AN INTELLIGENT TECHNIQUE FOR THYROID DISEASE DETECTION USING MACHINE LEARNING

BY

Md.Shahnewaj Bin Hafiz

ID: 201-15-3049

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Ms.Nasima Islam Bithi
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Rasel Sarker
Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

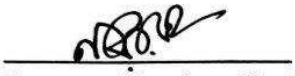
DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "An intelligent technique for thyroid disease detection using machine learning", submitted by **Shahnewaj Bin Hafiz** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 26-01-2024.

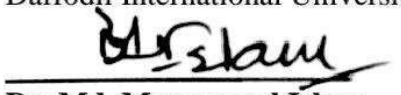
BOARD OF EXAMINERS



Narayan Ranjan Chakraborty — **Board Chairman**
Associate Professor & Associate Head
Department of CSE
Daffodil International University

Md. Sadekur Rahman — **Internal Examiner 1**
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Mr. Tanvirul Islam — **Internal Examiner 2**
Lecturer
Department of Computer Science and Engineering
Daffodil International University



Dr. Md. Manowarul Islam — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ms. Nasima Islam Bithi, Lecturer, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Bithi
13.07.24

Ms.Nasima Islam Bithi
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Rasel Sarker
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Md.Shahnewaj Bin Hafiz
ID: 201-15-3049
Department of CSE
Daffodil International University

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to Ms.Nasima Islam Bithi, Lecturer Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Ms.Nasima Islam Bithi, for her generosity for completing our project as well as to other faculty members and also the individual employee of the Department of CSE, Daffodil International University.

We appreciate our every course mate of Daffodil International University, who participated at this discussion while working on the courses.

Eventually, we have to acknowledge with due respect for our parents' patience and support.

ABSTRACT

An underactive thyroid gland is the hallmark of the common endocrine condition hypothyroidism, which can cause a variety of health problems. A timely and precise diagnosis is essential for the proper management and treatment of this illness. In this paper, we investigate how machine learning approaches—more especially, ensemble techniques like Bagging and Boosting—can be used to forecast hypothyroidism. We have taken two popular datasets from Kaggle and Figshare website. We use a wide range of data, such as laboratory and clinical characteristics, to train and assess various machine learning models. The Bagging method lowers variance and improves overall model stability by combining predictions from several base learners. By giving misclassified cases a larger weight, the technique known as "boosting" aims to repeatedly improve the model's accuracy. The most accurate classifier was the traditional technique, which achieved an impressive accuracy rate of 93.17% by Random Forest (RF). Other classifiers that were used included Logistic Gradient Boosting (GB), Regression (LR), Adaboost Classifier (ABC), K-Nearest Classifier (KN), Support Vector Machine (SVM), Decision Tree (DT), Ridge Classifier (RC), Quadratic Discriminant Analysis (QDA), Passive Aggressive (PA), Gaussian Naïve Bayes (GNB). In addition, 92.15% accuracy was obtained by the Boosting Gradient Boosting (GB), while Boosting Random Forest (RF) 91.86% accuracy was attained. Hyperparameter tweaking was used to maximize each classifier's performance. After conducting an experimental examination and reviewing prior research, it was determined that the Random Forest (RF) classifier performed very well, correctly diagnosing hypothyroid illness with an astounding accuracy rate of 93.17%.

CHAPTER 1
Introduction

1.1 Introduction

Managing daily obstacles comes with living with hypothyroidism, a disorder characterized by hormone deficits and a loss of overall function. Prompt diagnosis is essential in order to address the major worry of early prediction of this common condition. Machine learning, which examines a wealth of approved medical data and patient diagnostics, shows promise as a method for predicting hypothyroidism. Through a review of patient medical data, we were able to identify key markers of the illness and use this information to diagnose hypothyroidism. To do this, researchers have collaborated to develop machine learning algorithms.

One of the most common endocrine diseases in the modern world is thyroid disease (TD). Early diagnosis of TD can help minimize symptoms, even though the exact origin of the condition is still unknown [1] Given that TD is one of the endocrine illnesses that is developing the quickest in the current human population, predicting endocrine sickness is an essential challenge in the field of clinical data analysis. Many risk factors, including high blood pressure, high cholesterol, an irregular pulse rate, and other disorders, make it difficult to diagnose TD [2] On the other hand, machine learning (ML) has produced positive outcomes.

Applications of data mining are quite significant and helpful in the fields of medicine and healthcare. Unless it is converted into the most valuable knowledge and information that can support cost containment, profit maximization, and the high-quality maintenance of patient healthcare, the massive amount of data collected from healthcare organizations has no organizational value [7]. The process of identifying the sickness that is producing a person's symptoms is known as disease diagnosis. The most difficult issue is identifying an illness since some indications and indicators are non-specific [3]

Over time it has become obvious that machine learning techniques are extremely helpful in addressing the intricate and nonlinear problems associated with developing prediction

models. Models are needed for every illness and infection prediction in order to determine the important qualities that may be selected from various datasets and applied as accurately as feasible for a classification in the healthy patient. If not, a misdiagnosis might lead to a healthy patient receiving needless therapy [5].

Healthcare providers consistently prioritize prevention within the context of wellness care. It is crucial to do the right disease evaluation on a patient at the right time due to the related risk. These days, reports may be produced based on symptoms employing advanced diagnosis methods such as decision support systems, in addition to the traditional medical report [4]. Machine learning is a contemporary computer approach that builds a model that mimics human behavior using data and technology. The machine learning classification model will begin to predict the class of a given feature set once it has been trained [5].

Machine learning has made it possible for machines to learn without requiring explicit programming. An early illness diagnostic prediction model that provides treatment recommendations may be developed by applying machine learning techniques. Receiving top-notch treatment and an early diagnosis is the greatest way to lower the death rates from any condition. As a result, most medical professionals are interested in the recently created predictive model technologies that employ machine learning algorithms to forecast illnesses [6].

The thyroid is an organ that is shaped like a butterfly and is located in the neck, directly below the mouth. It produces the two primary hormones, T3 and T4, and is responsible for releasing hormones that regulate metabolism, including body temperature and heart rate.[5] classify thyroid diseases into two basic categories: hypothyroid and hyperthyroid. Thyroid illness is one of the most common lifestyle problems.

Higher than necessary hormone levels in the body cause hyperthyroidism. Hypothyroidism, the reverse of hyperthyroidism, reduces body metabolism and results in weariness, aches in the joints, and drowsiness. Numerous metabolic functions, such as body weight and heart rate, are regulated by these hormones. These effects may be interfered with if there are fluctuations in these hormone levels. Therefore, it's important to identify thyroid disease before starting treatment [26].

1.2 Motivation

Several academic institutions developed Machine learning algorithms to identify diseases of the human body, including hypothyroidism. However, it has been detected that the guess of the algorithms is not flexible or accurate in their prediction of hypothyroidism. Following this, we propose a novel strategy to improve the body's capacity to anticipate ailments. These methods for machine learning divided into two categories: supervised learning, which makes use of data which is labeled to produce outputs according to inputs via input-output, and unsupervised learning, which makes use of data which is unlabeled to reveal invisible sample and information. The techniques we have created are designed to specifically forecast the onset of hypothyroid disease.

1.3 Rational of the Study

The necessity to solve pressing issues in the fields of medical data and respiratory illness diagnostics is what led to the creation of the "An Intelligent Technique for Thyroid Disease Detection Using Machine Learning" project. The management and treatment of hypothyroid sickness, which is still a major global physical health concern, depend heavily on a prompt and precise diagnosis. The machine learning technique is a potent artificial intelligence technique that is properly integrated for

- Leveraging historical data from unrelated activities
- Improving the model's capacity
- recognizing patterns of ovarian sickness in medical data.

Using historical data from unrelated activities, machine learning is a powerful artificial intelligence technique that improves the capacity of the model to recognize patterns of ovarian disease in medical data when properly integrated. This study examines characteristics and prediction methods to go beyond simple prediction. Understanding the subtleties of these techniques in depth is essential to understanding the geographic scope and complexity of hypothyroidism patients. By carefully evaluating and predicting each approach, the research hopes to shed light on its advantages and disadvantages and add a substantial amount of new knowledge to the scientific community. This comparative study is essential for improving

the current diagnostic approach and encouraging the creation of more accurate and dependable systems.

The project's ultimate goal is to offer a dramatic improvement in the accuracy and effectiveness of prediction and assessment techniques, opening the door for later, substantial advancements in the diagnosis of hypothyroidism. Through this effort, we hope to advance the theoretical and applied disciplines of healthcare by better understanding how machine learning may be used in this setting. Improved diagnosis has a direct and positive impact on overall physical health as well as medical results.

The results of our research have led to the development of a prediction model for the prediction of human hypothyroidism, an illness that is gradually affecting our society. We have looked into machine learning as a possible treatment because we are aware of the lack of diagnostic information and resources, especially in our financially struggling nation where identifying hypothyroidism and evaluating symptoms can be expensive.

1.4 Research Question:

"An Intelligent Technique for Thyroid Disease Detection Using Machine Learning" is the subject of several important problems that the research seeks to answer. These inquiries function as directional markers, directing the research toward a thorough comprehension of the difficulties associated with implementing cutting-edge artificial intelligence systems for improved medical diagnostics. The following are the main research questions:

- **In what ways might machine learning enhance the effectiveness and precision of hypothyroid disease prediction in medical data?**

This study looks at the critical role that machine learning plays in quantifying data from unrelated activities to the unique challenges associated with diagnosing hypothyroid illness. In order to advance our knowledge of machine learning in artificial intelligence, the study attempts to identify the mechanisms by which it improves the accuracy and efficacy of identifying hypothyroid disease in medical datasets.

- **What part does the capacity of machine learning algorithms to predict hypothyroid sickness play in the whole diagnostic process, and how do they compare in this regard?**

This study looks into the precise role that machine learning—which is well known for its adeptness at data analysis—plays in hypothyroidism prediction. The study compares and evaluates machine learning's efficacy in predicting hypothyroid illness with the aim of advancing medical data analysis and improving overall diagnostic capabilities.

- **What is the impact of adding machine learning algorithm on the sensitivity and specificity of hypothyroid illness diagnosis compared to traditional methods?**

In order to investigate how machine learning and hybrid approaches perform in terms of sensitivity and specificity in the prediction of hypothyroid disease, this topic examines their performance. Through evaluating the impact of these sophisticated methods on diagnostic precision, the study seeks to identify areas for advancement and obstacles related to their use in the diagnosis of respiratory disorders.

- **Which preprocessing and prediction methods have benefits and drawbacks, and how might this knowledge be applied to enhance the accuracy and dependability of hypothyroid disease diagnostics in the future?**

This study includes a thorough analysis of the advantages and disadvantages of many methods for predicting and categorizing hypothyroid disease. By identifying and comprehending these variables, the study hopes to provide guidance for future investigations aimed at developing more reliable and accurate methods for hypothyroid diagnosis.

- **What implications do the study's findings have for medical dataset as a whole, and how can the information gained be applied to improve patient care and outcomes?**

The examination looks into the research findings' wider ramifications for the medical dataset industry. It attempts to evaluate the potential applications of the study's findings in

real-world settings, emphasizing the improvement of medical outcomes and treatment via more accurate diagnosis and prognosis of hypothyroid disease.

The conceptual foundation of the thesis is derived from these research issues, which also serve as a guide for delving into the intricacies of machine learning, evaluation techniques, hybrid machine learning, and hypothyroid prediction. Through a rigorous approach to these problems, the project hopes to add important knowledge that will affect future medical diagnoses and further the ongoing conversation about the relationship between artificial intelligence and health care.

1.5 Expected Output

This study aims to provide theoretical and practical outcomes in addition to expanding knowledge in medical science and technology. Among the expectations that have been voiced are:

Building a Sturdy Disease Prediction System:

The main outcome of this work is the development of an accurate and trustworthy method for hypothyroid prediction. Modern machine learning techniques will be used in this system, including our suggested model and models from Random Forest (RF), Gradient Boosting (GB), K-Nearest Classifier (KN), Passive Aggressive (PA), Decision Tree (DT), Support Vector Machine (SVM), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR), Ridge Classifier (RC), Adaboost Classifier (ABC), and Gaussian Naïve Bayes (GNB). The approach aims to provide high-accuracy diagnosis of various hypothyroid disorders to facilitate early intervention initiatives.

Assessment and Partial Comparison of Machine Learning Models:

There will be a thorough evaluation and analysis of the selected machine learning models. This research will detail each model's efficiency, accuracy rates, and performance indicators so that readers may determine whether or not a machine learning model is beneficial for them.

Integration of Predictive Strategies:

One important expected outcome is the integration of prediction methods into the disease

prediction system. The approach aims to forecast potential epidemics in their early phases as well as present diseases. This proactive approach, which aims to lower losses and increase the overall resilience of health management, is in line with medical best practices.

Application of Data Processing Techniques:

Project's ultimate is to boost the accuracy and efficacy of sickness prediction by utilizing sophisticated data processing techniques such missing value imputation, handling imbalance data, and categorical data encoding. The method integrates many approaches to offer thorough insights on the traits and geographic range of hypothyroid diseases.

Validation of Findings by Extensive Testing:

Strict procedures will be developed in order to verify the efficacy of the suggested illness prediction method. Numerous datasets will be used in the validation process to verify the system's robustness and suitability in a range of situations and changes in the environment that may affect the environments where hypothyroid disorders occur.

Contribution to Scholarly Conversation:

The goal of the project is to advance the academic discourse on the nexus between machine learning and medical knowledge in order for our suggested approach to provide the most accurate solutions to patients' issues in the real world.

1.6 Report layout

The suggested report uses a thorough structure to walk readers through the methods, conclusions, and consequences of the study. Every chapter has a distinct function and enhances and reinforces the work's overall cohesion.

Chapter 1: Introduction

An overview of the research and background information on the importance of detecting hypothyroid disease, machine learning,

Chapter 2: Background

An overview of the theoretical underpinnings, useful techniques, and technical developments pertaining to machine learning and the identification of hypothyroid illness are comprehensively explained in this section.

Chapter 3: Research Methodology

The study's execution strategy is described in the research methodology chapter, this chapter discusses the study's potential limits and moral dilemmas.

Chapter 4: Experimental Result

The results we experiment of the machine learning-based prediction of hypothyroid disease are presented in this chapter.

Chapter 5: Impact on Society

The research findings' larger implications for medical diagnosis and health management are examined in the chapter on social effect.

Chapter 6: Summary, Recommendation, Conclusion and Implications for Future Research

The whole report is summarized in this last chapter. It restates the research's results, highlights its major discoveries, and makes recommendations for further research and useful uses.

