Explainable AI Based Machine Learning Models for Thyroid Disease Prediction

BY

M. A. OMAR FARUQ
ID: 242-25-039

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

Supervised By

Dr. Md Alamgir Kabir
Assistant Professor & Coordinator, MIS
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor & Director, MSc. in CSE
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2025

APPROVAL

This Project/Thesis titled **"Explainable AI Based Machine Learning Models for Thyroid Disease Prediction"**, submitted by **M. A. Omar Faruq**, ID No: **242-25-039** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

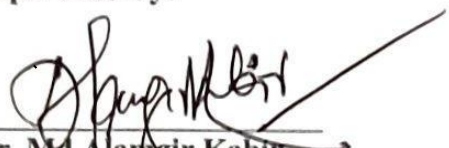
External Examiner

Mr. Nazibur Rahman
Head of IT Infrastructure
Networld Bangladesh PLC

DECLARATION

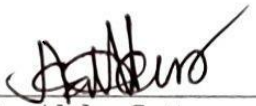
I hereby declare that this research has been done by me under the supervision of **Dr. Md Alamgir Kabir, Assistant Professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



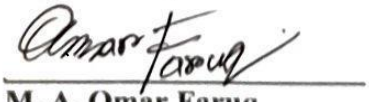
Dr. Md Alamgir Kabir
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:



M. A. Omar Faruq
ID: 242-25-039
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartfelt thanks and gratitude to Almighty Allah for His divine blessing, which makes it possible to complete the final year project/internship successfully.

I am grateful and wish to express my profound indebtedness to **Dr. Md Alamgir Kabir, Assistant Professor**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori**, **Head of the** Department of CSE, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

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

ABSTRACT

The incidence of thyroid disorders is among the most widespread endocrine diseases worldwide and accurate and timely diagnosis is an important part in patient management/procedures. The paper presents a proposal of an Explainable Artificial Intelligence (XAI) based solution for thyroid disease prediction which predicts using the different machine learning methods including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, CatBoost, and XGBoost. To address this challenge of class imbalance, pre-processing was carefully done to this data set in order to alter this issue as well as increase the representation as feature, and the performance has been evaluated in terms of precision, recall, accuracy, and F1-score. In comparison to all the models, XGBoost turned out to be the best model with accuracy 98.5% upholding classification of Hyperthyroid, Hypothyroid and Negative thyroid accurately. In order to help ensure clarity and confidence regarding the patient's clinical condition, a method that used SHAP XAI was utilized and interpretation of the trained model was derived into how and why the models arrived at such decisions, thus giving insight into features. This includes XAI which can be used to turn the predictive power into an explainable system to ensure that health practitioners understand the process of how powerful decisions are being made through a scoring rubric. All in all, this study demonstrates that the fusion of XGBoost with XAI techniques is an effective, correct and explainable method to detect thyroid disorders at the very earliest stage, which further assists in enforcing better diagnostic practices that improve patient outcomes.

Keywords: Explaninable AI (XAI), XGboost, LIME, Thyroid disease Detection

.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER	
CHAPTER 1: INTRODUCTION	**1-6**
1.1 Introduction	1-2
1.2 Motivation	3-4
1.3 Problem Statement	4
1.4 Research Questions	5
1.5 Research Objectives	5
1.6 Expected Output	5-6
1.7 Project Management and Finance	6
1.8 Report Layout	6
CHAPTER 2: BACKGROUND	**7-15**
2.1 Preliminaries	7-8
2.2 Related Works	8-10
2.3 Comparative Analysis and Summary	11-13
2.4 Research Gap	14
2.5 Challanges	14
CHAPTER 3: RESEARCH METHODOLOGY	**15-37**
3.1 Research Subject and Instrumentation	15
3.2 Dataset	15
3.2.1 ThyroidDF	17

3.3 Data Pre-Processing 19
3.3.1 Missing Value Handling 20
3.3.2 Attribute Dropping 22
3.3.3 Outlier Handling 22
3.3.4 Data Encoding 22
3.3.5 Feature Selection 24
3.3.6 Train Test Split 24
3.4 Exparimental Process 25
3.5 Machine Learning Algorithms 28
3.5.1 Logistic Regression 28
3.5.2 Decision Tree 29
3.5.3 Random Forest 30
3.5.4 Weighted Logistic Regression 31
3.5.5 Gradient Boosting 31
3.5.6 XGBoost 32
3.5.7 CatBoost 33
3.5.8 Explainable AI 34
3.5.9 LIME 34
3.5.10 SHAP 34
3.5.11 Application of Explainable AI 35
3.6 Implementation Requirements 36
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **38-58**
4.1 Experimental Setup 38
4.2 Metrics of Evaluation 39
4.3 Results & Analysis 41
4.3.1 XGboost Classifier 43
4.3.2 Decision Tree 44

4.3.3 Logistic Regression 46
4.3.4 Weighted Logistic Regression 48
4.3.5 Random Forest 50
4.3.6 Gradient Boosting 51
4.3.7 CatBoost Classifier 53
4.3.8 Best & Final Model 56
4.4 Discussion 57
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **59-60**
5.1 Impact on Society 59
5.2 Impact on Environment 59
5.3 Ethical Aspects 60
5.4 Sustainability Plan 60
CHAPTER 6: CONCLUSION AND FUTURE WORK **61-63**
6.1 Summary of the Study 61
6.2 Conclusions 62
6.3 Implication for Further Study 62
REFERENCES **64-66**

LIST OF FIGURES

FIGURES	PAGE NO
Fig 3.3: Count of target class instence	19
Fig 3.4: Data pre-processing steps	20
Fig 3.5: Proposed Methodology flowchart	27
Fig 3.6: Decision tree architecture	29
Fig 3.7: Random Forest Model Architecture	30
Fig 3.8: Gradient Boosting Model Architecture	32
Fig 3.9: XG Boost Classifier Architecture	33
Fig 3.10: XAI Explanation for a class decision	36
Fig 4.1: Confusion matrix of XGboost	43
Fig 4.2: Precision vs Recall curve of XGboost Classifier	44
Fig 4.3: Confusion Matrix of Decision Tree	45
Fig 4.4: Precision vs Recall curve of Decision Tree	46
Fig 4.5: Confusion matrix of Logistic Regression	47
Fig 4.6: Precision vs Recall curve of Logistic Regression	48
Fig 4.7: Precision vs recall curve of weighted Logistic Regression	49
Fig 4.8: Confusion matrix of Random Forest	50
Fig 4.9: Precision vs recall curve of Random Forest classification	51
Fig 4.10: Confusion matrix of Gradient Boosting	52
Fig 4.11: Precision vs Recall curve of Gradient Boosting	53
Fig 4.12: Catboost Confusion matrix (Under sampling)	54
Fig 4.13: Catboost Confusion matrix (Over-sampling)	55
Fig 4.14: Precision vs Recall curve of CatBoost (Under(left) & over(Right) Sampling)	56

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Comparative summary of related work	11-13
Table 3.1: Dataset description	16
Table 3.2: Data encoding map for target class	18
Table 3.3: List of Missing Value Count	21
Table 3.4: All data encoding map	23
Table 4.1: Results of all the trained model	42

CHAPTER 1

Introduction

1.1 Introduction

Thyroid disorders like hypothyroid or Hyperthyroid is an important health problem that the world is facing as millions of people get affected by each year. The thyroid is important in metabolism functioning and its malfunctions are dangerous in terms of health. Indeed, it has been shown through research that thyroid diseases particularly hypothyroidism is common among women and it may cause complications particularly cardiovascular diseases, depression, and sub fertility if not addressed. The treatment is dependency on early discovery and correct diagnosis - treatment may include the use of hormone replacement therapy in aim to restore normal thyroid functions.

The current advancements in the areas of machine learning (ML) has redefined the medical sector and enables to come up with more efficient and accurate diagnostic system. In terms of forecasting the thyroid disease, the decision trees, support vector machines (SVM) and k-nearest neighbours (KNN) have achieved an ubiquity of ML algorithms used to assess the various characteristics of a patient like age, gender and the thyroid hormone levels. However, the interpretability and transparency of the conventional models of ML are severely limiting and, therefore, it is an issue in the healthcare sector where credibility and accountability are given prime importance.

XAI is one of the most important ways of mitigating such issues. XAI methods are aimed to render machine learning to be more comprehensible and industrial to the medical fraternity. Such transparency is essential in earning AI-based systems trust, particularly in the high-stakes field, such as health care diagnosis. Thanks to the realisation of XAI methods, the capacity to achieve clinically-relevant features, interpret predictions made by the model, and determine the applicability in the clinics can be improved to a greater extent.

This thesis will endeavour to integrate XAI within existing machine learning systems in order to make thyroid disease prediction systems more accurate and transparent. The major research gap to be addressed by this article is that no interpretable AI-related solution to the prediction of thyroid diseases was provided thus far, and it may hinder the process of the implementation of AI in medical practice. By integrating XAI into this paper, this line of research will seek not only put forward precise predictions but also explainable prediction models, as the medical staff can base its decisions on the knowledge produced by an AI system.

The main objectives of this thesis are:

- To develop a thyroid disease prediction model using advanced machine learning algorithms, with a particular focus on XGBoost for its superior predictive performance.

- To integrate Explainable AI techniques into the model to enhance its interpretability and ensure transparency in the decision-making process.

- To evaluate the performance of the proposed system in terms of prediction accuracy and the ability to provide clear, understandable reasoning behind predictions.

Through this research, the integration of XAI aims to not only improve the diagnostic accuracy of thyroid disease prediction systems but also ensure that these systems are comprehensible and trustworthy for healthcare providers, ultimately contributing to better patient outcomes.

1.2 Motivation

Thyroid disorders constitute one of the most common endocrine health issues globally with millions of people being diagnosed each year with the disease by way of the non-specific symptoms which makes it quite hard to diagnose it at the right time in most cases. Other disease like high and low thyroid state, when not identified, may result in high-extreme metabolic changes, heart disorders and low quality of life thus effective and early detection is of paramount importance Convening diagnostic procedures, especially in cases that are asymptomatic and early-stage procedures are reliant on clinical examination, and biochemical tests of TSH, T3, and T4 although efficient are prone to interpretation errors, time lag, and mis-diagnosis, especially in early stages [5].

Machine learning (ML) in recent years is marking its presence as a revolutionary method in medical diagnostics due to its high accuracy in disease prognosis because it learns intricate and non-linear associations within multidimensional databases . Nevertheless, numerous top performing ML applications in thyroid disease are still black-box models, in the sense that they provide predictions without any explainable rationale. This lack of transparency reduces the trust that clinicians place in it, reduces its adoption in evidence-based care, and complicates its regulatory acceptance in safety-critical areas, such as healthcare [7]. Moreover, irrelevancy of features, redundancy of features, unbalanced data, and explainability issues in models compromise their performance and, as a result, their generalizability and the associated bias in decision-making has negative impact [1].

