

T3: - It is an active thyroid hormone. Measurement of T3 levels provides insights into thyroid function and potential abnormalities.

TT4: - represents the total amount of thyroxine hormone in the blood. TT4 measurement helps evaluate thyroid function and potential disorders.

T4U: - measures the binding capacity of thyroxine-binding globulin (TBG), a protein that transports thyroid hormones. It provides information about TBG availability and thyroid hormone binding.

FTI: - It is a calculated value that estimates the concentration of free thyroxine hormone in the blood. It combines measurements of TT4 and T4U to assess the biologically active form of T4.

Appendix A1: Sample Dataset

81	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	3.9	t	1.8	t	131	t	0.98	t
58	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.17	t	1.2	t	75	t	0.88	t
64	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.44	t	2.5	t	89	t	0.97	t
68	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	28	t	1.8	t	72	t	0.87	t
82	F	f	f	f	f	f	f	f	f	t	f	f	f	f	f	t	1.3	t	1.4	t	131	t	1.08	t
70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.5	t	0.8	t	99	t	0.79	t
62	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.86	t	1.9	t	78	t	0.85	t
62	M	f	f	f	t	f	f	f	f	f	f	f	f	f	f	t	2.2	t	2.8	t	122	t	1.13	t
72	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	9.7	t	1.1	t	77	t	0.79	t
67	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.1	t	2.3	t	114	t	1.1	t
24	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.6	t	1.3	t	67	t	0.88	t
24	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	3.2	t	2.8	t	91	f	f	
28	M	f	f	f	f	f	f	f	f	f	f	f	f	f	t	t	1.4	t	2.4	t	110	t	0.99	t
47	F	f	f	f	f	f	f	f	f	f	f	f	f	f	t	t	1.9	t	2.3	t	135	t	1.07	t
37	F	f	f	f	f	f	f	f	f	f	t	f	f	f	t	t	4.4	t	2.6	t	108	t	1.12	t
74	M	f	f	f	t	f	f	f	f	f	f	f	f	f	f	t	3.3	t	1.2	t	91	t	0.92	t
59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	7.1	t	2.1	t	79	t	0.96	t
69	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.6	t	1.9	t	90	t	0.99	t
41	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.03	t	2.1	t	187	t	0.89	t
41	M	f	f	f	f	f	f	f	t	f	f	f	f	f	f	t	0.015	t	2.5	t	22	f	f	
31	M	f	f	f	f	f	f	f	f	f	f	f	f	f	t	t	1.2	t	2.7	t	99	t	0.88	t
37	F	f	f	f	f	f	f	f	f	f	t	f	f	f	f	t	0.64	t	2.1	t	92	t	0.99	t
73	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.6	t	1.3	t	146	t	0.89	t
62	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.1	t	2.3	t	119	t	0.95	t
71	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.92	t	2.1	t	84	t	0.86	t
27	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.035	t	2.2	t	88	t	0.9	t
37	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.09	t	2.2	t	128	t	0.75	t
25	F	t	f	f	f	f	f	f	t	f	f	f	f	f	f	t	0.7	f		t	183	t	1.41	t

Appendix B: Sample Code

```
# Define the number of folds
n_splits = 10
# Initialize a StratifiedKFold cross-validator
cv = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)
confusion_matrices = []
# Define the parameter grid for random search
param_dist = {
    'max_depth': np.arange(3, 11),
    'learning_rate': np.linspace(0.01, 0.9),
    'n_estimators': np.arange(100, 1001, 100),
    'gamma': np.linspace(0, 5, 100),
    'reg_lambda': np.linspace(0, 2, 100),
}

# Create a SMOTE instance
smote = SMOTE(random_state=42)

# Resample the dataset
X_resampled, y_resampled = smote.fit_resample(X, y)

# Initialize Extra Trees Classifier for feature selection
extra_trees = ExtraTreesClassifier(n_estimators=100, random_state=42)

# Iterate through each fold
for train_idx, test_idx in cv.split(X_resampled, y_resampled):
    X_train, X_test = X_resampled.iloc[train_idx], X_resampled.iloc[test_idx]
    y_train, y_test = y_resampled.iloc[train_idx], y_resampled.iloc[test_idx]

    # Forward feature selection
    forward_sfs = SequentialFeatureSelector(estimator=xgb.XGBClassifier(),  # Create a dummy XGBoost model for feature selection
                                            direction='forward',
                                            scoring='accuracy',
                                            cv=cv)
    forward_sfs.fit(X_train, y_train)
    selected_features_forward = forward_sfs.transform(X_train)

    # Backward feature selection
    backward_sfs = SequentialFeatureSelector(estimator=xgb.XGBClassifier(),  # Create a dummy XGBoost model for feature selection
                                             direction='backward',
                                             scoring='accuracy',
                                             cv=cv)
    backward_sfs.fit(X_train, y_train)
    selected_features_backward = backward_sfs.transform(X_train)
```

```
    # Define the XGBoost model
    xgb_clf = xgb.XGBClassifier(objective='multi:softmax',
                                num_class=3,
                                missing=1,
                                eval_metric=['merror', 'mlogloss'],
                                seed=42)

    # Initialize RandomizedSearchCV object
    random_search = RandomizedSearchCV(
        xgb_clf,
        param_distributions=param_dist,
        n_iter=50,  # Adjust the number of iterations as needed
        scoring='accuracy',
        cv=cv,
        verbose=2,
        n_jobs=-1,
        random_state=42
    )

    # Perform Randomized Search for hyperparameter tuning with selected features
    random_search.fit(X_train[selected_features], y_train)

    # Get the best estimator from Randomized Search
    best_xgb_clf = random_search.best_estimator_

    # Train the XGBoost model with the best hyperparameters and selected features
    best_xgb_clf.fit(X_train[selected_features], y_train,
                     eval_set=[(X_test[selected_features], y_test)],
                     early_stopping_rounds=10, verbose=0)

    # Predict for the test set of the current fold
    y_pred_fold = best_xgb_clf.predict(X_test[selected_features])

    # Calculate evaluation metrics for the fold
    accuracy_scores.append(accuracy_score(y_test, y_pred_fold))
    precision_scores.append(precision_score(y_test, y_pred_fold, average='weighted'))
    recall_scores.append(recall_score(y_test, y_pred_fold, average='weighted'))
    f1_scores.append(f1_score(y_test, y_pred_fold, average='weighted'))

    # Calculate confusion matrix for the fold
    cm = confusion_matrix(y_test, y_pred_fold)
    confusion_matrices.append(cm)

    # Perform cross-validation to get additional scores
    cross_val_accuracy_fold = cross_val_score(best_xgb_clf, X_train[selected_features], y_train, cv=cv, scoring='accuracy')
```