CHAPTER 2

Background

2.1 Preliminaries

Researchers who are interested in building on the current work may choose a direction for future research by discussing the implications of the findings. The reader is guided through the study process from the opening to the broader ramifications and guidance by this well-organized arrangement, which guarantees a logical flow of data. It is now possible to thoroughly comprehend the study methodology and any possible effects it may have on the theoretical and practical elements of diagnosing hypothyroid illness thanks to machine learning. These terms are used repeatedly throughout the study to investigate the connections between hybrid and machine learning techniques. Before delving into the study's intricacies and comprehending the subtleties of the methods used to identify hypothyroid sickness, readers must be aware of these preparatory steps.

2.2 Related Works

One potential strategy has been the application of ML classifiers to the prediction of hypothyroid illness. Because machine learning methods work well in this field, they are frequently used in conjunction with tree architectures for decision models [9].

Machine learning models specifically designed for the categorization of heart diseases have been developed with the help of several researchers. A model with Decision Tree (DT) with an outstanding accuracy of 97.6% was introduced by [22].

Other models with impressive accuracy were Naive Bayes (NB) with 96.7%, Support Vector Machine (SVM) with 75.1%, and an Ensemble technique with 97.3%. Maheshwari and Afzal (2000) introduced a model with three tiers of accuracy: 98.50% for Random Forest (RF), 97.02% for Support Vector Machines (SVM), and 95.81% for K-Nearest Neighbor (KNN). A machine-learning model was suggested [14]

It utilized KNN to reach 92% accuracy and SVM to obtain an astounding 97.8% accuracy. On the other hand, KNN, with an accuracy of 93.44%, and Naive Bayes, with an accuracy

of 22.56%, were used in the model developed by Chandel and Khushboo [25]. P. Duggal and S. Shukla collaborated to conduct research using a variety of feature selection techniques and classification algorithms to detect thyroid illness. In addition to Naive Bayes, Support Vector Machine (SVM), and Random Forest, they also employed Tree-Based Feature Selection, Recursive Feature Elimination (RFE), Univariate Selection (UFS), and Random Forest techniques. Notably, combining the SVM approach with RFE produced the greatest accuracy of 92.92% [26].

Another set of researchers used machine learning approaches for classification, such as K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN), to predict the severity of hypothyroid illness. SVM produced the highest accurate results, at 99.08%, after data normalization and parameter adjustment, whereas logistic regression showed 96.08% accuracy [27]

In a different study contributed techniques including SVM, Decision Tree, Random Forest, Logistic Regression, and Naïve Bayes. While Naïve Bayes had a little lower but still respectable 94.23% accuracy, SVM, Decision Tree, Random Forest, and Logistic Regression all achieved an incredible 99.35% accuracy. These algorithms produced very high rates of accuracy. When combined, these findings demonstrate how machine learning algorithms might raise the accuracy and effectiveness of techniques for diagnosing thyroid and heart conditions[28].

acquired an accuracy of 99.62% with cross validation k=6 for the thyroid disease (LDA) technique using the linear discriminant analysis data mining strategy. Similar to this, Shivanee et al. (2016) developed a classifier for the diagnosis and classification of hypothyroid disease using different k-fold cross validation for the C4.5 classifier. To do this, a number of data mining classification techniques were used, such as the Mutlilayer Perceptron, RBF Network, Baileys net, C4.5, CART, Decision stump, and REP tree approaches. He was correct for various k-folds; nonetheless, 99.60% was the most accuracy he could get for k = 6, while 99.575% was the greatest for k = 10. Several machine learning methods were put out by for the diagnosis and treatment of thyroid conditions. [29]

In contrast, SVM produced the best accuracy result, coming in at 99.63%. Furthermore,

103 patients with suspected thyroid nodules who had thyroid ectomy and were not amenable to chemical analysis were included in a study by (training cohort–to–validation cohort ratio, ≈ 3:1). The most accurate model was cytokeratin-19, which demonstrated 84.4% accuracy for the training cohort and 80.0% accuracy for the validation cohort[25].

The training cohort accuracy of the thyroperoxidase and galectin 3 prediction models is 81.4%, whereas the validation cohort accuracy is 84.2% and 85.0%. With an accuracy of just 65.7%, the high molecular weight cytokeratin prediction model demonstrated poor performance and could not be confirmed. In an attempt to prevent unnecessary endoscopic treatments, proposed developing a predictive diagnosis model for esophageal varices that incorporates a minimal number of the most significant characteristics. Only 10 of the dataset's more than forty distinct clinical laboratory variables were determined to be significant; the results were enhanced by combining the p-value filtering method with the correlation coefficient[21]

This study developed a unique method called "Boosted-Naïve Bayes Tree" (B-NBT), which enhances the conventional Naïve Bayes Tree by adding a boosting strategy, hence improving the overall performance of the predictive diagnostic model. The dataset performed better when B-NBT was used, with an AUROC (Area under Receiver Operating Characteristic Curve) of 0.865 and an accuracy of 79%. An ensemble of bagging with J48 and an ensemble of bagging with SimpleCart was designed by in order to extract important information and diagnosis for the prediction of hypothyroid disorders.[5]

The experiment using an ensemble of bagging with simple cart produced the best prediction accuracy (99.6554%), whereas the accuracy using bagging with J48 was 99.6023%.investigation involved the enrollment of consecutive patients who had thyroid ectopy and neck MR scans performed during the study period. Pathological analysis of thyroidectomy tissues was used to evaluate the diagnosis and degree of aggressiveness associated with PTC. After manually segmenting the thyroid nodules on the MR images, radiomic characteristics were taken out. The experiment used a gradient boosting classifier to identify PTC aggressiveness and least absolute shrinkage and selection operator for radiomic feature selection, yielding an AUC of 0.92[19].

In the same study, Range et al. offered six categories for the diagnosis of thyroid fine-

needle aspiration biopsy The associated cancer risk for each group is crucial information for thyroid nodule therapy. To detect follicular cells and estimate the ultimate disease's malignancy, an MLA was created. Patients in the test set were blindly assessed and assigned to TBSRTC categories by a cytopathologist. The performance of the MLA was evaluated using the area under the receiver operating characteristic curve. 988 selwals met the requirements. With a sensitivity and specificity of 92.0% and 90.5%, respectively, the MLA predicted malignancy [7].

2.3 Comparative Analysis and Summary

Comparison study, which identifies the minor variations and complementary aspects of multiple approaches, this research on "An Intelligent Technique for Thyroid Disease Detection Using Machine Learning" would not be feasible. The effectiveness of machine learning in identifying hypothyroid illness is thoroughly assessed in this work, which makes use of both cutting-edge evaluation algorithms and conventional techniques. The research evaluates the advantages and disadvantages of each strategy through thorough testing and experimental result analysis, offering a thorough grasp of their distinct contributions to the diagnostic process.

A full investigation of prediction and hybrid algorithms' contributions to improving the precision and effectiveness of the diagnosis of hypothyroid illness is made possible by the extensive examination of these algorithms. The study greatly increases the understanding of the subject and helps practitioners and researchers choose the best solutions for their particular situations by comparing the results of various approaches. Beyond just highlighting differences, the study demonstrates how prediction and assessment might work together to improve our comprehension of the features and geographic distribution of hospitals impacted by hypothyroidism.

Ultimately, this work's comparison study provides a valuable perspective for evaluating the effectiveness of machine learning in hypothyroid prediction. It draws attention to the necessity of a comprehensive approach in medical diagnostics and presents a convoluted picture through the integration of conventional and cutting-edge methods. The findings

improve our knowledge of the prognosis of hypothyroid disease and provide valuable information for improving current diagnostic techniques. The comparative study serves as a strong basis for the discussion of the social effect, recommendations, and implications for more research in the following chapters.

Reference No	Dataset Name	Model	Limitations
[2]	Thyroid Disease Data (Kaggle)	Random Forest	They did not use any hybrid or ensemble techniques
[22]	Thyroid Disease Data (Kaggle)	Decision Tree	They had used a small dataset and the result analysis was not enough
[23]	Thyroid Disease Data (Kaggle)	Random Forest	They did not use any hybrid or ensemble techniques
[25]	Thyroid Disease Data (UCI)	KNN	They did not use any hybrid or ensemble techniques
[26]	Thyroid Disease Data (Kaggle)	SVM and Random Forest	They did not use any hybrid or ensemble techniques

TABLE 2.3.1: Comparative Analysis Table

2.4 Scope of the Problem:

The "Machine Learning Based Hybrid Technique to Predict Hypothyroid for Combined Data" purpose of this study expands the scope of the problem to cover medical diagnostics in general as well as patient problems in particular. This pursuit is driven by the need to close the large gaps in current knowledge on the diagnosis and subtype classification of hypothyroid disease. Conventional techniques sometimes fall short in accurately and efficiently differentiating between

the many forms of hypothyroid illness, leading to treatment outcomes that are not ideal and diagnostic delays. The goal of this work is to lay a strong basis for the use of visual data in machine learning to classify hypothyroid illness subtypes. In the process, we intend to revolutionize the diagnostic process and provide quick, trustworthy information to medical experts so they may tailor treatment programs and, ultimately, improve patient outcomes in the battle against illness. The challenge of detecting hypothyroidism with traditional methods, which can have limitations in fact of overall diagnostic sensitivity, specificity, and accuracy, is included in the issue's scope.

2.5 Challenges

The study explores the use of machine learning, a rapidly developing artificial intelligence paradigm, to improve ovarian diagnostics. By using information from unrelated activities, machine learning enables the model to possibly handle the complexity and unpredictability present in medical data related to hypothyroid disease. Because machine learning models are a great tool for extracting complicated patterns and features from complex visual input, integrating machine learning adds even more intricacy and sophistication.

The challenge's scope also includes a comparative examination of assessment and prediction techniques, which includes a thorough investigation of their advantages, disadvantages, and possible synergies. For a thorough diagnosis, an understanding of the architectural design and features of hospitals impacted by hypothyroidism is crucial, and this research attempts to provide insights that go beyond conventional screening techniques. The problem's scope essentially includes the difficulties in detecting hypothyroid disease, the possible improvements provided by machine learning, and the comparative analysis of different methods. The initiative aims to improve the precision of presently existing diagnostic tools and lessen the impact of hypothyroidism on public health by tackling these issues.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Research Subject:

Using cutting-edge AI approaches in medical diagnostics, the research project "Machine Leaning Based Hybrid Technique to Predict Hypothyroid for Combined Data" focuses on the prediction of hypothyroid illnesses. Our study's main focus is hypothyroidism, a prevalent respiratory ailment with potentially harmful effects. The study looks at the drawbacks of the present approaches to hypothyroid diagnosis and explores potential solutions using machine learning.

One of the distinctive concepts in artificial intelligence is machine learning, which enables one to use information from unrelated occupations to the difficult task of identifying hypothyroid sickness. The diagnosis procedure becomes more accurate and thorough when intricate patterns and features are extracted from medical photos using machine learning, which is well known for its effectiveness in data analysis.

The study's comparative analysis section examines the variations and complementarities among different prediction and assessment techniques. Prediction establishes if the illness is documented in medical records, whereas feature selection focuses on describing the features and geographic arrangement of hospitals impacted by hypothyroidism. The study aims to get a thorough understanding of the advantages and disadvantages of various approaches, which will facilitate the creation of more trustworthy and efficient diagnostic instruments.

The main focus of this work is the interaction of artificial intelligence, respiratory health, and medical data. It specifically looks at how machine learning may be used to improve the prediction of hypothyroid disease. The discoveries are intended to enhance health performance and boost diagnostic precision, which will benefit health care's theoretical and practical aspects.

3.2 Data Collection Procedure

The dataset, which had 26404 rows and 29 columns, was easily accessible for use and was obtained via Kaggle and Figshare. Out of these columns, the diagnostic feature was critical in classifying the frequency of hypothyroidism, and each feature was important in determining the presence of this ailment. Patients were categorized into two categories, 0 and 1, which stood for the presence and absence of hypothyroidism, respectively; Table 3.2.1 shows that 21954 patients were in the former group and 4450 in the latter. Training the test set were the two additional segments created from the dataset. Eighty percent of the candidates were chosen for the training set, while the remaining twenty percent made up the test set. This allowed for thorough model building and evaluation for the prediction of hypothyroidism and the attributes are Age, Sex, On thyroxine, Query on thyroxine, On antithyroid medication, sick, pregnant, Thyroid surgery, I131 treatment, Query hypothyroid, lithium, goitre, tumor, hypopituitary, psych, TSH measured, TSH , T3 measured, TT4, T4U measured, T4U, FTI measured, FTI, TBG measured and target output (binary class).