Explainable machine learning (XAI) is a new set of approaches to overcoming these drawbacks, by embedding transparency into prediction modeling to allow practitioners to analyse feature relevance, benefit, and decision justifications. Within the context of the thyroid disease, XA not only helps improve the diagnostic reliability but encourages the advent of informed clinical decision-making, patient-collaboration, and the adherence to the principles of developing ethical and legal practice of AI in healthcare.

The motivation of this research is the gap which exists between the predictive level and the interpretability of thyroid diagnose. It aims to deliver state-of-the-art, transparent models

based on a synergy of state-of-the-art machine learning algorithms, carefully built feature selection and explainable AI models. These models can help clinicians in the early diagnosis of the malfunction of the thyroid, reducing the circular of the misdiagnosis, improving the treatment's intervention, and eventually patient outcome. Through it, the study would be in line with the common healthcare goals of precision medicine, cost effectiveness, and trust-based machine integration in healthcare processes.

1.3 Problem Statements

Thyroid disorders are a widely spreading health problem that affects millions of people around the globe. Hyperthyroidism and hypothyroidism are the two general forms of these disorders that may result to severe problems when not detected and treated properly in time. Diagnosis of the thyroid-related disorders occurs by relying on clinical tests and the determination of doctors. Although such an approach can also be rather beneficial, there is still always a risk of failure or delay in diagnosing a condition. The question, however, is how to interpret the results of the test as a complex issue. Although a certain level of progress has been made in the field of ML application in medical diagnostics, the apparently general problem of interpretability of AI-driven prediction is still noted. This research paper attempts to implement Explainable AI (XAI) so that the decision-making process of AI models can be explained and be justified. Transparency of this kind will be necessary where the AI models and systems used require trust among healthcare practitioners, physicians in particular, as they rely on the predictions made by the system to make the necessary decisions and to speed up the diagnosis query.

1.4 Research Question

What are the ways in which Explainable AI (XAI) can be applied to a machine learning model to achieve higher interpretability and accuracy of thyroid disease predictions, and how can it assist medical professionals directly in the clinical decision-making process?

1.5 Research Objectives

- Developing a machine learning model that can correctly predict thyroid disease with a very large dataset, which contains major clinical characteristics like hormone levels or health history.
- Adopt specific XAI approaches such as SHAP or LIME to be able to convey the transparency into the process of making the predictions, providing for the clear understanding of what aspects contributed to the outcomes.
- To understand how well does the machine learning model work both in terms of accuracy and transparency compared to the traditional ways of the diagnosis, and also to learn what kind of performance is given to the AI model and the point of improvement or otherwise the limitation it encounters.
- To translate theoretical conclusions drawn from the XAI-model to an understanding by medical professionals not only of the XAI-model and XAI-model predictions but also how to verify and substantiate them in real-life clinical practice.

1.6 Expected Output

The goal of this project is to create a thyroid disease prediction model capable of yielding a highly accurate result whilst also explaining the reasoning behind every prediction in a manner that is easily understood by healthcare professionals. By adding XAI to the model, we hope to increase the level of trust and so that the medical team can make more informed decisions. This work has the potential to present a working solution to establishing better early diagnosis, minimizing errors, and ultimately raising the patient outcome in thyroid-related conditions. This study can also be used in the future as the baseline of AI-based diagnostic systems more focused on both performance and transparency.

1.7 Project Management and Finance

This study was not funded/sponsored. The entire implementation process of the projects such as data collection, preprocessing, model development and evaluation was done with the use of personal resources through the available facilities in the institution. The lack of dedicated funding meant that planning and effective use of budgeted computational and software resources had to be exercised in order to complete the research in a timely manner. Project management concentrated on ensuring a systematic flow of work, compliance with the schedule of the research, and the maximization of the use of resources, which subsequently allowed fulfilling all tasks, regardless of the reduced funding..

1.8 Report Layout

Chapter 1 has the introductory part, research objectives, and research questions of the study. Chapter 2 will have a literature review explained in detail. Chapter 3 explains the methodology, what is going to be the framework and processes used in the research. Chapter 4 presents and discusses experimental outcomes. Chapter 5 discusses the sustainability plan, the implications of the society, and the environmental implications of the work along with the ethical reasoning on the same. Lastly, Chapter 6 brings the study to an end, and points to possible future research avenues.

CHAPTER 2

BACKGROUND

2.1 Preliminaries

In this research project, a few areas are prioritised including the diagnosis of the thyroid disease and the classification of the thyroid problems as well. The ideas are in the guaranteed fast and accurate detection, which will lead to early stage diagnosis. Neck ultrasonography is a popular tool in healthcare that is applied to diagnose thyroid nodules as one of the main symptoms of endocrine diseases caused by disordered growth of thyroid cells. Patient conditions can be identified more efficiently with the help of the proposed model. Through the training of the system to a wide variety of patients, we are able to match incoming information on new patients to what was already observed and which data improves the accuracy of detecting thyroid disorders.

This model is projected to label three types of thyroid states hyperthyroidism, hypothyroidism and normal (negative). Hyperthyroidism is caused by excessive production of hormones by the thyroid gland and hypothyroidism by lack of adequate production of the hormones. A regular state is a sign of equal levels of hormones without an appearance of problems.

With this model, healthcare personnel will be able to identify thyroid conditions more quickly and make timely decisions to initiate suggested treatment. Furthermore, investing in learning about the diseases involved in the increase of noxious symptoms, their underlying mechanisms, other conditions that might arose due to the non-treatment of thyroid disorders, and other complications, could lead to an early diagnosis, and therefore, the prevention of late phases.

Moreover, early diagnosis in detecting thyroid diseases would allow inquis Nazi Space to recognize any threats to the health of the patient and prevent the impact on other organs of

the body which are considered irreversible and thereby reduce the exacerbation of the health risk. If thyroid problems are left untreated, one will face some life-threatening complications like heart disease, osteoporosis, infertility, and even thyroid cancer in extreme cases. This model saves a lot of money on medical bills, as well as long-term health consequences, because it helps people be quickly treated.

2.2 Related Works

Chaubey et al. compared three well-known machine learning models: logistic regression, decision trees, and k-nearest neighbors (kNN) to forecast the presence of thyroid disease, using the UC Irvine data. The analysis also stressed the promise of such algorithms in the classification of thyroid states and underlined the significance of early diagnosis. According to the study, KNN performed better than logistic regression and decision trees with a high accuracy in classification. This paper highlights the increasingly important role of machine learning in medical diagnostics and gives a fundamental comparison of algorithms to predict thyroid disease [1].

Sankar et al. studied the XGBoost algorithm to predict thyroid disease and compared the results with decision trees, logistic regression, and kNN. Their experiments showed that XGBoost was better than the other algorithm since it represented an increase in accuracy by 2 percent compared to kNN. The paper applied the UC Irvine dataset and showed the usefulness of feature selection with XGBoost, making it a powerful method of predicting thyroid disease. This work is consistent with the emerging direction of using sophisticated ensemble methods in healthcare predictions [2].

Salman and Sonuç evaluated several machine learning methods, such as SVM, random forest, decision trees among others, to classify thyroid disease. The researchers evaluated data obtained in hospitals located in Iraq and determined the decision tree algorithm to have the most accurate results compared to other algorithms. The article highlighted the attention that should be paid to the selection of classifiers and features in enhancing the performance of thyroid disease prediction systems [7].

In the work by Riajuliislam et al. the authors aimed to predict hypothyroidism at an early stage with the help of Recursive Feature Elimination (RFE), Univariate Feature Selection (UFS) and Principal Component Analysis (PCA). The experiment incorporated these methods with SVM, decision trees, and random forests, and Naive Bayes classifiers. They found that RFE produced the highest accuracy (99.35%) across all classifiers, showing the importance of selecting the features in enhancing the accuracy of classifying thyroid disease detection [5].

Razia et al. compared various machine learning models, such as SVM, multiple linear regressions, Naive Bayes, and decision trees on predicting thyroid diseases using the UCI dataset. The accuracy of decision trees was disclosed to be the highest at 99.23%. Nevertheless, the article admits that these algorithms were indeed promising but would require more development and take into account large volumes of data to be used in practice [8].

Shankar et al. presented a new method of classifying thyroid diseases with an optimal feature-based multi kernel support vector machine (MKSVM). They implemented feature selection through the Gray Wolf Optimization (GWO) approach in order to increase the accuracy of classification. The analysis noted considerable gains given that an accuracy of 97.49% was attained against other models. This study highlights the significance of feature selection with the contribution of kernel strategies in improving the accuracy of classification models in the diagnosis of thyroid disease [9].

Alyas et al. presented an empirical approach to classifying thyroid diseases through decision trees, random forests, KNN and artificial neural networks. The purpose of the study was to enhance the precision of thyroid diagnosis through the machine learning method in the IoMT setting. The results of the experiment were that the random forest classifier obtained a 94.8 percent accuracy. In this paper, the practical use of machine learning in the healthcare sector with the implication of the necessity of perfect, precise classification systems of thyroid diseases [11].

Jha et al. evaluated how dimension reduction approaches and data augmentation could positively impact the precision of the thyroid disease prediction. The research used a two phase methodology and showed a very high accuracy of 99.95 percent. Through the application of machine learning and deep learning solutions, the study showed that these state-of-the-art approaches could dramatically improve the prognostic abilities of the thyroid diseases, giving a way forward to more effective and sooner diagnosis in the medical field [10].

Rehman et al. examined the efficiency of different K-Nearest Neighbor (KNN) algorithms with a set of caliber functions in the classification of thyroid disorders. The paper considered KNN without feature selection and feature selection based on L1 and chi-square. They discovered that the use of chi-square-based feature selection performed best especially with the new column which had pulse rate, BMI, and blood pressure as new features. KNN is versatile and effective at classifying thyroid diseases in combination with suitable features selection methods [6].

Hosseinzadeh et al. offered a multiple multilayer perceptron (MMLP) with an adaptive learning algorithm to enhance the diagnosis of a thyroid disease in the Internet of medial things (IoMT). There we improved on the difficulties of slow convergence and local optima in older techniques of back-propagation. The MMLP model performed better in the accuracy of the classification of 99% as compared to single-layered models. In this study, the authors have shown the possibility of deep learning models to significantly improve diagnostics systems in IoMT health applications [3].

A study was conducted to predict thyroid disease using different machine learning algorithms, with the dataset collected from the UCI repository. The algorithms applied included Decision Tree, Random Forest, LightGBM, XGBoost, GaussianNB, KNN, ANN, SVC, CatBoost, and Extra-Trees, achieving an overall accuracy of 95.8%. However, the study was limited to a particular dataset without exploring broader generalization [12].

Another work aimed to categorize thyroid disease into different types such as hyperthyroidism, hypothyroidism, and normal, using data collected from around 1,250 Iraqi citizens across different genders and age groups. The algorithms used in this study were SVM, Logistic Regression, MLP, Decision Tree, Naïve Bayes, Random Forest, and KNN, achieving a high accuracy of 98.98%. Despite the strong results, the study relied on a relatively small dataset [13].

Similarly, researchers attempted to develop a classification-based system to detect euthyroidism, hyperthyroidism, and hypothyroidism using medical records collected from two hospitals in Haryana between January 2020 and July 2020. The models applied included Logistic Regression, Decision Tree, Naïve Bayes, Random Forest, and a proposed algorithm, achieving an accuracy of 94%, with precision of 96, recall of 97, and F1-score of 96. Although the study reported promising outcomes, its limitation lies in the use of a small dataset [14].

Existing studies on thyroid disease prediction have explored a range of machine learning models, from traditional algorithms like Logistic Regression, Decision Trees, and SVM to advanced approaches such as XGBoost, Random Forest, and Deep Neural Networks, often achieving accuracy levels above 90%. Most of these works, however, emphasize accuracy while paying limited attention to interpretability, dataset imbalance, scalability, and real-time clinical use. Some studies also rely on small or artificially augmented datasets, raising concerns about generalizability. In my opinion, while these methods demonstrate strong predictive capabilities, they fall short in providing transparency and trust, which are essential in healthcare. This highlights the need for Explainable AI-based approaches that balance predictive performance with interpretability, ensuring that thyroid disease prediction models are not only accurate but also reliable, transparent, and applicable in real-world medical practice.

2.3 Comparative Analysis and Summary

The Table 2.1 reviews several studies on thyroid disease prediction, outlining datasets, algorithms, results, and limitations. Most research used UCI, KEEL, or hospital data with varying sample sizes. Common models included Logistic Regression, Decision Trees, kNN, SVM, Random Forest, Naïve Bayes, MLP, and deep learning, achieving accuracies between 81% and nearly 100%. While advanced methods with feature selection or neural networks showed high accuracy, most studies faced challenges such as small datasets, limited features, reliance on artificial data, and lack of scalability for real-time applications.

Table 2.1: Comparative summary of related work

Author and Year	Dataset	Algorithms	Results	Limitations
Chaubey et al. (2021)	UCI dataset (215 instances)	Logistic Regression, Decision Trees, kNN	Logistic Regression: 81.25%, Decision Tree: 87.5%, kNN: Not provided	Limited dataset size, only 2 features used
Sankar et al.	UCI dataset	XGBoost, Decision Trees, Logistic Regression, kNN	XGBoost outperforms other methods by 2%	Focused on accuracy, limited analysis on other metrics
Sonuç et al (2021)	UCI dataset, others from hospitals	SVM, Decision Trees, Random Forest, Naïve Bayes, kNN, MLP, LDA	Various classifiers showed accuracy between 91-99%	Doesn't address real-time implementation or model scalability
Riajuliislam et al. (2021)	Diagnostic data from Dhaka, Bangladesh	Recursive Feature Selection (RFE), SVM, DT, Random Forest, LR, Naive Bayes	RFE + SVM/DT/RF/LR/Naive Bayes all achieved 99.35% accuracy	Small dataset (519 instances)

Razia et al. (2018)	KEEL dataset, UCI dataset (3152 records)	SVM, Decision Trees, KNN, Random Forest, Naïve Bayes	Random Forest: 94.8% accuracy	Dataset imbalance, feature set not optimized
Jha et al. (2022)	UCI dataset	Dimension reduction, data augmentation, Deep Neural Networks, Decision Trees	Achieved 99.95% accuracy using a two-stage approach	Focused on artificial data augmentation, may not generalize well
Hosseinzadeh et al. (2021)	UCI thyroid disease dataset (7200 instances)	Multiple MLP (neural network), adaptive learning rate algorithm	99% accuracy using multiple MLP networks	Computational cost, risk of overfitting with too many networks
Rehman et al. (2021)	KEEL dataset and registered hospital data	KNN, Chi-square, L1 feature selection techniques	KNN with chi-square-based feature selection achieved highest accuracy	Results based on small dataset with added features for specificity
Kumar et al. (2023)	Hospital data from Haryana, India	Logistic Regression, Decision Tree, Naïve Bayes, Random Forest, Proposed Algorithm	94% accuracy, Precision 96, Recall 97, F1-score 96	Small dataset
Al-Azzawi et al. (2023)	Iraqi citizens dataset (1250 records)	SVM, Logistic Regression, MLP, Decision Tree, Naïve Bayes, Random Forest, kNN	98.98% accuracy	Small dataset

2.4 Research Gap

Most studies that explain the process of thyroid detection with the help of machine learning have certain limitations to them One of the limitations relates to the size of the dataset that is relatively low in most cases with fewer than 600 instances. Moreover, a major proportion of the studies emphasize only on accuracy keeping the number of features extremely few in some cases even to two features. In this way constraining the perceived richness of the data, and losing much of the diagnostic information. Most of these works have been highly concerned with obtaining high accuracy, but hardly any attention has been placed upon measuring performance in terms of other important measures, such as precision, recall, F1-score, or ROC-AUC, particularity within an imbalanced framework. They also rely on some artificially generated data, such as augmentation. In none of the studies was the cause of detecting the reason stated. It is those Limitations that I will overcome, as part of my research.

2.5 Challenges

In order to carry out the research, several difficulties must be overcome. To start with, an adequate dataset to feed the model. The majority of the datasets present several issues, including missing, incomplete and inaccurate data, imbalanced data, and low-value attributes. Also, in most cases the size of the database is small. To curb these errors, a variety of techniques must be applied on the preprocessing of data and these include missing data, outlier techniques, data cleansing, data standardizing and so many more. Selecting features is another challenge that has beset us It is one of the most significant sections of a research study because the model operates based on the features that are chosen. Moreover, the incorporation of techniques related to Explainable AI, e.g. SHAP or LIME, complicates the matter, as they need to not only explain accurately, but maintain the ability to be interpreted by medical professionals. Having sufficient computation resources, time management, keeping a constant communication with the supervisor, sticking to ethical norms such as privacy of data and referencing accuracy are also problems that occur over time in the research.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The study topic is based on the idea of creating and testing the effectiveness of the usage of Explainable Artificial Intelligence (XAI)-based machine learning model to predict thyroid diseases precisely and transparently. The proposed study will rely on an externally published thyroid diseases dataset, which will consist of the wide-ranging set of clinical and demographic features, such as the age of the patient, gender, history, laboratory testing results, and diagnostic indicators. These variables are used as major characteristics of training and evaluation of predictive models.

For instrumentations, this study uses some of the software tools and programming libraries as well as the use of computational facilities. The primary programming language is Python and libraries of machine learning, such as Scikit-learn, XGBoost and CatBoost are applied to train and refine models. Explainable AI approaches are incorporated, specifically, SHAP (Shapley Additive exPlanations), to provide an explanation of why models make a particular decision and improve model transparency. The processing and representation of data are also finished with the help of Pandas, NumPy, and Matplotlib. The experiments are conducted on Google Colab, and make use of cloud GPU acceleration to provide maximum performance of models.

3.2 Dataset

In this research, the three-class data was chosen to identify the thyroid condition in clear and hierarchical manner- hyperthyroidism, hypothyroidism and normal (no disease). This strategy will give the model an opportunity not only to detect a patient who has a thyroid disorder but also to conclude that a patient is healthy.

To access more data, the source of data in this case was the UCI Machine Learning Repository, which is credible in supplying research data. It has 31 attributes and 9,173 patient records composed of demographic data, thyroid-related lab data (thyroxine thyroid stimulating hormone T3, T4, FTI and TBG), and clinical symptoms (tumor presence, goitre, overall health status). Target attribute is the thyroid condition of patient which is the predictive outcome.

With a combination of the clinical characteristics and laboratory data, the dataset is rich enough that can be used to develop an accurate and transparent model detecting thyroid diseases.

Table 3.1: Dataset description

No	Feature Name	Description	Missing Values	Unique Values
0	age	Age of the patient	0	100
1	sex	Patient gender	307	2
2	on_thyroxine	whether patient is on thyroxine	0	2
3	query_on_thyroxine	whether patient is on thyroxine	0	2
4	on_antithyroid_meds	whether patient is on antithyroid meds	0	2
5	sick	whether patient is sick	0	2
6	pregnant	whether patient is pregnant	0	2
7	thyroid_surgery	whether patient has undergone thyroid surgery	0	2
8	I131_treatment	whether patient is undergoing I131 treatment	0	2
9	query_hypothyroid	whether patient believes they have hypothyroid	0	2

10	query_hyperthyroid	whether patient believes they have hyperthyroid	0	2
11	lithium	whether patient is lithium	0	2
12	goitre	whether patient has goitre	0	2
13	tumor	whether patient has tumor	0	2
14	psych	whether patient is psych	0	2
15	TSH	TSH level in blood from lab work	842	369
16	T3	T3 level in blood from lab work	2604	85
17	TT4	TT4 level in blood from lab work	442	287
18	T4U	T4U level in blood from lab work	809	176
19	FTI	FTI level in blood from lab work	802	323
20	TBG	TBG level in blood from lab work	8823	66
21	target	hyperthyroidism medical diagnosis	0	32

3.2.1 ThyroidDF

This research utilized the dataset known as ThyroidDF retrieved through the UCI Machine Learning Repository that is a long-established source of public research data. It is in the form of 9,172 patient records and 31 attributes that include demographic data (age, sex), medical history (i.e. use of thyroid medication, pregnancy status, illness), and clinical test data (i.e. TSH, T3, TT4, T4U and FTI).

The main purpose of utilization of this data is identification and classification of a patient belonging to one of three groups: hyperthyroidism, hypothyroidism, or no thyroid pathology. The data were initially obtained in form of medical diagnosis reports and

hospital records which has provided a wide scope of evaluation of cases in various conditions.

The integration of these two characteristics, demographic and laboratory test, in ThyroidDF gives a richer area to develop and validate machine learning-based predictive models of detecting thyroid problems.