The data count plot is shown below:

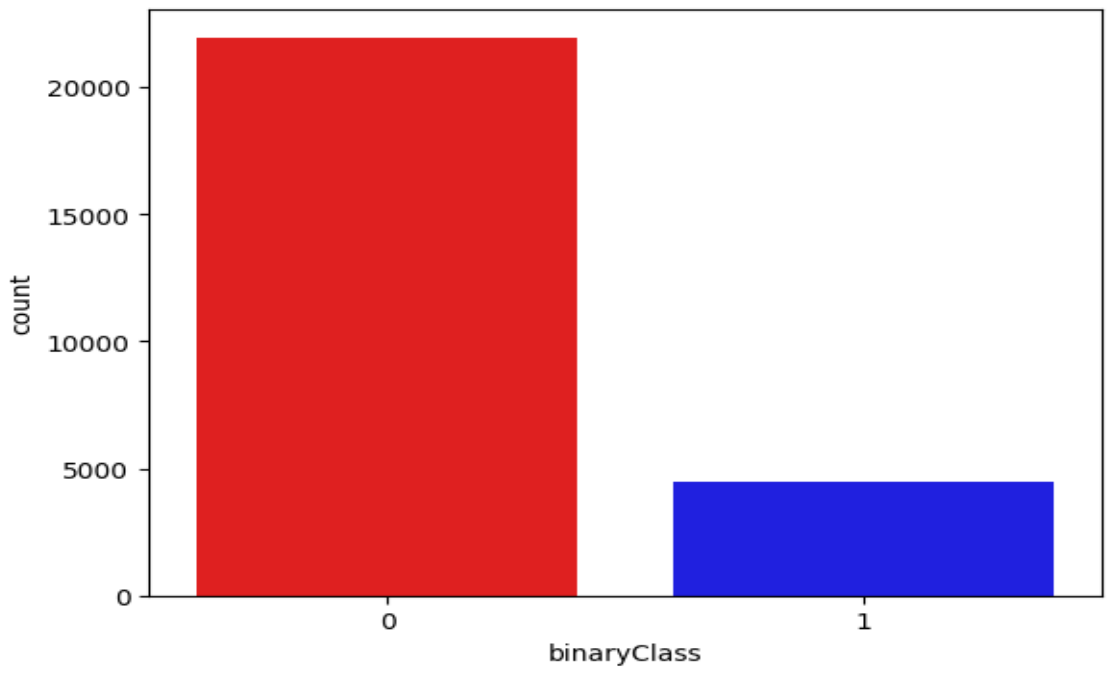


Figure 3.2.1: Hypothyroid Disease Data Count before SMOTE

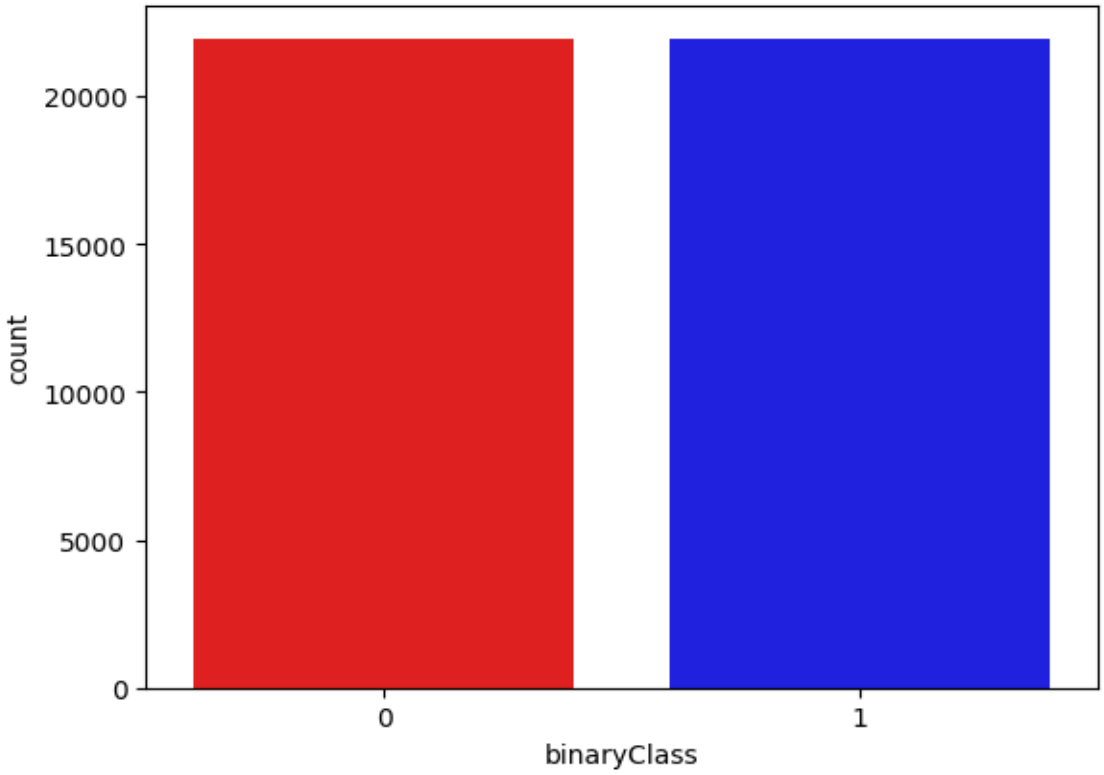


Figure 3.2.2: Hypothyroid Disease Data Count after SMOTE

3.3 Working principle of the proposed model

The initial stage in the recommended architecture is to extract information from the hypothyroid dataset. In situations when machine learning algorithms require more data to train and function more effectively, data augmentation techniques are employed to address the problem. The working principle of the proposed model have been given as follows

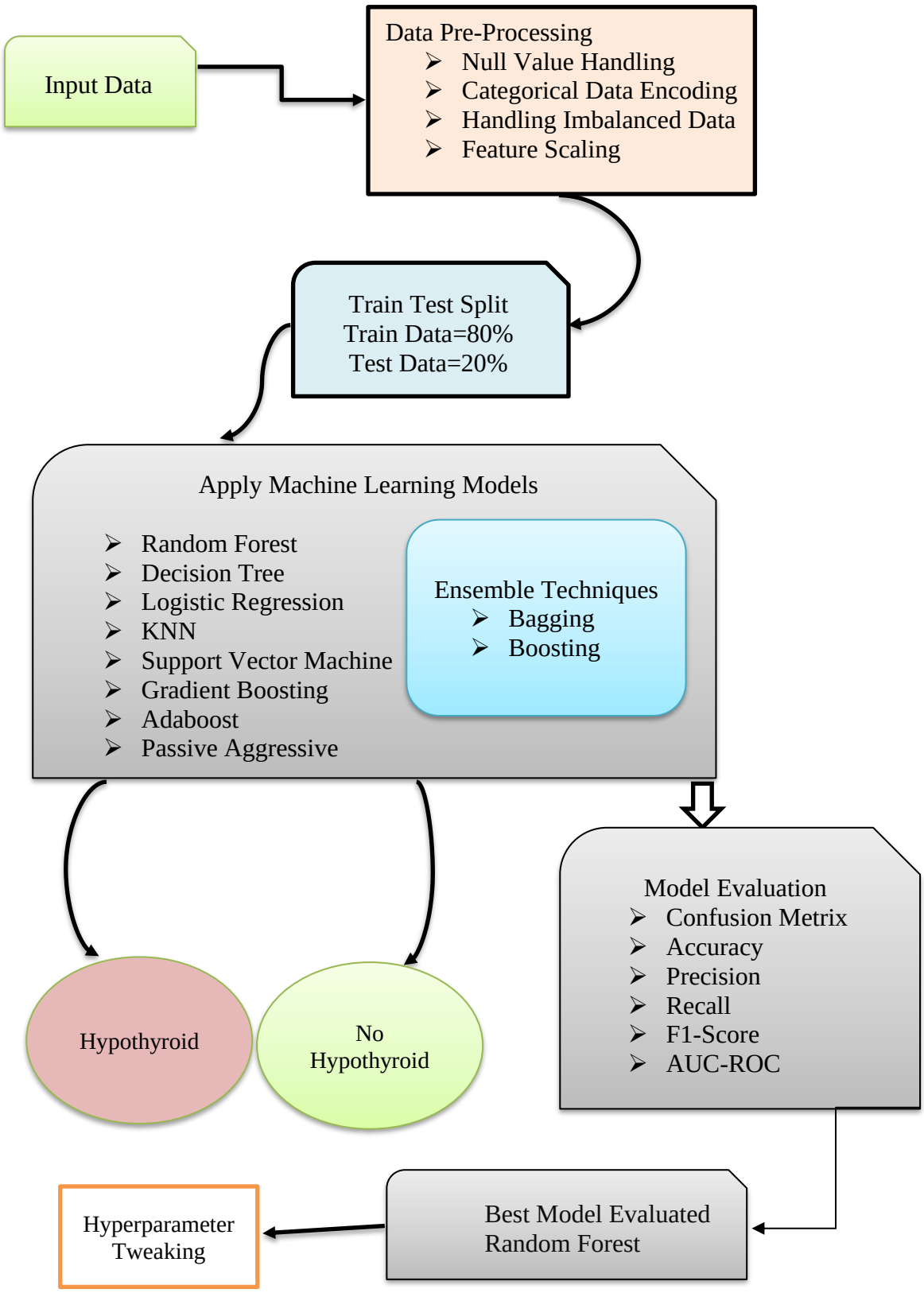


Figure 3.3.1 Methodology of the work

Data Pre-Processing

Four kinds of pre-processing techniques have been applied in this paper and there is null value handling, categorical data encoding, imbalance class handle by SMOTE and feature scaling which are described below

Categorical Data Encoding

This technique, which turns category data into numerical values, is essential to our study. Since machine learning is dependent on numerical input and output, this method's application was essential. We have to change the dataset's columns in order to make it easier to use the categorical data encoding approach and include categorical data into our analytical framework.

Missing Value Imputation

In this procedure, values that are imputed—derived from the examination of data from other datasets—replace missing or incomplete data. It's important to note that there were missing values in our dataset, but we were able to replace them.

Handling Imbalanced Data

In this method, the distributed class within a dataset is manipulated. By methodically adding new instances to the dataset, it manages the data in an efficient manner. Consequently, it uses the complete dataset as input and improves minority data representation. To balance the dataset, we employed the SMOTE approach.

Feature Scaling

This technique involves handling data with negative values by varying the scale of many independent data variables. Changes were made as part of this procedure to guarantee consistent scalability. Furthermore, it was determined that two columns—TBG and referral source—had to be eliminated from the dataset.

A correlation subplot was used to examine the complex relationships between two variables and showed how changes in one variable affected the behavior of the other. The

probability of correctly predicting one element from another depended heavily on the degree of interaction among the variables. Our comprehension of the dataset has expanded, which has enhanced our capacity to pinpoint the primary determinants of hypothyroidism. Figure 3.2.3 provided an all-encompassing perspective of all the characteristics linked to the anticipated attribute "Hypothyroid Disease," illuminating the interconnections within the dataset and facilitating the development of more precise forecasts.

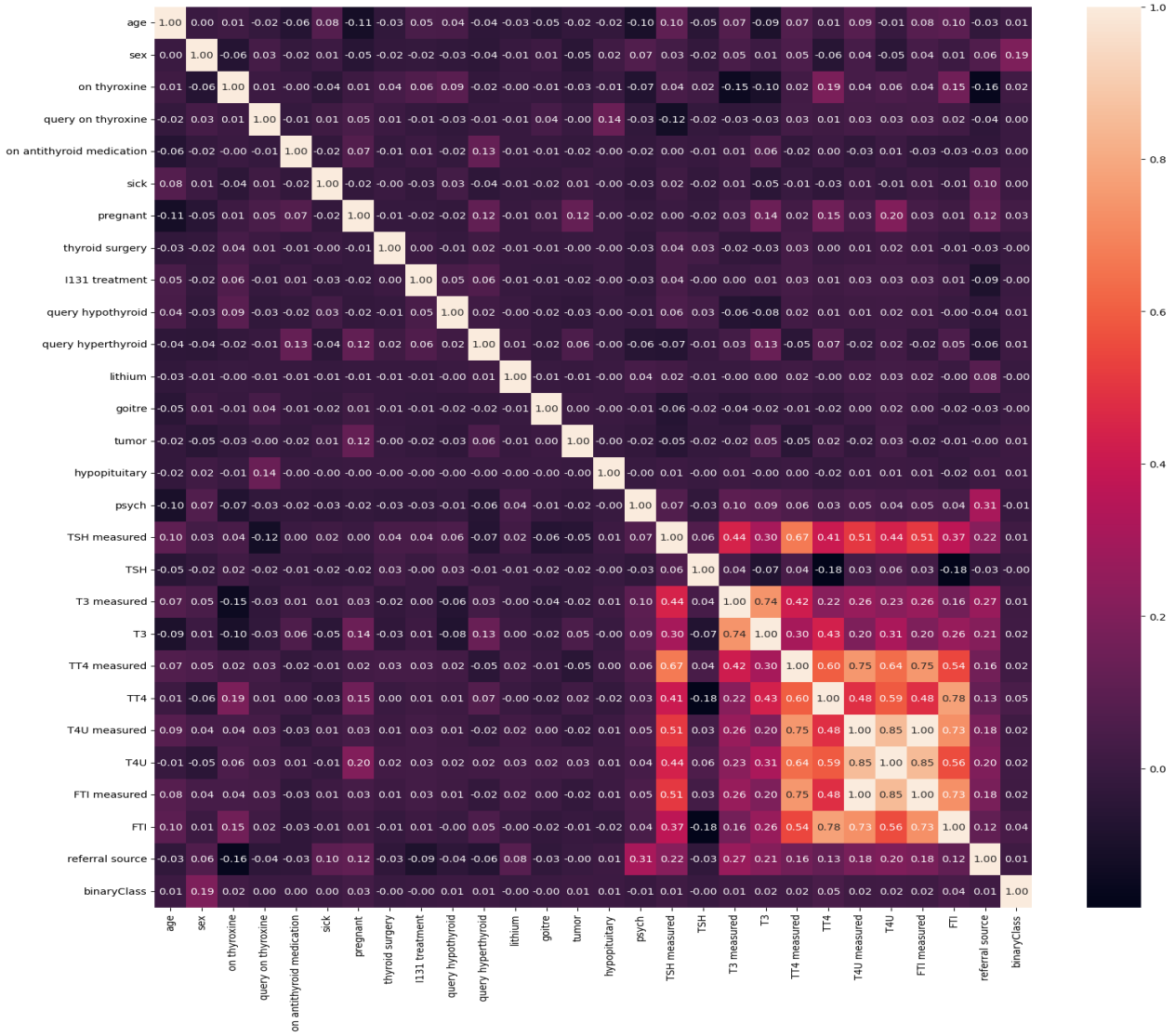


Figure 3.3.2: Correlated Features of Hypothyroid Disease Dataset

3.4 Applying Machine Learning Models

In this section, we demonstrated how to efficiently forecast hypothyroidism using a process diagram Significant data pre-processing techniques were applied when the data is used for train and test in our system were displayed, including the Standard Scaler Transform., Categorical to Numeric conversion, and Feature Selection. In an effort to maximize our forecast accuracy, we utilized ensemble algorithms, which include bagging and boosting. This made it possible for us to get the most out of the combined algorithms and produce thoroughly examined outcomes. To find out how well the models used in this phase predicted hypothyroidism, outcome analysis was performed on the models.

Random Forest

The ability of Random Forest can handle high-dimensional data, maintain outlier resistance, and provide insight into feature importance for model interpretability. Furthermore, because it employs bagging (Bootstrap Aggregating) and feature bagging techniques, the Random Forest algorithm has a lower tendency to overfit than single decision trees. The Random Forest classifier is a robust and flexible machine learning algorithm that is widely used for challenges including regression and classification both. It constructs a set of decision trees, each of which is made using a subset of the training data and available features chosen at random. its average of the predictions from each tree. Compared to other techniques, it requires less hyperparameter adjustment and is especially useful for handling complex and noisy datasets. Furthermore, Random Forests are able to identify important characteristics and evaluate how they affect the predicted accuracy of the model. Many industries, including banking, healthcare, and image analysis, have adopted it widely because to its robust performance, scalability, and adaptability. Higher computational costs and complexity are a trade-off for the algorithm's flexibility and power, this could be problematic for applications which require to use limited resources or real-time. Random forests continue to be a solid machine learning technique in spite of these difficulties, offering accurate predictions and insightful analysis for a variety of issues [13].