Table 3.2: Data encoding map for target class

Class	**Encoded Class**	**Instances**
Negative	0	182
Hypothyroid	1	593
Hyperthyroid	2	6,767

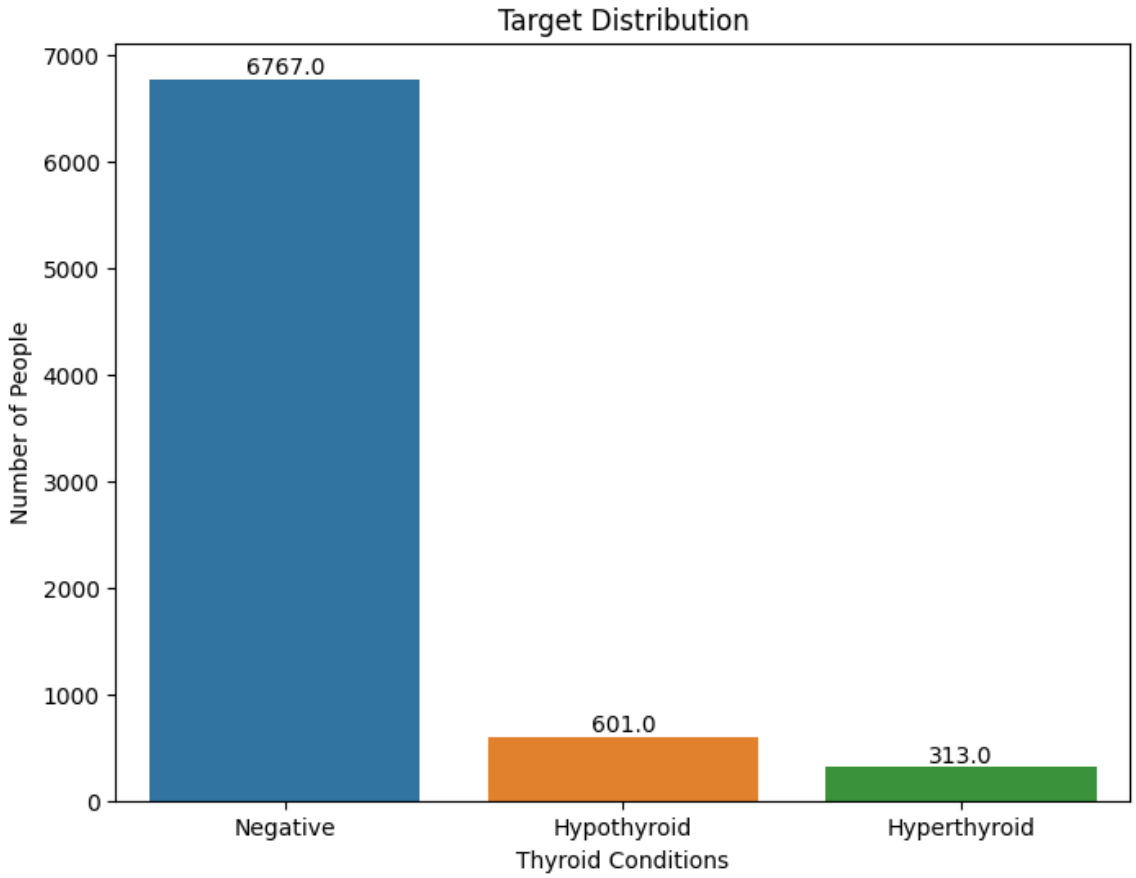


Fig 3.3: Count of target class instence

3.3 Data Pre-Processing

This was done prior to training the predictive models to make sure that the quality and usefulness of the dataset is assured. This included imputation of missing data, encoding categorical features, scaling real-valued features and overcoming class imbalances (if any). Data was also cleaned to eliminate inconsistencies in the data or any extreme outliers that may cause the faulty performance of the models. These procedures preconditioned the data to efficient and sound machine learning analysis.

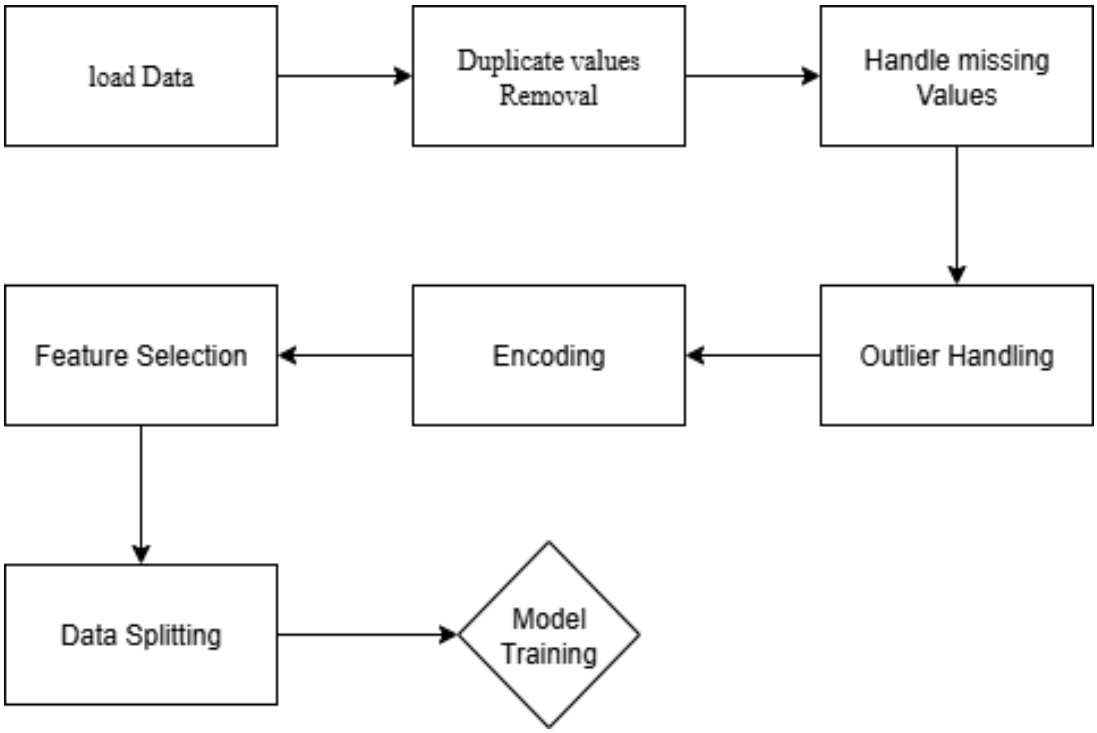


Fig 3.4: Data pre-processing steps

3.3.1 Missing Value Handling

Missing data is a widespread problem in a medical dataset, in which incomplete or missing data can affect the relative performance and accuracy of predictive models. The missing data can be received because of the unlikeable factors, including data collection mistakes, radiance hardware misfunction, or inconsistent donor records. Missing data should be addressed with caution since neglecting this problem can result in biasing the models or making wrong forecasts, which is highly undesirable in such a delicate domain as healthcare.

Table 3.3: List of Missing Value Count

Attribute	No of missing Values	Used Method
sex	307	Mode Imputation
hypopituitary	0	-
TSH_measured	0	-
TSH	842	Replaced By 0
T3_measured	0	-
T3	2604	Replaced By 0
TT4_measured	0	-
TT4	442	Replaced By 0
T4U_measured	0	-
T4U	809	Replaced By 0
FTI_measured	0	-
FTI	802	Replaced By 0
TBG_measured	0	-
TBG	8823	Replaced By 0
target	0	-

We used imputation in dealing with the missing values in our dataset. Specifically:
Mode Imputation: Missing values of categorical features (e.g. Gender) were substituted with the most frequently-occurring value within that feature.
Replacing missing: Missing values were replaced by 0 and only the numeric variables were considered here like TSH, T3, TT4, T4U, FTI, TBG.

3.3.2 Attribute Dropping

The inclusion of nulled out values in the TSH, T3, TT4, T4U, FTI, and TBG columns can be directly related to the presence of f in the related TSH_measured, T3_measured, TT4_measured,T4U_measured, FTI_measured, and TBBG_measured columns. This implies that since there were no TSH, T3, TT4, T4U, FTI, and TBG test values measured in the blood test, there were no test indications to the same. To this reason such columns are removed. There is also only one type of value in a column corresponding to a hypopituitary condition; it is doubtful whether it will ever add value to pattern recognition or improve performance of a data model. It is also dropped so.

3.3.3 Outlier Handling

The outliers, which are values highly aberrant to the rest of the data, were detected inaccurate values that were not to affect the model learning.In this study, there was outliers handling, where the extreme and unrealistic values could not skew the model learning. In particular, an analysis of the age characteristics was performed, and those values greater than 100 years were treated as outliers, which is unlikely to occur in the context of the data collection. Patient ages that produced such records were dropped in the dataset, giving a more realistic distribution of the data.

3.3.4 Data Encoding

Machine learning algorithms usually accept only numerical values, and categorised attributes require conversions to the corresponding suitable numerical representation to be compatible and to enhance the model behaviour. Appropriate encoding would prevent misinterpretation of categorical values as being ordinal or continuous thereby resulting to biased learning results.

To make the data set ready to be used by machine learning algorithms, the categorical variables were encoded into sequential numeric format. This was necessary since most machine learning models are unable to directly operate on categorical/textual data Some of the categorical variables in the data included binary variables (f/t) and gender representation (M/F) and multi-category categorical variables like referral source.

In case of binary categorical attributes embodied by the letters f (false) and t (true), a manual mapping technique was used, with f being assigned zero and t being assigned one. This mapping provided uniformity across all of the binary features including on_thyroxine, query_on_thyroxine, on_antithyroid_meds, and sick, and so on. The gender characteristic (sex) was coded independently with simple correspondence, F and M were assigned, respectively, the value 0 and 1, as the variable is binary.

In the attribute referral_source, which included more than one nominal level, Label Encoding was used. This process systematically gave each category of values a unique integer thus making sure that no information was lost and at the same time introducing no bias. Label Encoding was a good option because the variable was discrete, categorical and unordered in nature.

Table 3.3: Data encoding map

Variable Type	**Attributes**	**Encoding Method**	**Mapping / Transformation**

Binary (f/t)	on_thyroxine, query_on_thyroxine, sick, pregnant, etc.	Manual Mapping	f → 0, t → 1
Gender (M/F)	sex	Manual Mapping	F → 0, M → 1
Multi-class (Nominal)	referral_source	Label Encoding	Each unique category → integer (0, 1, 2, …)

3.3.5 Feature Selection

In medical machine learning the selection of features is necessary in order to enhance the predictive accuracy, avoid overfitting, and create the explanations. In clinical datasets, irrelevant, redundant, or noisy features potentially can negatively affect model behavior, and a systematic selection is, therefore, required.

In this study, hybrid approach was used. To eliminate the redundancy between variables that had high inter relational coefficients correlation analysis was initially used. The feature importance was then obtained, by using the XGBoost model based approach, summarising both linear and non-linear correlations. Lastly, to give transparency through feature contribution to predictions, SHAP values were used.

The selected methods were made to accommodate the goals of accuracy and transparency: correlation analysis would increase efficiency, XG Boost would ensure sound ranking, and SHAP would increase the interpretability.

The last feature set encompassed clinical relevant data points such as the thyroid-stimulating hormone (TSH), triiodothyronine (T3), and total thyroxine (TT4) in conjunction with demographic values such as age and gender. These characteristics provided statistical value and medical considerations that the predictive model was precise, and interpretable, and had clinical relevance.

3.3.6 Train Test Split

The data was divided into an 80-20 split where by 80 percent of the records were used to train the models and 20 percent were used as a test. A static seed (random_state = 42) was used so as to ensure that the results are reproducible. This process gave four arrays; x_train which was the training features, y_train which was the labels of the training sort, x_test which was the testing features, and finally, y_test which were the testing labels.

This plan enabled the models to get ideas of the training data, and individually measure it on the test data minimizing the danger of overfitting and guaranteeing the accurate assessment of generalization performance.

3.4 Exparimental Process

The methodology of the current study is meant to take a procedure approach, i.e. build on the intensification of the dataset and end with the interpretation and explanation of outcomes. Thyroid data were initially brought into the working environment and then a thorough data preparation process took place. In the given step, any entries of duplicates were excluded to avoid duplicate records, missing values were addressed by appropriate methods of imputation, and outliers were processed to reduce the influence of abnormal records on the analysis performed. To train the data in machine learning, the categorical data were transformed into numerical data, and feature selection techniques have played a role in eliminating all insignificant features to keep only significant features currently pertinent to thyroid disease classification.

The data was used to create predictive models after preprocessing and then the data was split into the training and testing subsets. Several modeling approaches to this task were attempted, where Image Logistic Regression, Weighted Image Logistic Regression, Decision Tree based model, Random Forest, XGBoost, CatBoost and Gradient Boosting models are used. The models were all trained using the training subset and then tested using

the testing subset to give an insight on the effectiveness of each one in diagnosing thyroid disorders.

After every instance of the models used, their results were then compared in order to determine the most effective method. The estimated value to measure the perimeter considering the measures of accuracy, precision, recall and F1-score was weighted, which meant the resulted in an unbiased picture of classifiers performance. The final model of the study was determined as a most performers in general among the models. To further increase the level of interpretability and a dependable model the selected arsenic used approaches of Explainable Artificial Intelligence (XAI). XAI enabled gaining a better understanding of the way the model carried out the decision-making process by informing on the importance of specific features, therefore providing a visibility and confidence level for the diagnostic results.

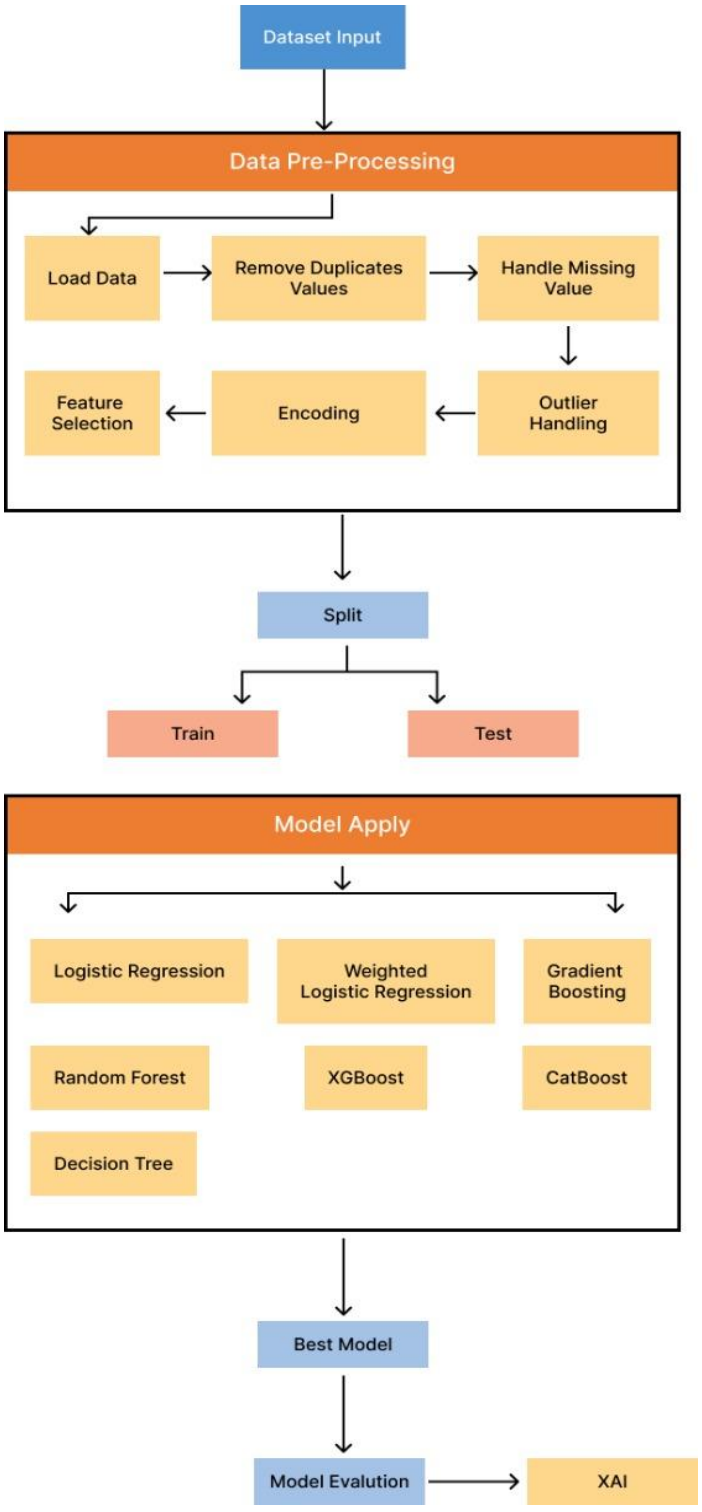


Fig 3.5: Proposed Methodology

3.5 Machine Learning Algorithms

This paper uses a variety of machine learning tasks like Random Forest, Support Vector Machine (SVM), and Logistic Regression, and they were first assessed in terms of their accuracy with respect to the thyroid data. In terms of how well it handled the data and the brevity of distortion, XGBoost proved to be the most promising model, even suggesting good performance specificities not seen in the other models. To ensure transparency and intelligibility of medical decision-making, the SHAP (SHapley Additive exPlanations) and LIME was added into the XGBoost model. This ensemble model was not only able to accurately predict but also gave straight forward information about feature importance hence reliable and explainable.

3.5.1 Logistic Regression

Logistic Regression is one of the most basic and the most popular statistical methods in classification problems. It is trained based on the likelihood (probability) estimation of target event based on input feature values given by logistic (sigmoid) sigmoidal function as follows Unlike linear regression in which the values used to predict the outcome are continuous whereas in the case of logistic regression, the outcome is the probability of the objects belonging to a class. It is suitable for the binary classification and extensions can be developed to apply the identified criteria in multiclass case. Logistic regression has important advantages owing to interpretability through coefficients which can be derived which describe the impact that each predictor variable has on the outcome variable of interest. But it performs badly when the data forms some nonlinear functions, or the variables under consideration interact with one another. Use of logistic regression while not majorly affected by these shortcomings has nonetheless made logistic regression very popular as the starting point in predictive analytics, and find repeated use in medical uses, social science and financial applications because of its simplicity and stability.

3.5.2 Decision Tree

A Decision Tree is a supervised learning method; it is founded on a model of choices that come in the form of a tree to classify information. Its functionality is based on the recursive division of the dataset into subsets according to the values of the features and, eventually, on the leaf nodes corresponding to the estimated class. The model is interpretable since it resembles the decisions made by humans through the rules of the IF-then. Decision trees give a more flexible nonlinear relationship with data, unlike logistic regression. Their sensitivity to minor changes in the data set is high and they often over fit when trees become too deep. This effect is dealt with by pruning, maximum depth limit and minimum samples per split. Simple decision trees lack robustness, but these can form the basis of more sophisticated ensemble techniques such as Random Forest and Gradient Boosting, which are much more powerful and generally-purpose in terms of their predictive performance.

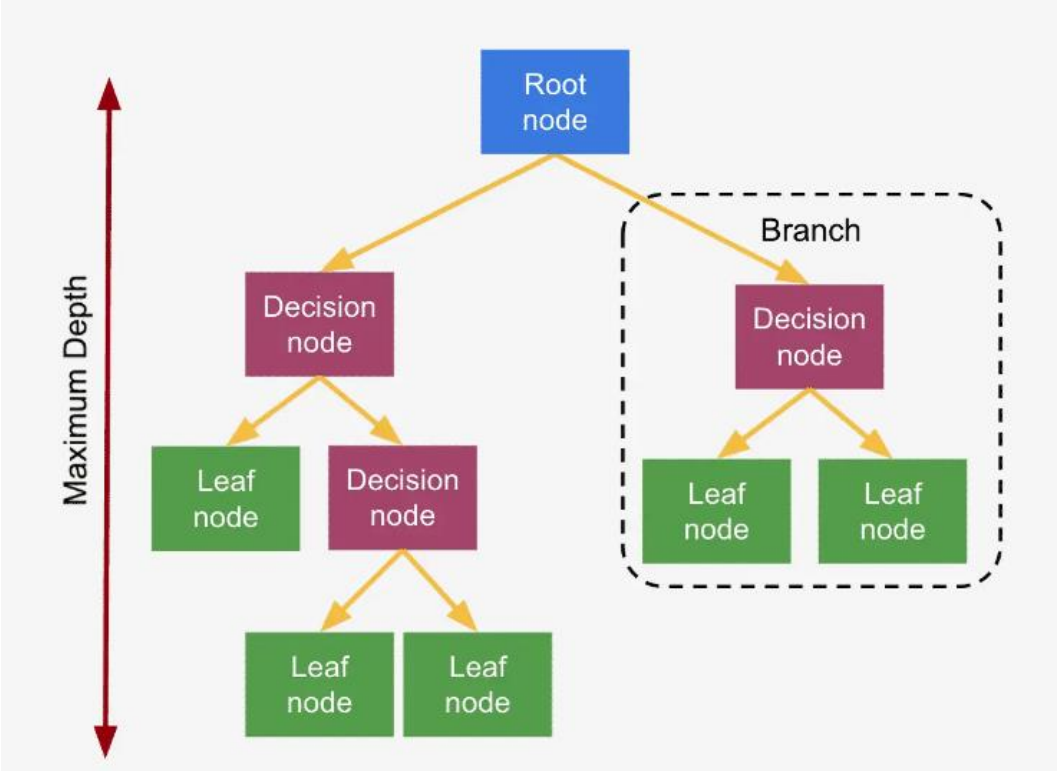


Fig 3.6: Decision tree architecture [27].