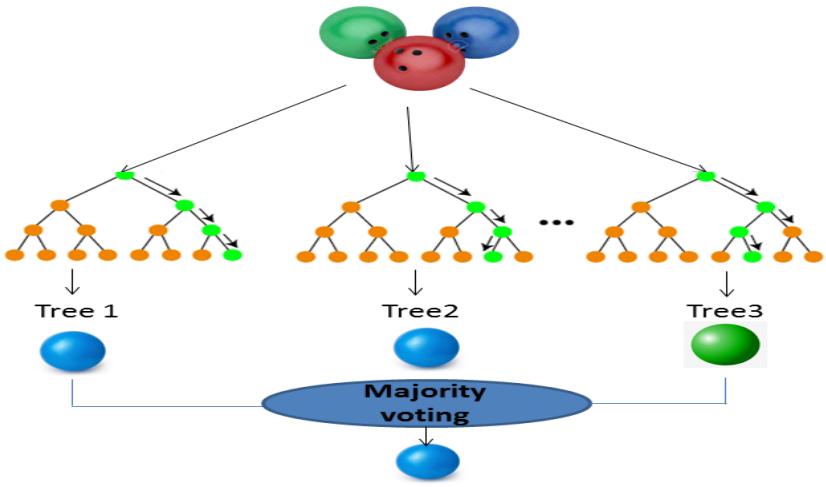


Figure 3.4.1 Random Forest

Decision Tree

To categorize an instance, at first, we must examine characteristic and then the representation of tree node. Secondly, we comply with a section of the structure whose value is determined by that feature. Most popular and successful among prediction methods is the Decision Tree methodology, that only requires two number classes. An attribute test is represented by each inner node of a decision tree, which is an ordered data structure in which each leaf hierarchy node indicates a different class. A tree structure called Decision Trees (DT) is widely used based on decision trees. The method may be applied to regression and classification problems. The "Splitting" process is used to choose the "Best Features" or "Best Attributes" from potential characteristics pool when the tree develops from root node. Finding the "Best Attribute" usually involves computing two additional metrics: "Entropy," as stated in (1), and "Data Gain," as noted in (2) [13].Entropy examines a dataset's consistency, but data collection gauges the rate at which changes in an attribute's volatility occur.

E(D)=-P (positive)log2P (positive)-P(negative) logf_0 ⟦2P (negative) (1)⟧

Gain (Attribute X)=Entropy (decision Attribute Y)-Entropy (X,Y) (2)

Naïve Bayes

The name "GNB" refers a group of classification algorithms based on the Bayes Theorem that determine the likelihood of an event occurring depending on the likelihood of another event occurring. The underlying assumption of all the algorithms in this category is that any two traits that are being recognized have no relationship with one another (equation 3).

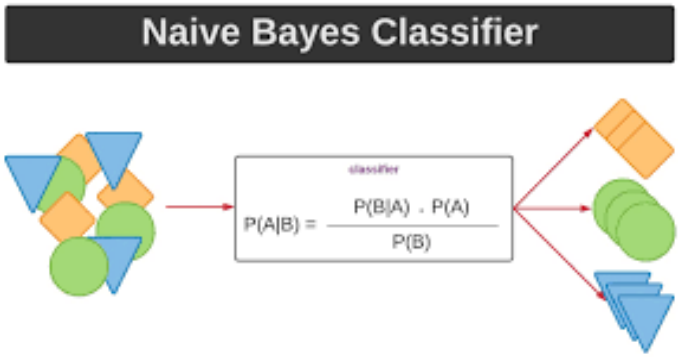


Figure 3.4.2 Naïve Bayes

$$P(A \mid B) = (P(B \mid A)P(A))/(P(B)) \quad (3)$$

The constant value is taken to represent a Gaussian distribution for every characteristic in Gaussian NB. The term “Normal distribution” is often used interchangeably w_(i^th).

$$P(x\_i \mid y) = 1/\sqrt{(2\pi\sigma y^2)}\ \exp(-(x\_i-\mu\_y)^2/(2\pi\sigma y^2)) \quad (4)$$

Logistic Regression

A popular and easily understood machine learning classifier, logistic regression works well for both binary and multiclass classification problems. In contrast to linear regression, which forecasts continuous values, logistic regression uses the logistic function (sigmoid) to represent the likelihood that an instance will belong to a certain class. It can clearly separate classes because it transfers the estimated probability of an event happening to a range between 0 and 1. Gradient descent and other iterative optimization techniques are used to minimize the logistic loss or cross-entropy loss in order to train the model. The benefits of logistic regression are its rapid training, simplicity, and interpretability. Through polynomial or

interaction terms, linear and non-linear correlations between the target variable and the features are both can be handled. While It is primarily a binary classifier, however one-vs-rest or softmax regression techniques can be applied to extend it to multiclass situations. When working with high-dimensional data or complicated connections, one of its limitations is that it is prone to overfitting. This may be lessened by using regularization techniques such as L1 (Lasso) and L2 (Ridge) regularization. Because of its transparency and efficacy, logistic regression is a foundational model in many machine learning pipelines and is useful in a variety of fields despite its simplicity, such as healthcare (disease outcome prediction), and natural language processing (text classification) [1].

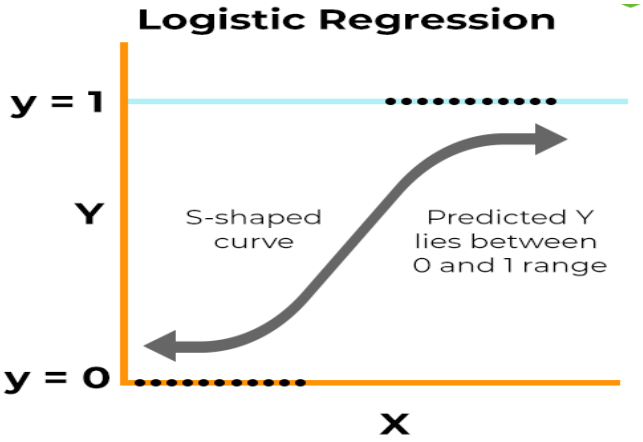


Figure 3.4.3 Logistic Regression

Support Vector Machine

The Support Vector Classifier (SVC) can be used to handle issues related to both of the regression and classification. Still, the most common use of artificial intelligence is in problems involving classification. To accurately classify new data points, the SVM technique searches toward a straight line, also a judgmental limit, which splits the region turns into categories in all variables. This maximum utility bound is a hyperplane. It is possible to generate a hyperplane by using SVM, that selects extreme locations, vectors Consequently, the name of the technique's, "support vector machine," originates from phrase "support vector," which is used to characterize these dire conditions [1]

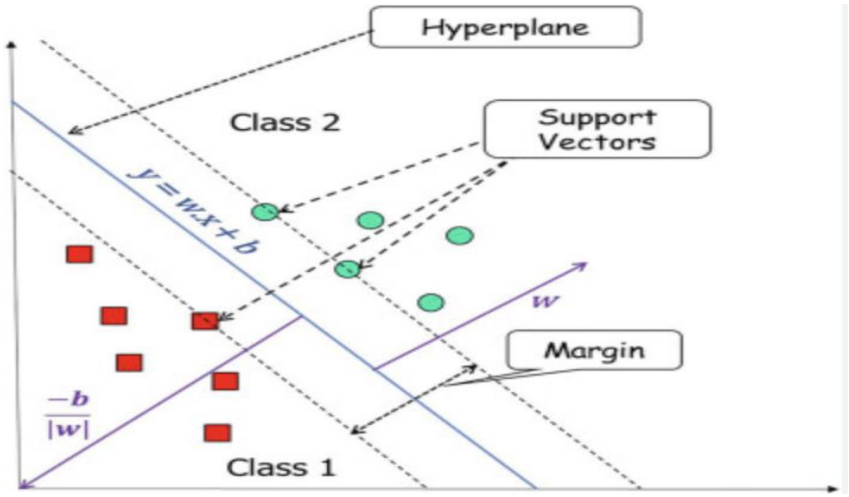


Figure 3.4.4 Support Vector Classifier

Gradient Boosting

A strong and adaptable machine learning technique, the gradient boosting classifier performs exceptionally well in predictive modeling, especially in classification problems. It works by combining many weak models—usually decision trees—in a sequential fashion to iteratively create a strong prediction model. The method gives the incorrectly identified data points from the previous step more weight at each iteration. The algorithm may constantly improve its predictions thanks to this iterative process, which eventually results in the creation of a strong ensemble model. The Gradient Boosting Classifier's capacity to handle complicated, high-dimensional data and grasp nuanced correlations within variables is one of its main features. It can obtain better prediction performance by aggregating the outputs of several weak learners. But, compared to certain other techniques, training a Gradient Boosting model might take longer because of its computing expense. Gradient Boosting implementation requires meticulous hyperparameter adjustment and cross-validation to reduce the danger of overfitting. Important parameters that affect the model's performance are the maximum depth of trees, the quantity of iteration used to boost (trees), and the learning rate. Gradient Boosting is a widely used method in numerous fields like data mining, finance, and biology, though it can accurately handle classification problems and yield precise answers. [14].

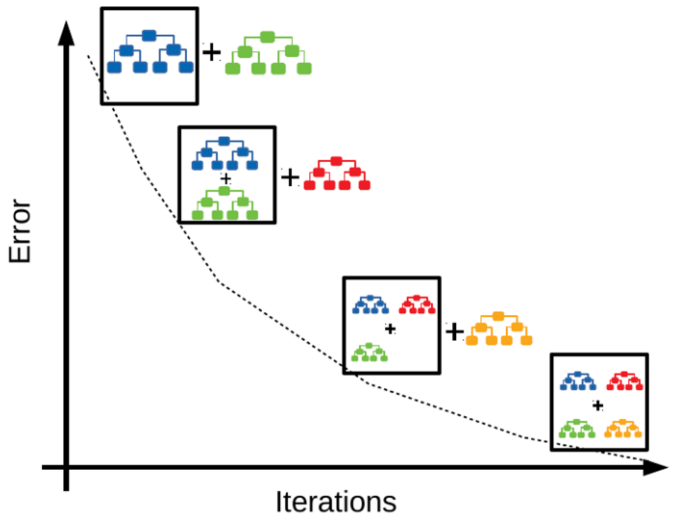


Figure 3.4.5 Gradient Boosting

K-Nearest

For classification problems, a popular and user-friendly machine learning technique is the K-Nearest Neighbors (KN) classifier. It focuses on the idea that data points with comparable characteristics typically belong to the same class. An input data point is categorized using the KN method according to the majority class of its K closest neighbors in the feature space. The number of neighbors to take into account, K, is a crucial hyperparameter that affects how well the method works. Since KN is non-parametric which can't make any significant assumptions regarding the distribution of the underlying data, it may be used in a variety of situations. It is a well-liked option for machine learning jobs that are basic because of its simplicity and ease of implementation. Due to the need to compute the distances between each data point in the dataset and the one under consideration, KN's computational efficiency may provide a challenge for huge datasets. Furthermore, dimensionality can impair KN's accuracy as the number of characteristics or dimensions rises, and its performance is susceptible to the choice of distance measure. Techniques like dimensionality reduction, feature selection, and meticulous hyperparameter tweaking are frequently used in concert with KN to overcome these issues. Even with its drawbacks, KN is still a useful tool for a lot of classification issues, especially when the particular dataset is small and the algorithm's presumptions match the actual data distribution.

Adaboost

Adaptive Boosting, or AdaBoost, is a potent ensemble learning technique that combines weak classifiers into a robust and accurate model, improving their performance. AdaBoost works in an iterative manner, successively changing each training instance's weight according to how well the preceding weak classifiers performed. By giving misclassified instances a larger weight, this enables weaker classifiers to concentrate on them and increase the accuracy of their classifications. The weighted outputs of these weak classifiers are then combined to get the final prediction. AdaBoost may be used with a variety of primary classifiers, most often shallow decision trees or decision stumps, which gives it flexibility to handle a variety of classification tasks. Since it places more attention on difficult data points during training, it works especially well with complicated datasets and overcomes problems like overfitting. AdaBoost is also renowned for its proficiency with high-dimensional unique feature spaces. Even though AdaBoost contains strong algorithm, noisy or out-of-the-ordinary data can nevertheless negatively impact its performance. Its sequential learning process, however, strengthens its ability to minimize these problems. In several domains, such as face prediction, and bioinformatics and also text classification, where impressive performance and flexibility are crucial, AdaBoost's mixture of weak learners is leveraged to produce a powerful and accurate classifier.

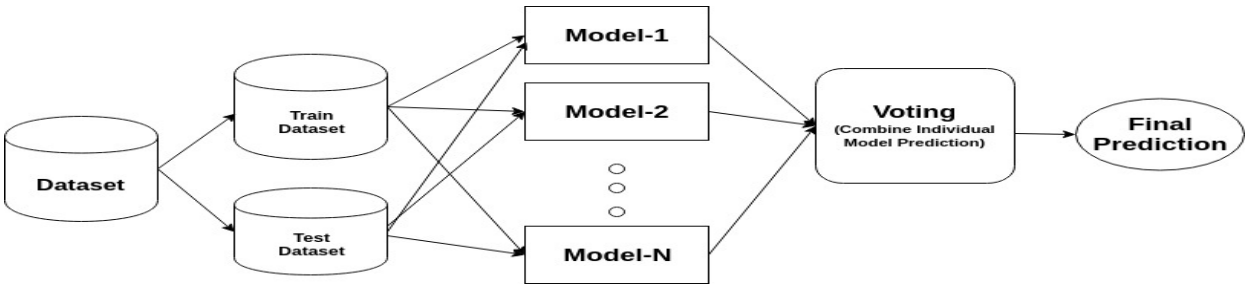


Figure 3.4.6 Adaboost

Quadratic Discriminant Analysis

Statistical classification method utilized in pattern recognition and machine learning is called quadratic discriminant analysis, or QDA. This technique is a continuation of Linear Discriminant Analysis (LDA) and is especially useful in situations when this condition of

the equal matrices which is covariance across classes don’t meet the satisfactory requirement. With QDA, every class has a unique covariance matrix that defines it, making the model more adaptable and better able to depict the data's underlying distribution. By calculating the probability distributions of the input features for each of the class, QDA seeks to determine the decision boundaries that optimally divide various classes. Compared to linear boundaries, quadratic decision limit (QDA) allows for further intricate interactions among variables when modeling of data point which located to a certain class. In QDA, the mean and covariance matrix for each class is estimated, each class's posterior probability of a given data point then determined by applying Bayes' rule. The class along with the largest posterior probability which is allocated to the data’s point during the classification process. Although QDA contains useful for capturing non-linear decision boundaries, overfitting may occur if there are too many features, as it necessitates estimating additional parameters. The underlying distribution of the data and specific assumptions about the matrix of covariance determine which of both Bayesian distribution methods to use: QDA or LDA. In general, QDA provides a more flexible method than LDA is a useful classification tool where class-particular covariances are uneven.