3.5.3 Random Forest

Random forest is yet another form of ensemble technique where a group of decision trees is produced and prediction of trees will be combined into a final solution. It also utilizes what is known as the bootstrap aggregation or bagging whereby several trees are constructed from random subsets of the data and attributes. This is balanced by the diversity of the trees which prevent the problem of overfitting that is typical of single decision trees. Random Forest is robust and can work effectively in large datasets, missing values and the high-dimensional feature space. It is also a strong point that they deliver measures of feature importance, which can help a researcher to know which variables provide the most contribution to the prediction. Although it provides a near perfect predictive performance, the interpretability of the model is not as good as that of the simple decision trees or logistic regression. Still, Random Forest is a highly popular choice in classification tasks after striking a balance between accuracy and generalization.

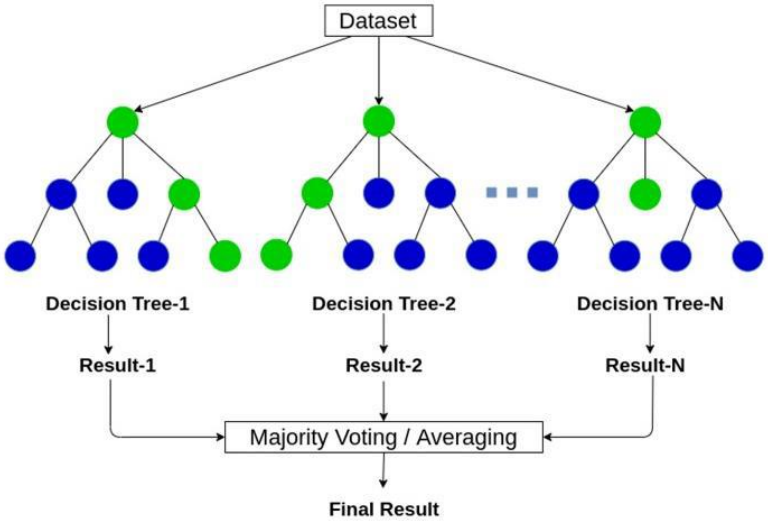


Fig 3.7: Random Forest Model Architecture [28].

3.5.4 Weighted Logistic Regression

Weighted Logistic Regression is an alteration to the basic logistic model that supplies classes or instance weights in order to rectify the lopsided nature of the data. In practice, in a number of contexts, one of the classes may be substantially more frequent than the other, and it may be of interest to model that skew. Such imbalance may lead to the biasing of a model towards the majority class at the expense of the minority, which is more often significant. The misclassification of minority classes incurs higher penalties in a weighted logistic regression model, which means that the model will not neglect them. This substitution increases the sensitivity and recall of the classifier with limited reduction in the accuracy of the classifier. Weighted logistic regression has been used in healthcare analytics, fraud detection and risk assessment research because of its interpretability and imbalance handling behavior.

3.5.5 Gradient Boosting

Gradient Boosting is sequential model prediction approach, an ensemble learning method of machine learning. However, in comparison to Random Forest, the trees used in Gradient Boosting do not train separately but are sequentially introduced, each aiming at minimizing the errors of the last compound. It achieves optimum predictions that would minimize a loss function through gradient descent This renders Gradient Boosting extremely flexible and potent as it is capable of learning non-linear associations in the data. Nonetheless, the algorithm is computationally expensive, and can become prone to overfitting when the number of trees or the depth is not well-regulated. These challenges notwithstanding, the large number of areas which Gradient Boosting has been applied is due to its flexibility and high classification and regression accuracy.

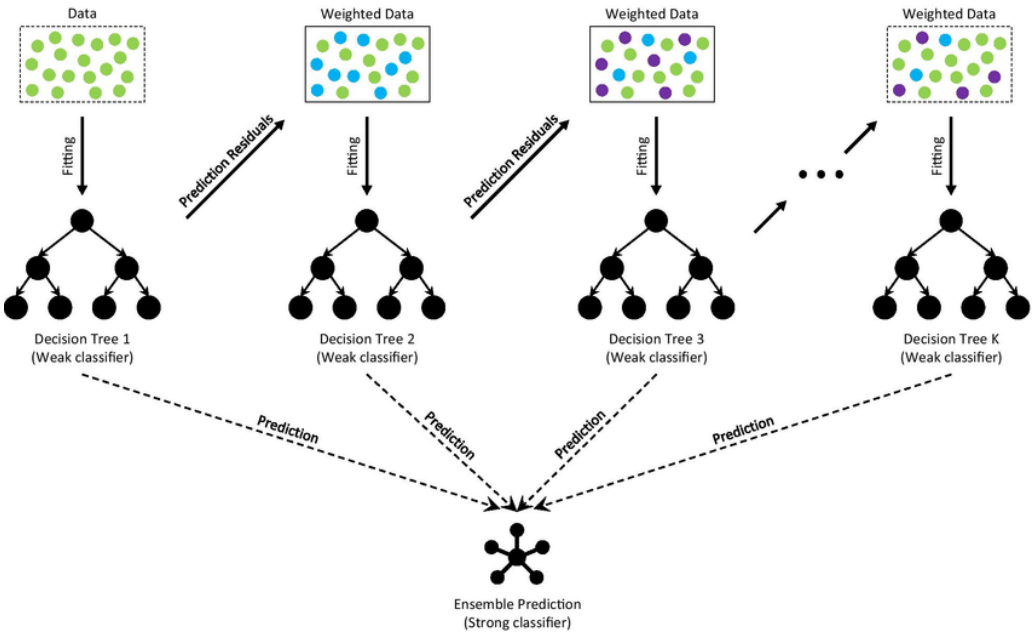


Fig 3.8: Gradient Boosting Model Architecture [29].

3.5.6 XGBoost

XGBoost is the short form of Xtreme Gradient Boosting, which is an improved version of gradient boosting that leads to one of the most potent algorithms on structured data. It builds an ensemble tree which relies on decision trees in a sequential order such that each new tree 'corrects' the mistakes made by a given tree. XGBoost uses state of the art regularization techniques to prevent overfitting and has a number of optimizations - such as parallelism, memory efficiency, and the ability to handle sparse datasets that allow it to be efficient as it scales to large datasets. One of its most important features is that it has the ability to capture high predictive accuracy on a wide variety of machine learning competitions and research. Although more complex than other models available, XGBoost has been popular due to their tradeoff of these three parameters: performance, scalability, and flexibility. However, it requires a tedious parameter-tuning problem that is computationally expensive. The reason why XGBoost has been successfully applied in the field of medicine and diagnostics and is often a state of the art solution is due to its ability to capture complex data relationships effectively.

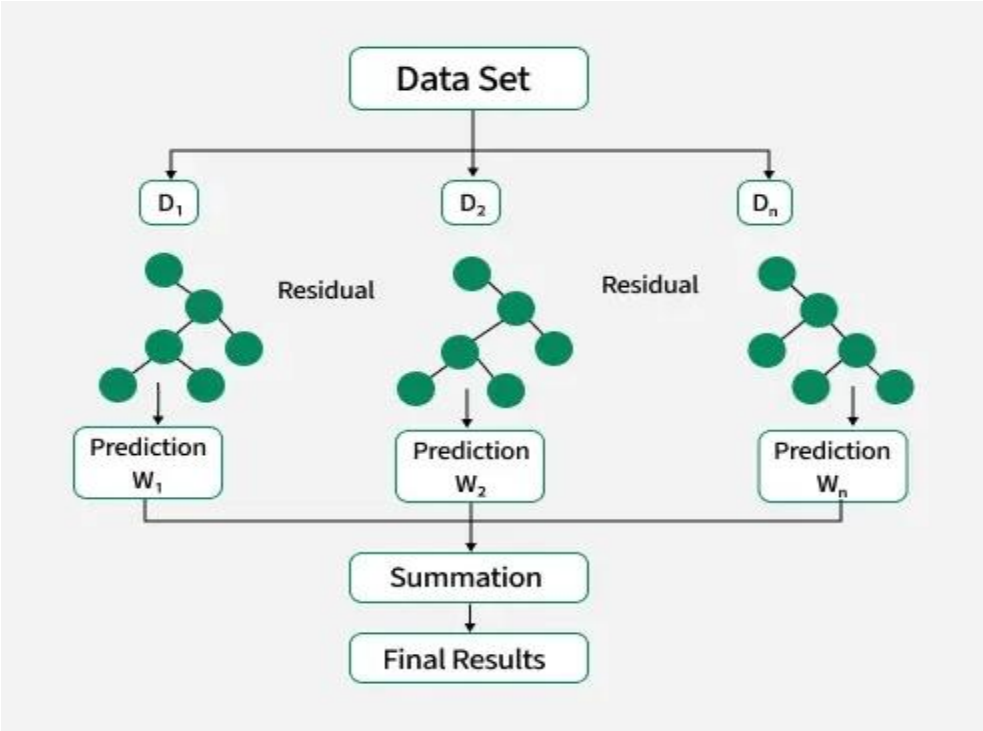


Fig 3.9: XG Boost Classifier Architecture [30]

3.5.7 CatBoost

CatBoost is a gradient boosting algorithm in which additional attention is paid to the efficient treatment of categorical variables that are often challenging with their non-linear presentation. Unlike with other boosting algorithms that need a lot of preprocessing, CatBoost can directly deal with categorical features via ordered boosting and target statistics. This saves preprocessing time and the chances of data exposure. CatBoost optimises symmetric trees which increase robustness of models and reduces the risk of overfitting. It has been known to provide solid performance in a large amount of datasets without too complex implementation. When compared to XGBoost and Gradient Boosting, CatBoost can be easier to use without hyperparameter tuning as it has a good level of accuracy nonetheless. It is especially appealing in real-life issues with mixed types of data, like healthcare, finance, and recommendations [26].

3.5.8 Explainable AI

Explainable artificial intelligence (XAI) is the structural attempt tomartyrize and explain the choices of machine-learning algorithms. High-stakes contexts, such as healthcare, reflect specific value on this transparency since it is important to know why AI-driven predictions tend to make decisions to trust and use them. XAI bridges the gap between complex systems and the non-experts by rendering it more intelligible and trustworthy, using multiple methods [12].