Ridge Classifier

By adding L2 regularization, also referred to as Ridge regularization, to conventional linear models, the Ridge Classifier is a method for linear classification. The purpose of adding this regularization element to the normal linear regression cost function is to reduce overfitting and enhance the model's capacity for generalization. When it comes to classification, binary or multiclass problems are frequently handled by the Ridge Classifier. It stimulates the coefficients the limitation of linear decision to reduce by applying Ridge regularization on them. Large values are penalized by this regularization factor, which is based on the coefficients of L2 norm square. A trade-off is introduced by the regularization term between maintaining modest model parameters and providing a good fit to the training set. One hyperparameter that regulates the regularization intensity is alpha. Greater regularization strength results from higher alpha values, which also lead to a simpler model with lower coefficients. As a member of the regularization family of classifiers, the Ridge Classifier improves model stability and guards against overfitting. When working with datasets that include correlated features and multicollinearity problems, it is very helpful. Ridge

regularization evens out the influence of correlated characteristics, contributing to the model's stabilization. The Ridge Classifier is implemented in Scikit-learn, a well-known Python machine learning package, enabling practitioners to use it in a variety of categorization settings.

Passive Aggressive Classifier

An online learning system called the Passive-Aggressive (PA) Classifier was created for large-scale, dynamic datasets. It adapts to situations with changing information by gradually updating its model when it comes into contact with fresh input since it learns in a slow manner. One training example at a time, the algorithm evaluates them all and modifies its model in response to inaccurate predictions. The aggressiveness parameter guides the update process, which modifies the model to correct faults; more aggressiveness results in faster adaptations. Because of its adaptability, the Passive-Aggressive method can be used for problems that involve both regression and classification. It is an important tool for situations where there is a steady and significant inflow of data because of its applicability in a various field, for example, text categorization and natural language processing. Various variations, such PA-I and PA-II, provide flexibility for adjusting to the particular data features and also learning needs.

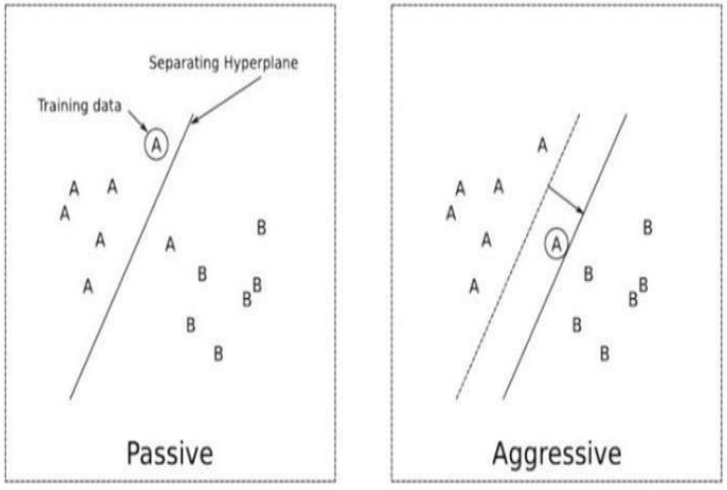


Figure 3.4.7 Passive Aggressive

Ensemble Learning Algorithms

The process of merging many machine learning models to increase overall prediction accuracy and resilience is known as ensemble learning [13].

Bagging Classifier

Specifically for decision tree algorithms, bagging is a potent method which lowers variance as well as boosts stability in machine learning algorithms. Bagging reduces problems like overfitting and handling missing variables by using bootstrapping to create numerous subsets of the training data and training different models on each subset. To create an ensemble model, the predictions come from each of the distinct models are aggregated applying methods like voting the majority or averaging. This ensemble model demonstrates improved performance and resilience. It is created by combining many model predictions. Bagging determines useful technique for enhancing the precision as well as dependability of the algorithms, offering better durable answer for classification problems [14].

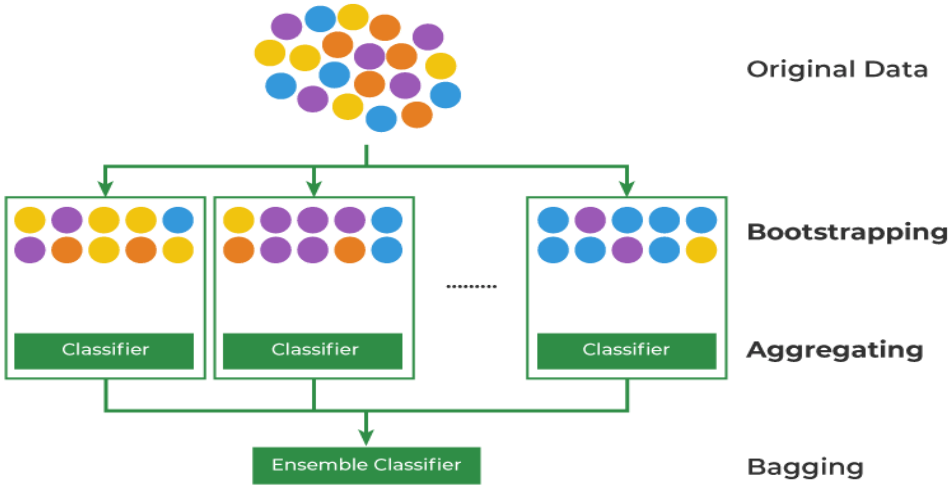


Figure 3.4.8 Bagging Classifier

Boosting Classifier

Boosting is a potent method that combines many algorithms using a weighted average to make weak learners stronger and improve the accuracy of independent models. The creation of loss functions that direct each model's learning process is the main goal of this method.

The algorithm's repetitive nature serves as an illustration of the boosting principle. In this work, utilized the boosting approach in both the testing and training stages to build a hybrid model which capitalizes on the advantages of each distinct model. The following represents the equation is used to boost algorithm, which reflects the model creation's iterative nature. By iteratively modifying the weights and integrating the calculation of several weak learners can turn into a more reliable and accurate ensemble model, boosting allows us to significantly increase the accuracy and performance of our models [26].

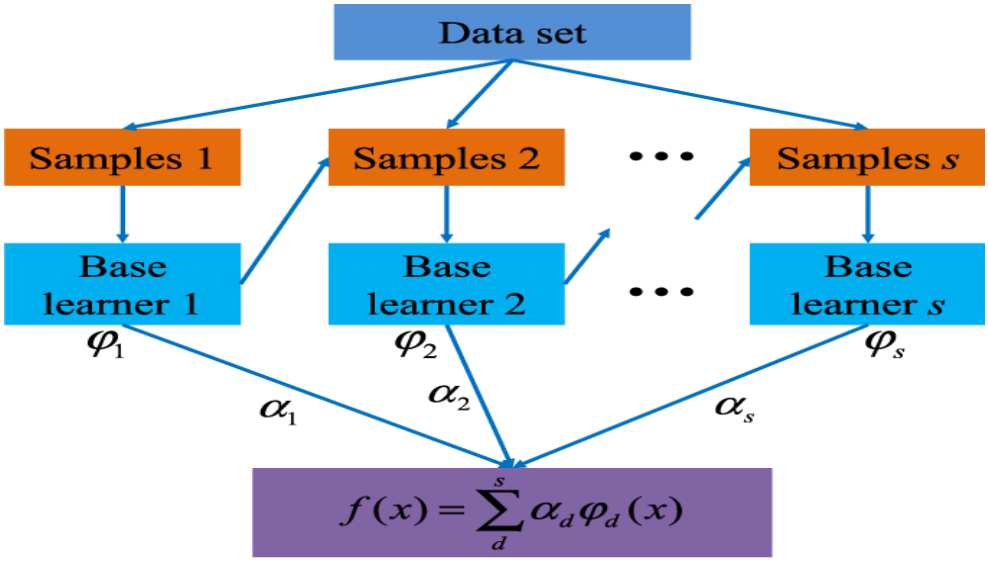


Figure 3.4.9 Boosting Classifier

3.5 Statistical Analysis

A variety of criteria are familiar to assess the effectiveness of machine learning classification models while evaluating basic approaches. The evaluation criteria include the following: precision, accuracy, recall, F1-score, and confusion matrix (CM).

Accuracy:

When it indicates the classification model of the accuracy it is called as one metric. Projected and estimated percentage of the particular model indicates its accuracy.It can be expressed as the proportion of accurately classified images to total samples.

$$Accuracy=(TP+TN)/(TP+TN+FP+FN) \quad (1)$$

Precision:

Positive label's likelihood and quantity make up accuracy. The accuracy of positively anticipated situations is shown by the precision. When FP is more essential than FN, precision is helpful. Mathematically speaking, this equation (1):

$$Precision=TP / (TP+FP) \quad (2)$$

Recall:

Sensitivity and Recall quantify the quantity of positively predicted outcome which are accurately classified. When recall become higher than FP, it is a valuable metric. The F1-score is an additional classification accuracy measure that takes precision and recall into consideration. Since accuracy and recall are harmonic means, the F1 score can offer a thorough comprehension of these two metrics. When Precision and Recall are equal, it performs optimally. To compute it, apply the below formula:

$$Recall=TP/(TP+FN) \quad (3)$$

F-1 Score:

Measuring recall-precision of categorizing accuracy is the F1-score. Which provides a throughout view. They are optimal when recall and accuracy are equal. Method:

$$F\text{-}1\ Score=2*(Precision*Recall)/(Precision+Recall) \quad (4)$$

Specificity is the proportion of patients with a negative test result who do not have the illness. It is helpful to exclude the majority of people who are not ill using a very precise test. This is the reason a good result on an incredibly specific test might exclude a certain person's illness. The equation is as follows:

$$Specificity =TN/((TN+FP)) \quad (5)$$

The model's capacity to generalize is limited by how closely it can resemble the training set of data. A specific table format known as the confusion matrix (CM) can be used to

evaluate the performance of an algorithm. Critical prediction requirements including accuracy, precision, specificity, and memory are highlighted by CM. Confusion matrices are especially helpful since they provide quick evaluations of variables such as TP, FP, TN, and FN. Based on the research questions, the subsequent sections respond to the research questions of this investigation.

3.6 Implementation Requirement

For the research project "Machine Leaning Based Hybrid Technique to Predict Hypothyroid for Combined Data" to be implemented effectively, a number of requirements must be satisfied. These implementation requirements encompass not just hardware and software components, but also activities such as data collect, model evaluate

Hardware Requirements:

- **Advanced computing resources:** the availability of a computer infrastructure with enough memory and processing power to handle the difficult tasks involved in data analysis and machine learning. **Graphics Processing Unit (GPU):** Accelerating the training of hybrid model with a GPU is crucial to cut down.

Software Requirements

- **Frameworks of DL:** Use popular deep learning frameworks for machine model building, training, and assessment, such as TensorFlow.
- **Python Programming Language:** Utilize Python's ubiquity and versatility to build machine learning algorithms and do data analysis.

Data Collection:

- **Diverse and Representative Dataset:** Assemble an extensive combine of medical data that includes both healthy and sick instances in order to ensure the models that are built are stable and applicable.
- **Balance Dataset**: Ensure that labeled data, clearly balanced to show whether hypothyroid illness is present or absent, is available for model testing and training.

Model Training and Evaluation

- **Machine Learning Models:** It is possible to use machine learning frameworks by utilizing models, like those acquired from libraries. This approach uses the most recent data to enhance the diagnosis of hypothyroid illness.

- **Evaluation Metrics:** For both prediction and evaluation tasks, establish and use relevant measures, such as accuracy, recall, precision, and F1 score, to evaluate the performance of the model.

Ethical Considerations:

- **Hospital confidentiality and data protection:** Compliance with legal and ethical standards to safeguard the confidentiality and integrity of healthcare data utilized for research.

- **Informed consent:** where appropriate, getting the informed consent of the people whose medical data are part of the collection.

Documentation and Reproducibility:

- **Code Documentation:** Complete documentation of the implementation code to ensure consistency, repeatability, and collaboration in the future.

- **Version Control:** use version control systems (such as Git) to monitor changes to the codebase and maintain an organized development process.

The study may be conducted systematically and comprehensively by meeting these implementation requirements, ensuring the correctness of the results and expanding the application of machine learning to diagnose hypothyroid illness.

CHAPTER 4

Experimental Result

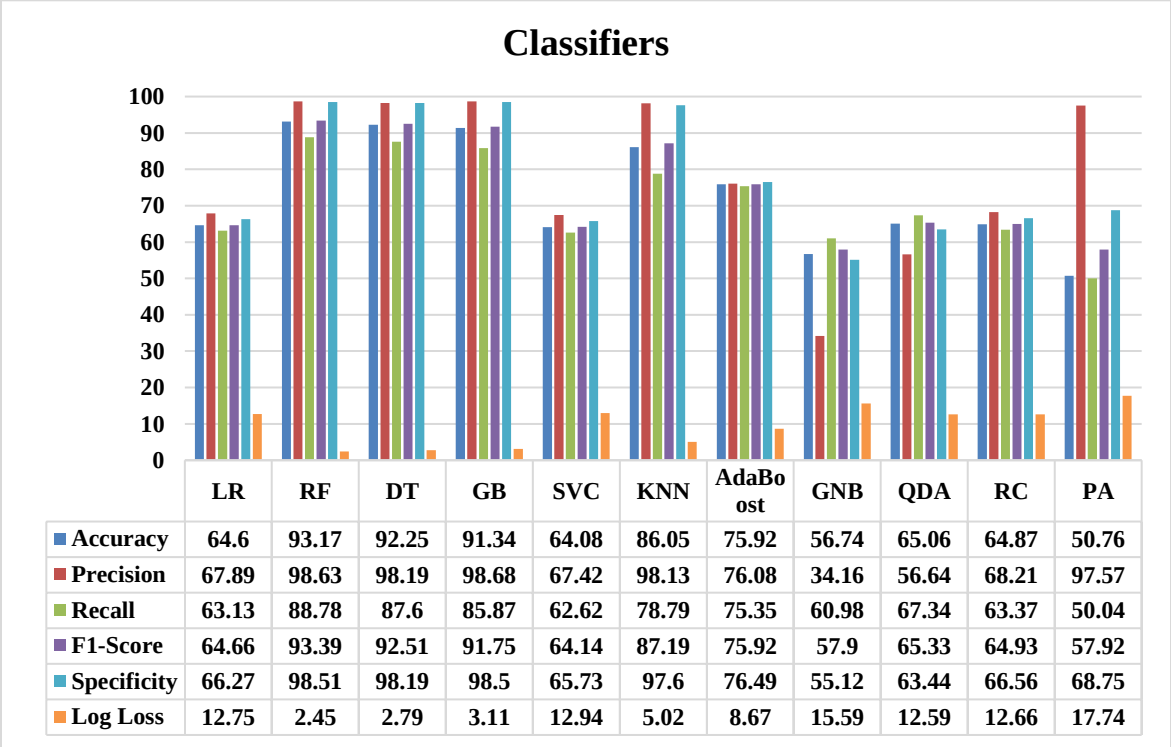
4.1 Experimental Setup

In order to conduct comprehensive investigations, the experimental setup for the study "Machine Leaning Based Hybrid Technique to Predict Hypothyroid for Combined Data" uses a carefully thought-out configuration of hardware, software, and data resources. High-performance hardware, such as a powerful graphics processing unit (GPU) and a robust central processor unit (CPU), comprise the computing core. This work used a training-and-testing-based supervised learning approach. Using the training dataset, which the algorithm used to identify patterns and correlations in the data, build the classification of the model. The dataset for testing was then subjected to the trained model's application in order to forecast results or categorize fresh occurrences. The following sections will go into more detail on the particular machine-learning algorithm that was used in this study. Throughout the experiment, a great deal of documentation is maintained, including code details, model architecture, hyperparameters, and conclusions, with the goal of encouraging future collaboration and transparency in the field of advanced hypothyroid illness prediction.

4.2 Experimental Results & Analysis

Using the chosen dataset, the effectiveness of the suggested model targeting hypothyroidism was evaluated in this part of the study by a critical analysis of previous models. The procedure started with the original deployment of the selected dataset, which was then subjected to a thorough analysis to find and correct any incorrect or missing point of data, thereby guaranteeing the dataset's accuracy. After that, a wide variety of machine learning algorithms were implemented and their results were carefully examined. Confusion matrices were utilized to do a thorough evaluation of the suggested algorithms. This evaluation included important metrics including Accuracy, F1Score, Recall, and Precision which gave a full picture of the algorithms' prediction skills. Furthermore, conventional algorithms were examined in the same way, allowing for a more thorough comparison. The assessment went on to investigate the possibilities of other ensemble methods, including boosting and bagging to take use of the combined advantages of several

models for improved prediction accuracy. Seven different classical classifiers were used in all, and the results obtained from them, when carefully evaluated, helped determine which methods were the most successful in predicting hypothyroid illness. This thorough assessment procedure was essential for determining the suggested model's effectiveness and optimizing its prediction accuracy for real-world use.



	LR	RF	DT	GB	SVC	KNN	AdaBoost	GNB	QDA	RC	PA
Accuracy	64.6	93.17	92.25	91.34	64.08	86.05	75.92	56.74	65.06	64.87	50.76
Precision	67.89	98.63	98.19	98.68	67.42	98.13	76.08	34.16	56.64	68.21	97.57
Recall	63.13	88.78	87.6	85.87	62.62	78.79	75.35	60.98	67.34	63.37	50.04
F1-Score	64.66	93.39	92.51	91.75	64.14	87.19	75.92	57.9	65.33	64.93	57.92
Specificity	66.27	98.51	98.19	98.5	65.73	97.6	76.49	55.12	63.44	66.56	68.75
Log Loss	12.75	2.45	2.79	3.11	12.94	5.02	8.67	15.59	12.59	12.66	17.74

Figure 4.2.1: Comparison of classifiers for Hypothyroid Disease Prediction

The evaluation of various classifiers reveals significant differences in performance across multiple metrics, including accuracy, precision, recall, F1-score, specificity, and log loss. Random Forest (RF) emerges as the best performer with an accuracy of 93.17%, high precision (98.63%), recall (88.78%), and a low log loss of 2.45, indicating both high predictive performance and robustness. Decision Trees (DT) and Gradient Boosting (GB) also show strong performance, with accuracies of 92.25% and 91.34% respectively, and similarly high precision and specificity. K-Nearest Neighbors (KNN) demonstrates a good balance with an 86.05% accuracy but slightly higher log loss (5.02). On the other hand, algorithms like Logistic Regression (LR), Support Vector Classifier (SVC), and Ridge Classifier (RC) exhibit moderate performance with accuracies around 64-65%, and higher

log losses around 12-13, reflecting less reliable predictions. AdaBoost and Quadratic Discriminant Analysis (QDA) perform moderately, whereas Gaussian Naive Bayes (GNB) and Passive Aggressive (PA) classifiers perform the worst, particularly PA with a high log loss of 17.74, highlighting issues with prediction stability and accuracy. These results underscore the effectiveness of ensemble methods (RF, DT, GB) in handling complex datasets compared to simpler classifiers.

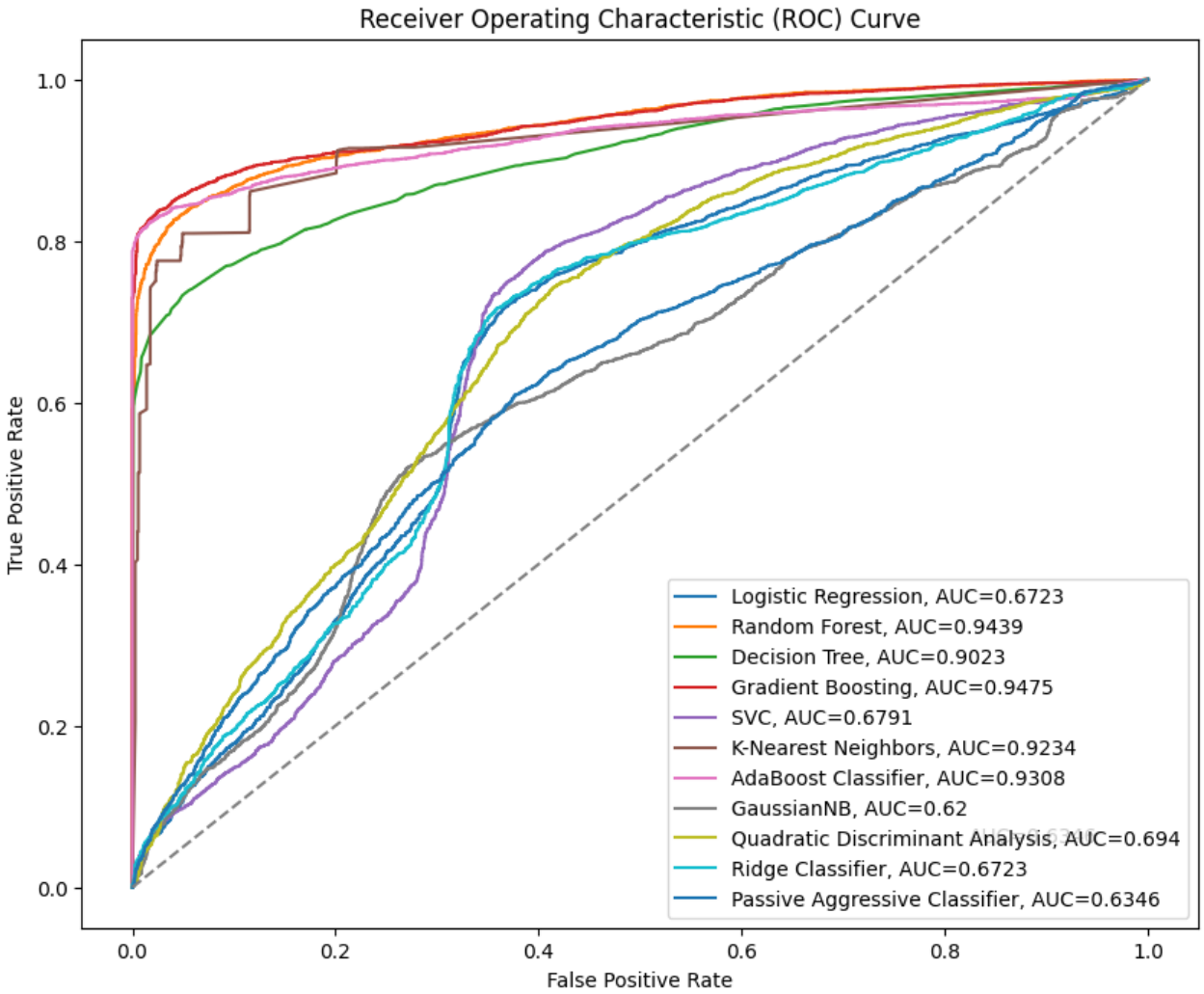


Figure 4.2.2: Classifiers model AUC-ROC evaluation

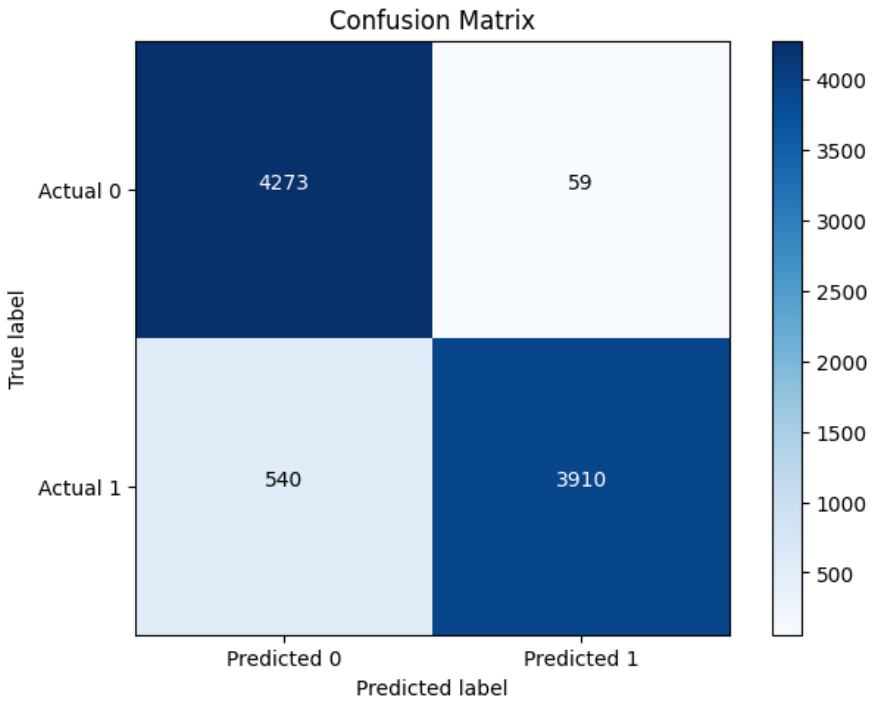
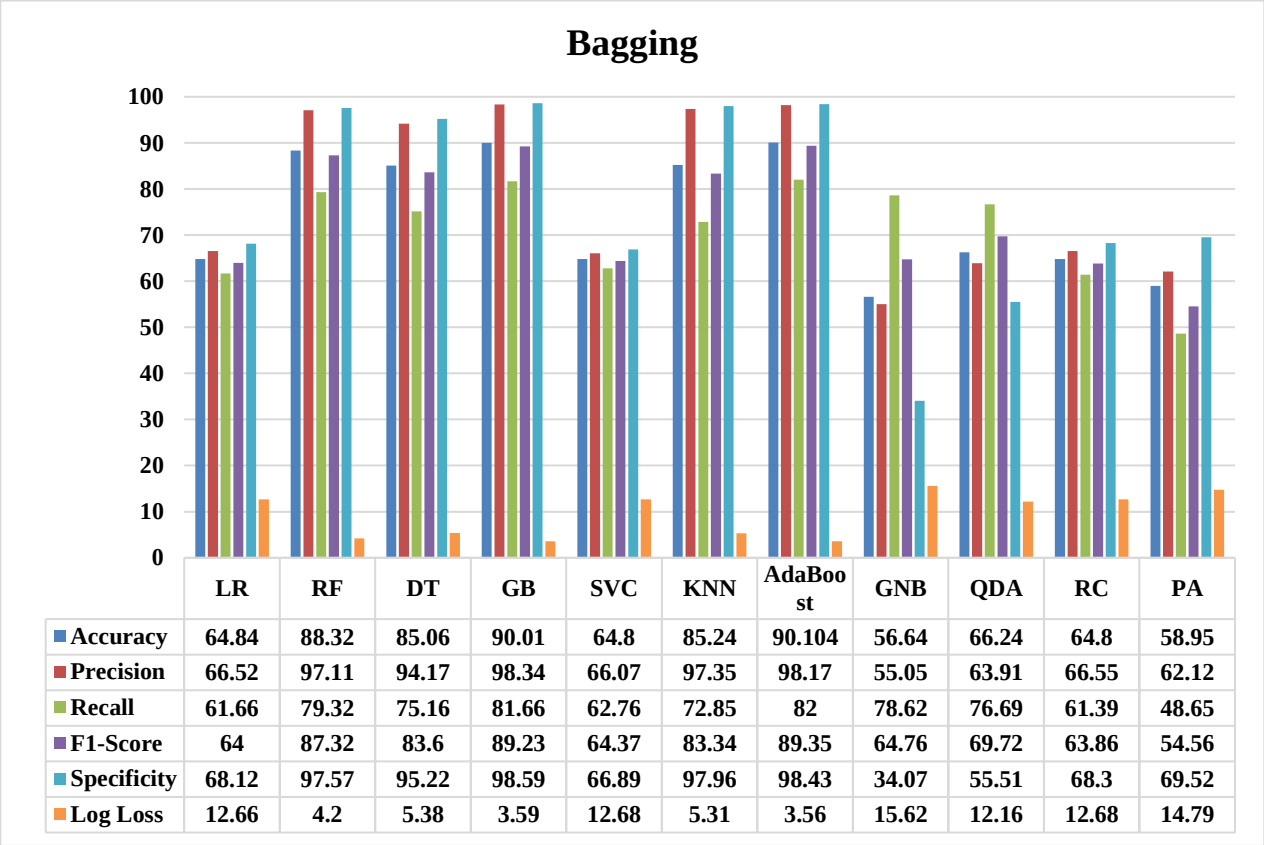


Figure 4.2.3 Confusion Matrix of RF

Random Forest showed very impressive performance in training and validation accuracy reaching 87.97% accuracy and which indicates it correctly identified 3910 positives, 4273 negative instances while it made 59 and 540 mistakes for false positives and false negatives.