3.5.9 LIME

Local Interpretable Model-agnostic Explanations or Lime is one of the popular XAI methods that are used to explain the predictions of individual models based on complex machine learning models. LIME seeks to give explainability to an adverse prediction by approximating the behavior of a black-box model with simpler, interpretable models in the local vicinity of a prediction. This plays an essential role when it comes to environments where transparency is required, and in the case of healthcare diagnostics this is essential with regard to providing clear understanding about how predictions have been constructed [13] Russians aimed at maintaining various levels of power, socio-economic stability and preserving the existing order [14].

3.5.10 SHAP

The other XAI method that uses a cooperative game theory is SHAP (SHapley Additive exPlanations) where each feature receives an importance score on how it contributed to the model prediction. HAP offers local and global explainability where stakeholders are able to visualize the individual impact of features on the individual estimates, as well as on the overall model. The technique proves exceptionally useful to interpret the predictions of complicated models such as tree-based models, as it ensures that the predictions made by the model can be probed and validated realistically by the stakeholders [14].

3.5.11 Application of Explainable AI

Explainable AI is instrumental in aiding the interpretation of the outputs of the final selected model, the XG boost classifier. The LIME and SHAP techniques are employed to explain the features that lead to the predictions in terms of thyroid diseases, such as the level of hormones and the age of the patient. This openness guarantees that the healthcare professionals can sufficiently trust decisions made with the help of AI when determining specific conditions, such as hyperthyroidism or hypothyroidism. XAI simplifies the practical use of AI in clinical scenarios by presenting clear reasoning regarding any decision [12-14].

Explainable Artificial Intelligence (XAI) is important in this research to fill the gap between how well a model performs and how explainable the model is. Although high accuracy in the prediction of thyroid disease can be obtained via machine learning models, especially ensemble and deep learning models, their black-box nature makes them difficult to use in clinical practice. Good medical practitioners require not only proper predictions; they require an explicit explanation of the process of making predictions.
Several XAI techniques (specifically, SHAP and LIME methods) are also integrated in this study and make the proposed model mature, credible and robust. These approaches provide the information of feature importance, the explanation of each individual patient and the impact of clinical variables on the model as a whole.. As an example, one might emphasize attributes like age or TSH level to show how the given characteristics affected the results of the classification to be better understood by medical workers.

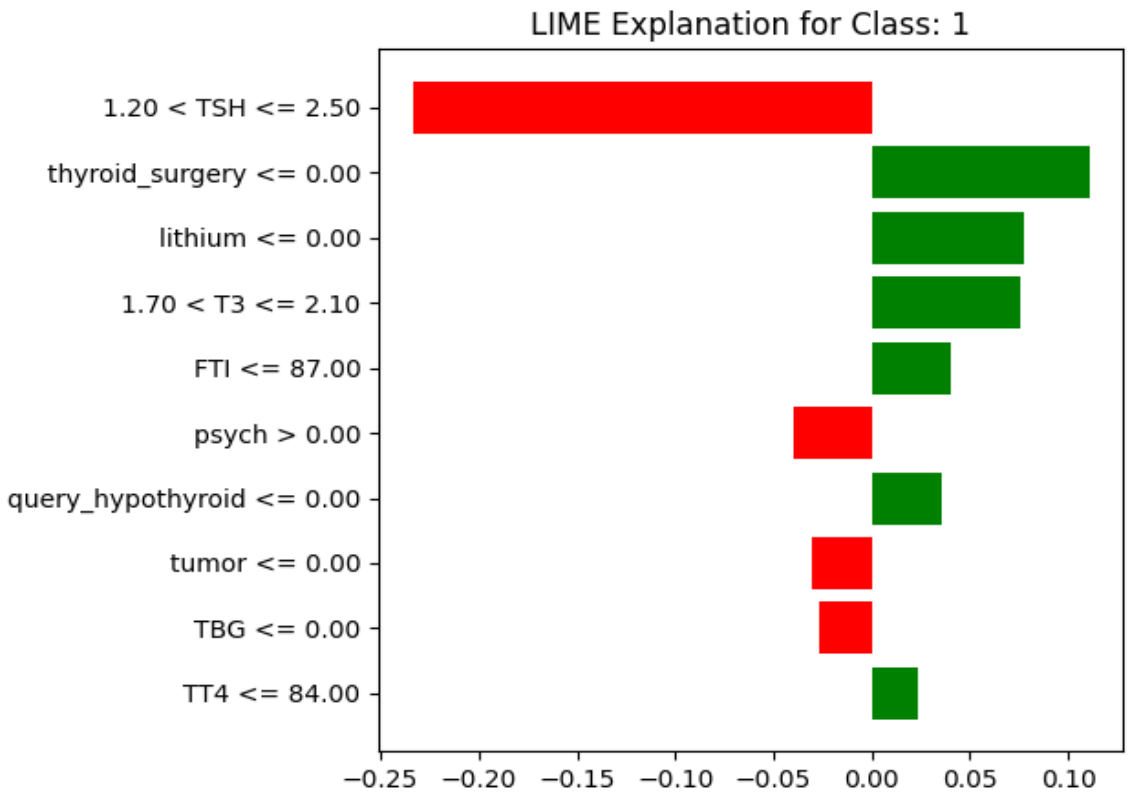


Fig 3.10: XAI Explanation for a class dicision

3.6 Implementation Requirements

It took various implementation requirements to ensure that this research was carried out successfully. All these requirements can be divided broadly into hardware, software and data related requirements:

Hardware Requirements

- Processing system that has a minimum of i5/i7 processor (or its equivalent).
- Minimum 8 GB RAM
- Transfer rates of at least 1 GB per minute to transfer data and output of experiments.
- Support of GPUs acceleration during model training and explainability methods.

Software Requirements

- Programming Language: Python
- Libraries: Scikit-learn, XGBoost, CatBoost, TensorFlow to perform modeling & SHAP and LIME to provide explainability.
- Tools: Google Colab as an experimentation platform, Pandas and NumPy to work with data, Matplotlib and Seaborn to visualize it.

Data Requirements

- An access to the thyroid disease dataset (with the clinical features and diagnosis outcomes).
- Effective pre processing such as cleaning, encoding of categorical variables as well as balancing of the classes proportion
- Splitting data used in training, test, and validation to guarantee in sound performance evaluation.
- All these implementation requirements serve as the basis of developing, testing, and interpreting the proposed thyroid disease prediction model and guarantee that the research will be efficient, reproducible, and scalable.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

This research framework was designed to make comparisons between the performance and interpretability of the recommended headache predictive model. All the experiments were conducted on Python as a primary programming language with the deployment in Google Colab platform. The physical architecture was comprised of ryzen5 3500x processor, memory of 16GB, and graphics card to carry out the model training and testing.

The preparation of data had been undertaken in the following ways to ensure that results are more valid; data cleaning, categorical encoding and the topic of class imbalance. The resulting data was divided into test and training segments with an 80:20 Thus a combination of the %ages maintained in the class distributions. Several machine learning libraries were tested and compared on their metrics of accuracy, precision, recall, F1-score and AUC: XGBoost and CatBoost were utilized.

Interpretability of the models was also taken into consideration by employing SHAP and LIME methods in the experiments. These tools were used because they allow exploring feature importance both at aggregate level and at individual level, to gain further insight on the way that the model makes a decision.

This approach is a nice compromise as it not only offers good performance tests but also has well-understood interpretation on the results of the model, which fits into the goal of cross-state comparison studies.

4.2 Metrics of Evaluation

The inner objective of machine learning is coming up with a model that precisely predetermines the results of new information. Evaluation measures are necessary in order to determine the extent to which the model meets this goal. Using these measurements, it will be possible to pinpoint some of the causes of errors and make certain changes to improve the accuracy and effectiveness of the model in general. In the present work, the effectiveness of the introduced intrusion detection approach was checked with a series of evaluation metrics.

Accuracy was adopted as the major performance indicator which reflects the percentage of items correctly categorized against the total number of items in the data set. Albeit providing a brief idea of how the performance is, precision is inaccurate in class imbalanced instances. Consequently, more specific measures including accuracy, recall, and F1-score were used in the assessment also.

- **Accuracy**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where True Positives: TP, True Negatives: TN, False Positives: FP and False Negatives: FN.

- **Precision**

$$Precision = \frac{TP}{TP + FP}$$

Precision Metric calculated in the ratios of how well the positive predictions are correct with respect to the total positive predictions made.

- **Recall (Sensitivity)**

$$Recall = \frac{TP}{TP + FN}$$

Recall is the capability of the model to successfully recognize the real positive cases.

- **F1-Score**

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall}$$

The F1-score provides a harmonic balance between precision and recall, especially useful when classes are imbalanced.

Besides these measures, a confusion matrix was obtained to give a visualization of the classification accuracy in the three classes: hyperthyroid, hypothyroid, and negative. This was assisted by the matrix in discovering the trends of misclassification.

Moreover, Receiver Operating Characteristic (ROC) curve and its AUC was used to test sensitivity-specificity trade-off at various decision thresholds. The greater the AUC the better the model at classifying data between classes.

Based on such evaluation metrics, a good performance comparison was established among all models such as, Logistic Regression, Decision Tree, Weighted Logistic Regression, Random Forest, Gradient Boosting, CatBoost, and XGBoost. This intense evaluation model selected a fair-play in modeling comparison and allowed the accuracy and reliability of the XGBoost model combating thyroid disorders with great precision and diligence.

4.3 Results & Analysis

In the following, the findings of using both the normal and weighted machine learning models over the thyroid data are provided. The main rationale to use a weighted model was its ability to counter the problem of unbalanced classes whereby, minority classes like Hyperthyroid were not dwarfed by the Negative class. All models were trained and tested on the processed dataset and performances compared based on accuracy and class-wise precision.

The baseline Logistic Regression model produced a mediocre performance and was not able to capture the feature relationship (nonlinear association) which caused it to perform weaker in predicting the minority classes. Using weighting, there was an improvement in performance particularly on Hyperthyroid classification, showing the advantage of punishing the misclassification of underrepresented classes. There was a slight benefit of the Decision Tree model because of class weighting, but it was subjected to some overfitting problems. Random Forest and Gradient boosting performed better and their weighted versions were more balanced in their performance over the classes. CatBoost retained a high value of predictive accuracy, and weighting enhanced its level of Hyperthyroid cases classification.

The largest increments were seen in XGBoost model. XGBoost performed better than all the other algorithms even in absence of weighting. When class weights were added, this resulted in a drop in overall accuracy of 0.5 percentage along with a better balance (though by small margins) amongst all the values of precision. This shows that, although XGBoost is immune to imbalance, weighting provides additional explanatory power in improving fairness in predictions.