Hybrid Bagging Result analysis:



	LR	RF	DT	GB	SVC	KNN	AdaBoost	GNB	QDA	RC	PA
Accuracy	64.84	88.32	85.06	90.01	64.8	85.24	90.104	56.64	66.24	64.8	58.95
Precision	66.52	97.11	94.17	98.34	66.07	97.35	98.17	55.05	63.91	66.55	62.12
Recall	61.66	79.32	75.16	81.66	62.76	72.85	82	78.62	76.69	61.39	48.65
F1-Score	64	87.32	83.6	89.23	64.37	83.34	89.35	64.76	69.72	63.86	54.56
Specificity	68.12	97.57	95.22	98.59	66.89	97.96	98.43	34.07	55.51	68.3	69.52
Log Loss	12.66	4.2	5.38	3.59	12.68	5.31	3.56	15.62	12.16	12.68	14.79

Figure 4.2.4: Bagging model evaluation

The evaluation of various classifiers, including both bagging and non-bagging methods, shows diverse performance metrics across accuracy, precision, recall, F1-score, specificity, and log loss. Gradient Boosting (GB) and AdaBoost stand out with the highest accuracy rates of 90.01% and 90.104% respectively, along with impressive precision and low log loss values, indicating strong overall performance. Random Forest (RF) also performs well with an accuracy of 88.32% and high specificity, although its recall is slightly lower, reflecting a balance in predictive capabilities. Decision Trees (DT) and K-Nearest Neighbors (KNN) offer solid accuracy around 85%, but their higher log loss indicates less stability in their predictions compared to GB and AdaBoost. Logistic Regression (LR), Support Vector Classifier (SVC), Ridge Classifier (RC), and Quadratic Discriminant Analysis (QDA) show moderate performance with accuracies in the mid-60s and higher log losses, suggesting limitations in handling complex data. Gaussian Naive Bayes (GNB) and Passive Aggressive (PA) classifiers underperform with the lowest accuracy rates, particularly GNB with a high log loss of 15.62,

highlighting its struggle with the dataset. Overall, the results underscore the superior performance of ensemble methods like GB and AdaBoost in delivering reliable and accurate predictions.

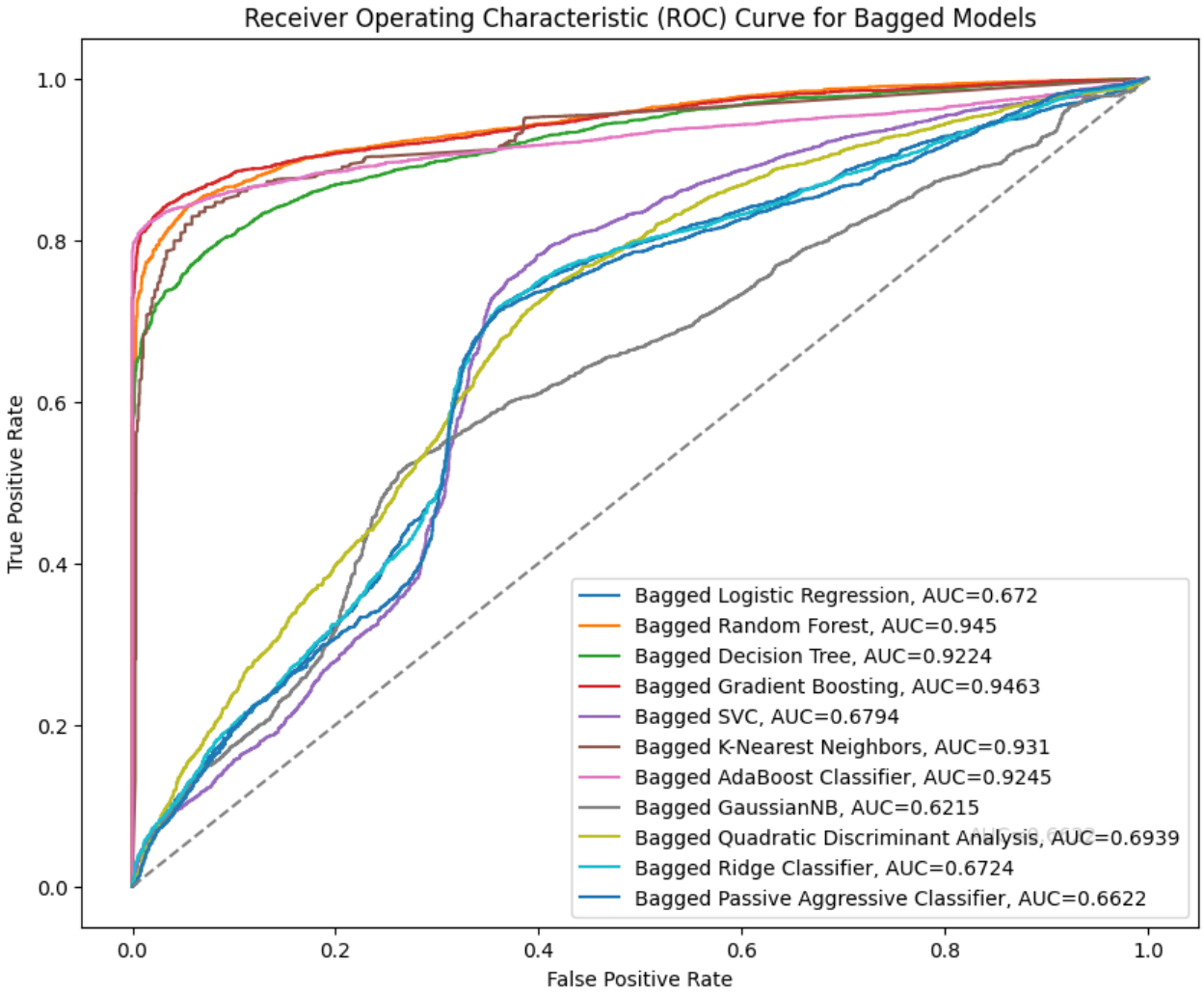
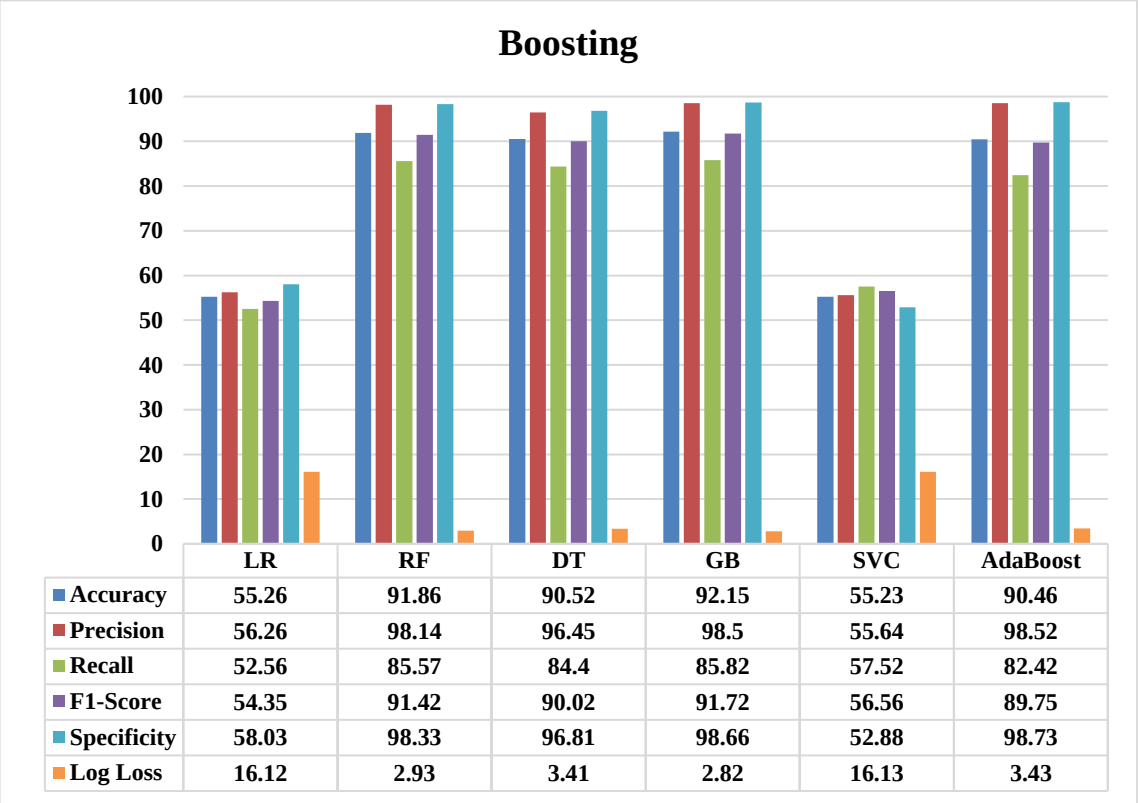


Figure 4.2.5: Bagging model AUC-ROC evaluation



	LR	RF	DT	GB	SVC	AdaBoost
Accuracy	55.26	91.86	90.52	92.15	55.23	90.46
Precision	56.26	98.14	96.45	98.5	55.64	98.52
Recall	52.56	85.57	84.4	85.82	57.52	82.42
F1-Score	54.35	91.42	90.02	91.72	56.56	89.75
Specificity	58.03	98.33	96.81	98.66	52.88	98.73
Log Loss	16.12	2.93	3.41	2.82	16.13	3.43

Figure 4.2.6: Boosting model evaluation

The evaluation results for various boosting classifiers indicate robust performance in terms of accuracy, precision, recall, F1-score, specificity, and log loss. Gradient Boosting (GB) emerges as the top performer with an accuracy of 92.15%, high precision at 98.5%, and a commendably low log loss of 2.82, suggesting highly reliable predictions and minimal errors. Random Forest (RF) follows closely, exhibiting strong performance metrics with 91.86% accuracy, 98.14% precision, and a log loss of 2.93. Decision Trees (DT) and AdaBoost also demonstrate high accuracies of 90.52% and 90.46%, respectively, with AdaBoost showing slightly better precision (98.52%) and specificity (98.73%) than DT. Both methods, however, have slightly higher log losses compared to GB and RF, indicating minor reductions in prediction stability. On the other hand, Logistic Regression (LR) and Support Vector Classifier (SVC) exhibit lower performance, with accuracies around 55%, higher log losses above 16, and generally lower precision and specificity scores. These results underline the superiority of ensemble methods like Gradient Boosting and AdaBoost in handling complex data and providing accurate, stable predictions.

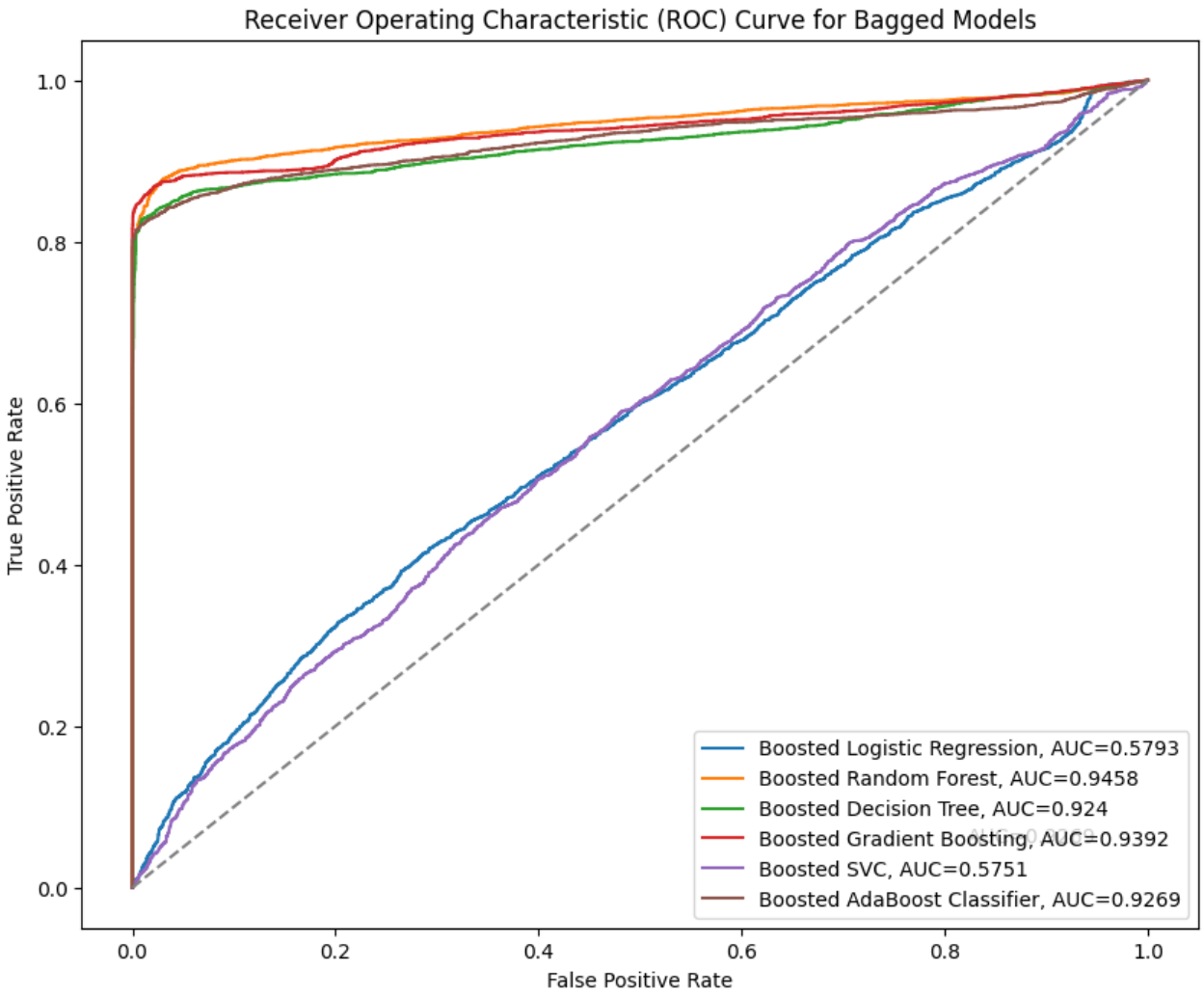


Figure 4.2.7: Boosting model AUC-ROC evaluation

The best performer in boosting is adaboost and it showed immersive performance in the ensemble with an accuracy of 90% which means it identified 3668, 4277 instances and to do that it made 55 true negative and 782false positive mistakes accordingly

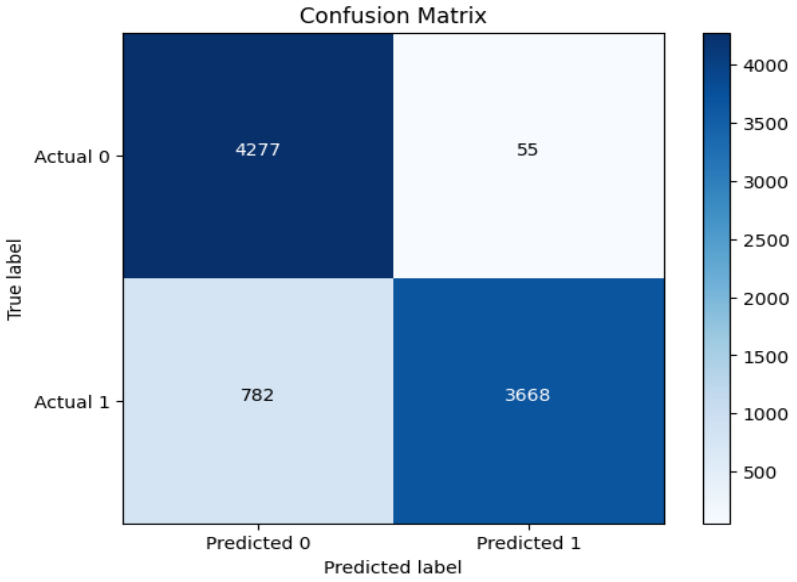


Figure 4.2.8 Confusion Matrix of Boosted ABC

4.3 Discussion

The Bagging method lowers variance and improves overall model stability by combining predictions from several base learners. By giving misclassified cases a larger weight, the technique known as "boosting" aims to repeatedly improve the model's accuracy. The most accurate classifier was the traditional technique, which achieved an impressive accuracy rate of 93.17% by Random Forest (RF). Other classifiers that were used included Gradient Boosting (GB), Support Vector Machine (SVM), Passive Aggressive (PA), ,K-Nearest Classifier (KN),Decision Tree (DT),Ridge Classifier (RC),Quadratic Discriminant Analysis (QDA),Adaboost Classifier (ABC),Gaussian Naïve Bayes (GNB), Logistic Regression (LR). In addition, 92.15% accuracy was obtained by the Boosting Gradient Boosting (GB), while Boosting Random Forest (RF) 91.86% accuracy was attained. Hyperparameter tweaking was used to maximize each classifier's performance. After conducting an experimental examination and reviewing prior research, it was determined that the Random Forest (RF) classifier performed very well, correctly diagnosing hypothyroid illness with an astounding accuracy rate of 93.17%.

CHAPTER 5

Impact on Society

5.1 Impact on Society

There are a lot of important advantages to our suggested approach, both in terms of the economy and society. Our model, which is based on an actual dataset, was carefully designed to look into and identify the essential elements and traits of people who have hypothyroidism. Numerous social benefits flow from this research, the most important of which is its ability to inform the public about the incidence of hypothyroidism and the possible preventative strategies. Because of our model's accurate diagnosis and consistent monitoring, therapy may be started earlier and people are better equipped to make educated healthcare decisions and foresee future problems. Notably, our method's simplified and efficient nature drastically cuts down on time and computational needs, making illness prediction remarkably accurate while remaining simple.

Our thorough data analysis uses cutting-edge diagnostic methods to identify the underlying causes of hypothyroidism. We hope to see our suggested course of action widely adopted and put into practice on a social level. Our objective is to foster a more educated and health-conscious society via the dissemination of knowledge and the promotion of proactive healthcare practices. The ultimate objective is to enable people to take responsibility for their health and well-being, therefore reducing the effects of hypothyroidism and other health-related issues. In conclusion, our approach presents a viable path for improving public health and healthcare awareness generally, as well as for accurate illness prediction.

5.2 Impact on Environment

It’s our belief that our paradigm has outstanding potential for underserved and distant places, providing time- and complexity-saving streamlined diagnostic techniques. Because of its simplicity and lack of invasiveness, the environment will gain from it without suffering any negative consequences. Our concept makes healthcare easier and more accessible by eliminating the need for those living in distant areas to come to metropolitan areas to find out if they have hypothyroidism. In addition to predicting anticipated results,

this predictive model may be used to seamlessly supplement a patient's diagnostic report, removing worries around the expense of prediction of hypothyroidism or the cost of local therapy. Because of its intuitive design, users of all ability levels may easily use it.

The implementation of our proposed method would enable the diagnosis of hypothyroidism in patients, a development that will significantly alter the social and political context of healthcare. We are certain that putting our recommended approach into practice will result in a major advancement in technology of medical science, finally enhancing the caliber of medical technology.

In conclusion, even if the research's main participation is to the advancement of medical diagnostics, it is nevertheless important to take the environment into account. As long as efforts are made to achieve a balance between technological innovation and environmental sustainability, health care advancements will continue to be socially and environmentally responsible.

5.3 Ethical Aspects

It is crucial to adopt ethical measures before putting our system into use to avoid the accidental release of personal data, diagnostic results, or inadvertent humor. We believe that our suggested method will be helpful for the diagnosis and treatment of hypothyroidism, both in the actual world and in future research projects, as this problem is worldwide in nature. Our approach provides a useful tool for proactive healthcare management globally by enabling people, whether they are patients or well-informed folks, to predict the development and evolution of their hypothyroid disease.

5.4 Sustainability Plan

We are certain that our suggested paradigm will work flawlessly with the current global technological environment for the diagnosis and study of hypothyroid disease. We are certain that our suggested course of action will be quite beneficial to ladies who are susceptible to hypothyroidism. We are excited to provide our help to underprivileged rural communities. Our suggested paradigm is intended to be long-lasting and useful, having a significant effect on the management of hypothyroidism and healthcare accessibility.

CHAPTER 6

Summary, Recommendation,

Conclusion, and Implication for Future Research

6.1 Summary of the Study

In the insightful article, we use algorithms to assess people's potential for effect. Our model provides a trustworthy way to predict future changes. The diagnostic approach used has great potential as it allows one to forecast a person's prospective impact. This insight is helpful not just for those who may erroneously think that they should be cognizant of their hypothyroidism, but also for comprehending the many phases of hypothyroidism. Our suggested approach brings its benefits to the domain of medical diagnostics, supported by the use of proven, quick-to-implement, low-training, and very accurate algorithms. This comprehensive strategy fosters a greater awareness of health-related issues and their potential effect, empowering both individuals and healthcare providers.

6.2 Conclusions

Modern society, which is distinguished by its harmonious fusion of cutting-edge technology and simplicity, makes access to cutting-edge inventions almost widespread. Our approach takes use of this technology environment to provide a surprisingly quick and simple method for predicting hypothyroid disease in individuals. We are dedicated to guaranteeing the viability and ongoing improvement of the model, with designs to add more features and handle more comprehensive healthcare issues down the road. Our model has the potential to help people all around the world. This vision, based on the level of technology today, lays the groundwork for a dynamic and ever-changing approach to illness prediction and healthcare.

6.3 Implication for Further Study

Being mortal is a natural element of being human, and we deal with a variety of ailments on a daily basis. Many of us suffer with hypothyroidism, yet some of us are equipped to fight it. Even though we live in a developing country, we have access to sophisticated and

accurate medical diagnosis and treatment technology. These developments have expedited and improved the time and efficacy of the hypothyroid disease diagnosis procedure. We hope that others will adopt our strategy in our endeavor to offer innovation. In order to promote advancements in healthcare and illness management, we have consistently improved the efficacy of our current algorithms and are dedicated to growing our product line in the next years.

REFERENCES

[1] Mir, Y. I., & Mittal, S. (2020). Thyroid disease prediction using hybrid machine learning techniques: An effective framework. International Journal of Scientific & Technology Research, 9(2), 2868-2874.

[2] Pandey, S., Miri, R., & Tandan, S. R. (2013). Diagnosis and classification of hypothyroid disease using data mining techniques. International Journal of Engineering Research and Technology, 2.

[3] Ibrahim, I., & Abdulazeez, A. (2021). The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends, 2(01), 10-19.

[4] Ioniţă, I., & Ioniţă, L. (2016). Prediction of thyroid disease using data mining techniques. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 7(3), 115-124.

[5] Awujoola Olalekan, J., Ogwueleka, F., & Odion, P. (2020). Effective and accurate bootstrap aggregating (Bagging) ensemble algorithm model for prediction and classification of hypothyroid disease. Int. J. Comput. Appl, 975, 8887.

[6] Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalized medicine, 10(2), 21.

[7] Selwal, A., & Raoof, I. (2020). A Multi-layer perceptron based intelligent thyroid disease prediction system. Indonesian Journal of Electrical Engineering and Computer Science, 17(1), 524-532.

[8] Tyagi, A., Mehra, R., & Saxena, A. (2018, December). Interactive thyroid disease prediction system using machine learning technique. In 2018 Fifth international conference on parallel, distributed and grid computing (PDGC) (pp. 689-693). IEEE.

[9] Banu, G. R. (2016). Predicting thyroid disease using linear discriminant analysis (LDA) data mining technique. Commun. Appl. Electron.(CAE), 4, 4-6.

[10] Gothane, S. (2020). Data mining classification on hypo thyroids detection: Association women outnumber men. International Journal of Recent Technology and Engineering (IJRTE), 8(16), 601-604.

[11] Mustapha, R. A. B. I. (2019). An Enhanced Computational Integrated Decision Model for Prime Decision Making in Driving. University utara Malaysia, 7.

[12] Sindhya, K. (2020). Effective prediction of hypothyroid using various data mining techniques. EPRA International Journal of Research and Development, 5(2).

[13] Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Alamin Dhaly, Md Nour Hossain. “A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease”. In April 2022 The 10th International Conference on Computer and Communications Management in Japan.

[14] Aunik Hasan Mridul, Md. Jahidul Islam, Mushfiqur Rahman, Mohammad Jahangir Alam, Asifuzzaman Asif. “A Machine Learning-Based Traditional and Ensemble Technique for Predicting Breast Cancer”, In December, 2022. Conference: 22th International Conference on Hybrid Intelligent Systems (HIS 2022) online, 2022At: Auburn, Washington, USA.

[15] “Logistic Regression for Machine Learning”, Accessed: August 6, 2021, Available:

https://www.capitalone.com /tech/machine-learning/what-is-logistic-regression/

[16] Ghosh, Pronab, Asif Karim, Syeda Tanjila Atik, Saima Afrin, and Mohd Saifuzzaman. "Expert cancer model using supervised algorithms with a LASSO selection approach." International Journal of Electrical and Computer Engineering (IJECE) 11, no. 3 (2021): 2631

[17] emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn."Journal of Marketing Research43, no. 2 (2006): 276-286.

[18] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." ArtificialIntelligence Review 54, no. 3 (2021): 1937-1967.

[19] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." In International Conference on Machine Learning, pp. 51335142. PMLR, 2018

[20] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301

[21] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: theory and practice,” Neurocomputing, vol. 415, pp. 295–316, 2020

[22] H. M. Almahshi, E. A. Almasri, H. Alquran, W. A. Mustafa and A. Alkhayyat, "Hypothyroidism Prediction and Detection Using Machine Learning," 2022 5th International Conference on Engineering Technology and its Applications (IICETA), 2022, pp. 159-163, doi: 10.1109/IICETA54559.2022.9888736

[23] Alam, S., Afzal, K., Maheshwari, M., & Shukla, I. (2000). Controlled trial of hypo-osmolar versus World Health Organization oral rehydration solution. Indian pediatrics, 37(9), 952-960.

[24] Akgül, G., Çelik, A. A., AYDIN, Z. E., & ÖZTÜRK, Z. K. (2020).Hipotiroidi Hastalığı Teşhisinde Sınıflandırma AlgoritmalarınınKullanımı. Bilişim Teknolojileri Dergisi, 13(3), 255-268

[25] Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., & Mukherjee,S. (2016). A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSItransactions on ICT, 4(2), 313-319

[26] P. Duggal, and S. Shukla, “Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques”, 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp.670-675, 09 April 2020

[27] V. S. Vairale, and S. Shukla, “Classification of Hypothyroid Disorder using Optimized SVM Method”, Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019), pp. 258-263,10 February, 2020

[28] Riajuliislam, Md & Rahim, Khandakar & Mahmud, Antara. (2021). Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques. 60-64. 10.1109/ICICT4SD50815.2021.9397052

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title: Machine Leaning Based Hybrid Technique to Predict Hypothyroid for Combined Data

Student ID: 202-15-

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to identify a hypothyroid disease prediction problem for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**

CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing machine learning based hybrid techniques, and data processing and classification tasks. The project demonstrates **specialist knowledge (K4)** by conducting machine learning with advanced hybrid technique, enhancing hypothyroid disease prediction accuracy in medical data, crucial for computer-aided diagnosis.	**Page no:** [16-20] **Section:** [3.2] **Page no:** [1-9] **Section:** [1.1]
				The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice & technology (K6)** by employing hybrid model.	**Page no:** [27] **Section:** [3.2] **Page no:** [33]

					Section: [3.4]
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, to advance hypothyroid disease prediction using machine learning, showcasing a comprehensive understanding of current methodologies.	**Page no:** [18-22] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in machine learning disease prediction, including limitations of traditional methods and the complexities of integrating hybrid learning. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining diagnostic methodologies.	**Page no:** [22-23] **Section:** [2.4, 2.5]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting Machine Learning as the chosen significant solution for enhancing hypothyroid disease prediction amidst multiple potential approaches.	**Page no:** [39-45] **Section:** [4.2]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting medical diagnostics in respiratory diseases like hypothyroid disease, contributing to advancements in health and hospital care practices which indicates **EP-4**.	**Page no:** [18-22] **Section:** [2.2, 2.4]

5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in medical diagnostics which ensures **EP-7**.	**Page no:** [27-38] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, deep learning frameworks, feature datasets, and ethical considerations to ensure systematic research and contribute to advancements in hypothyroid disease prediction through machine learning approach.	**Page no:** [36-38] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving health care through advanced hypothyroid disease prediction methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for hospital data privacy.	**Page no:** [47-49] **Section:** [5.1, 5.2, 5.4]

5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in hypothyroid disease prediction through machine learning, demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [18-22] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [14-16] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing hospital privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced health care technologies which comply with **CO9**.	**Page no:** [48] **Section:** [5.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [27-38, 39-45] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

Thyroid.docx

ORIGINALITY REPORT

16%	15%	7%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	2%
3	www.scienceworldjournal.org Internet Source	1%
4	123dok.com Internet Source	1%
5	Submitted to University of Technology, Sydney Student Paper	<1%
6	thesai.org Internet Source	<1%
7	katalog.ub.uni-paderborn.de Internet Source	<1%
8	www.ijmer.com Internet Source	<1%
9	"Intelligent Computing for Sustainable Development", Springer Science and Business	<1%