Table 4.1: Results of all the traied model

Model	Accuracy	Hyperthyroid Precision	Hypothyroid Precision	Negative Precision
Logistic Regression	92.4%	0.781	0.956	0.982
Weighted Logistic Regression	93.8%	0.835	0.962	0.985
Decision Tree	94.6%	0.812	0.963	0.987
Random Forest	96.7%	0.889	0.981	0.993
Weighted Random Forest	97.0%	0.902	0.983	0.994
Gradient Boosting	97.1%	0.902	0.985	0.996
Weighted Gradient Boosting	97.5%	0.917	0.987	0.996
CatBoost	97.8%	0.918	0.989	0.997
Weighted CatBoost	98.0%	0.928	0.991	0.997
XGBoost (Final Model)	98.5%	0.944	0.995	0.999

Table 4.1 compares different machine learning models for thyroid disease prediction. Simpler models like Logistic Regression show good accuracy (92.4%), but advanced ensemble methods such as Random Forest, Gradient Boosting, and CatBoost achieve higher precision across all classes. The final XGBoost model outperforms others with 98.5% accuracy and the highest precision, making it the most reliable choice.

4.3.1 XGboost Classifier

XGBoost classifier was fitted to the thyroid disease data and its performance illustrate exceptional predictive performance. The confusion matrix has given below:

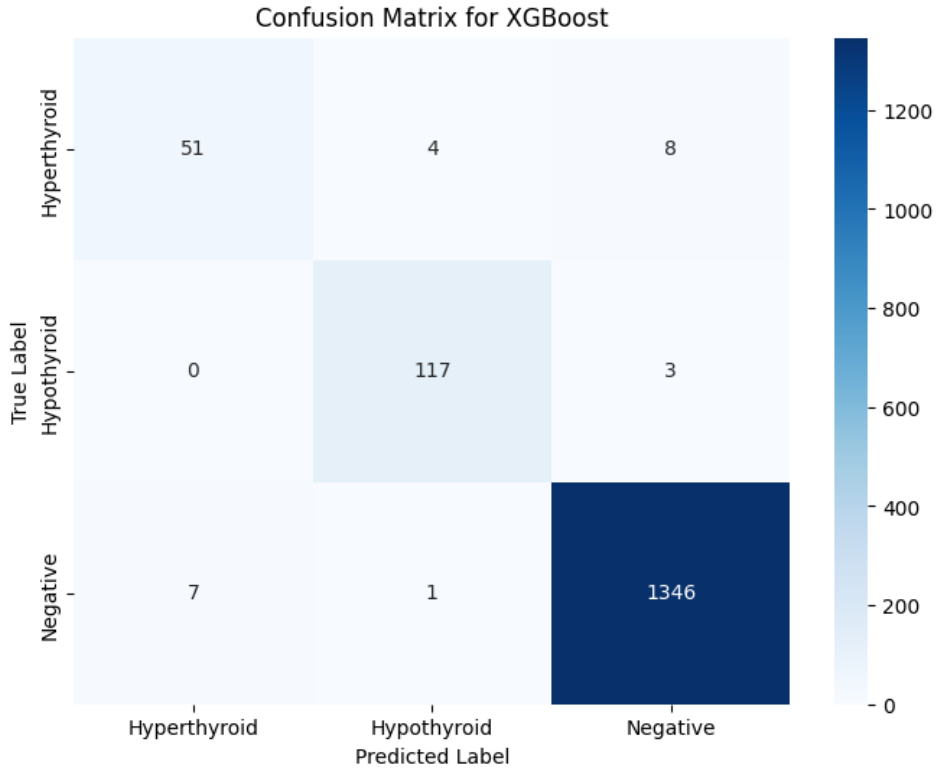


Fig 4.1: Confusion matrix of XGboost

Fig 4.1 indicates that the model accurately classified a very high ratio in each of the three categories, especially when it comes to case of Negative and Hypothyroid. The total precision was 98.50%, which indicates the quality of the model to recognize the thyroid disorders. Hyperthyroid and Hypothyroid averaged 0.944 and 0.996 respectively, and Negative had 0.999 average precision whereas micro-average precision was 0.999. These findings demonstrate that XGBoost outperforms all other models in classification accuracy and best fits the minor categories such as the case of Hyperthyroid which had high precisions as well as accuracy. In general, XGBoost turned out to be the most stable and consistent model in this paper that can provide highly accurate and interpretable results in medical diagnostics.

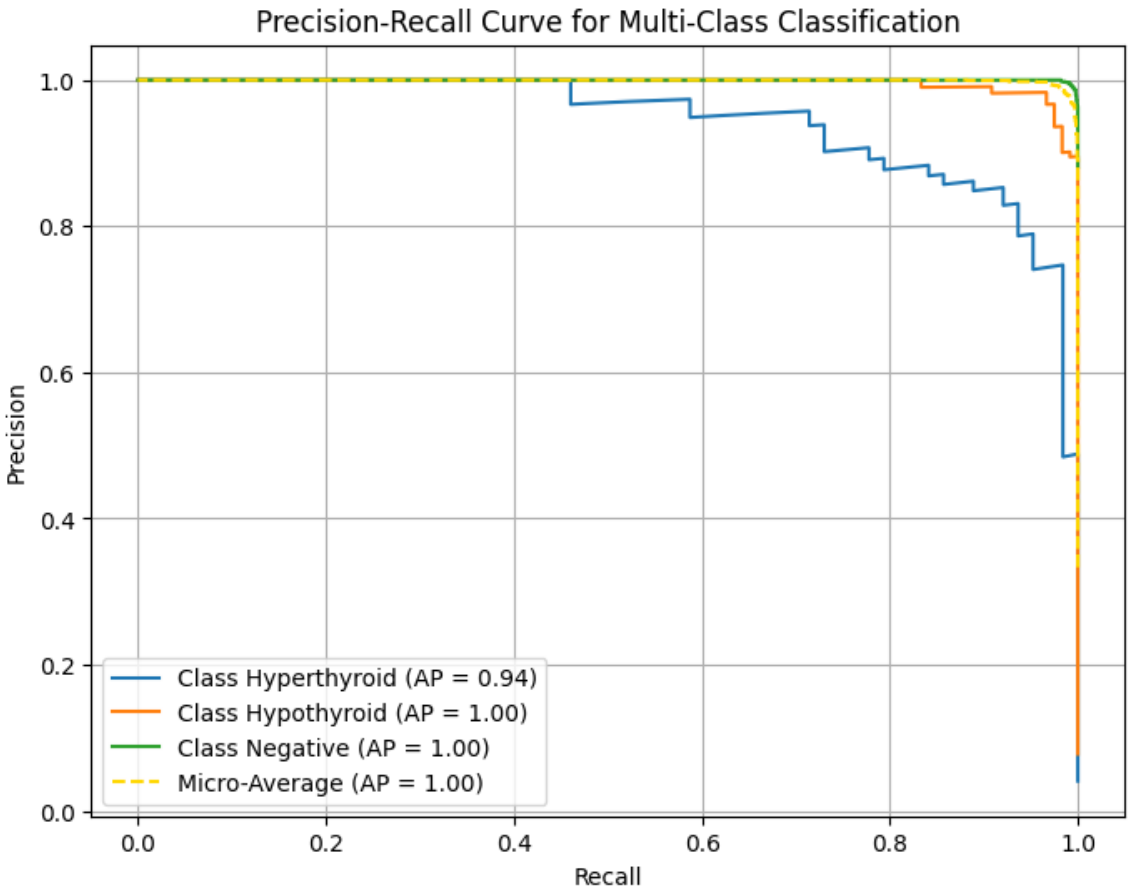


Fig 4.2: Precision vs Recall curve of XGboost Classifier

4.3.2 Decision Tree

Decision Tree classifier is used to classify patients with the thyroid diseases into three categories Hyperthyroid, Hypothyroid, and Negative. Several measures have been used to analyze the performance of the model. The accuracy of the Decision Tree model is 98.37% in general, and this fact shows that the implemented model can be successfully used to generalize thyroid diseases. The precision of the Hyperthyroid class is 74.72 percent, Hypothyroid one is 89.97 percent and the Negative one is 98.96 percent. The model has the best scores in predicting the Negative cases, with a precision of 100 percent. Nevertheless, Hyperthyroid has a comparatively lower accuracy indicating that the model is more likely to overestimate more cases of this category.

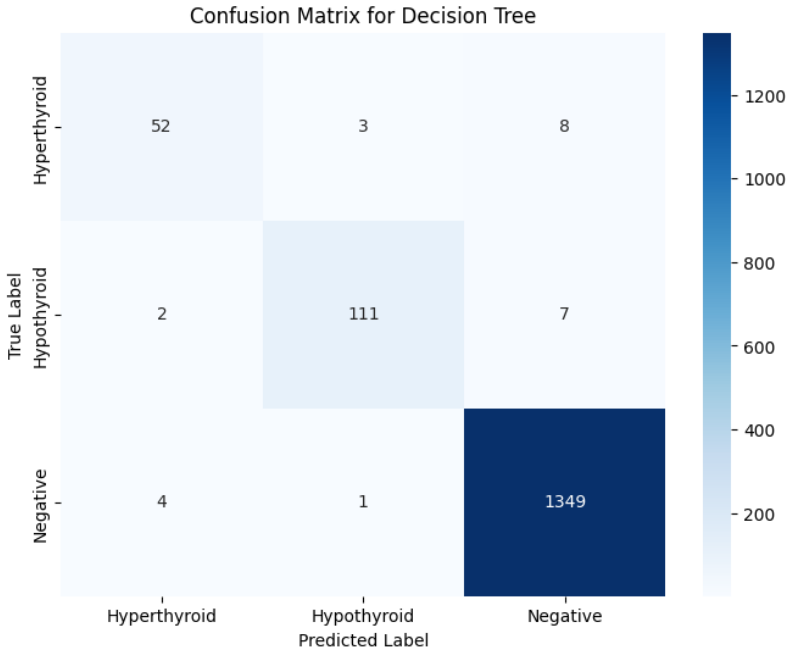


Fig 4.3: Confusion Matrix of Dicision Tree

These precision figures show that the model works extremely well at detecting Negative and Hypothyroid cases, but the lower figures in the Hyperthyroid case demonstrate that unusual hyperthyroidism cases may have been slightly under-detected.

In general, Decision Tree is a solid foundation (accuracy) to predict the screening results of thyroid diseases. It is rather precise in all fields except Hyperthyroid cases, marking a possible need to refine the models to detect less common cases.

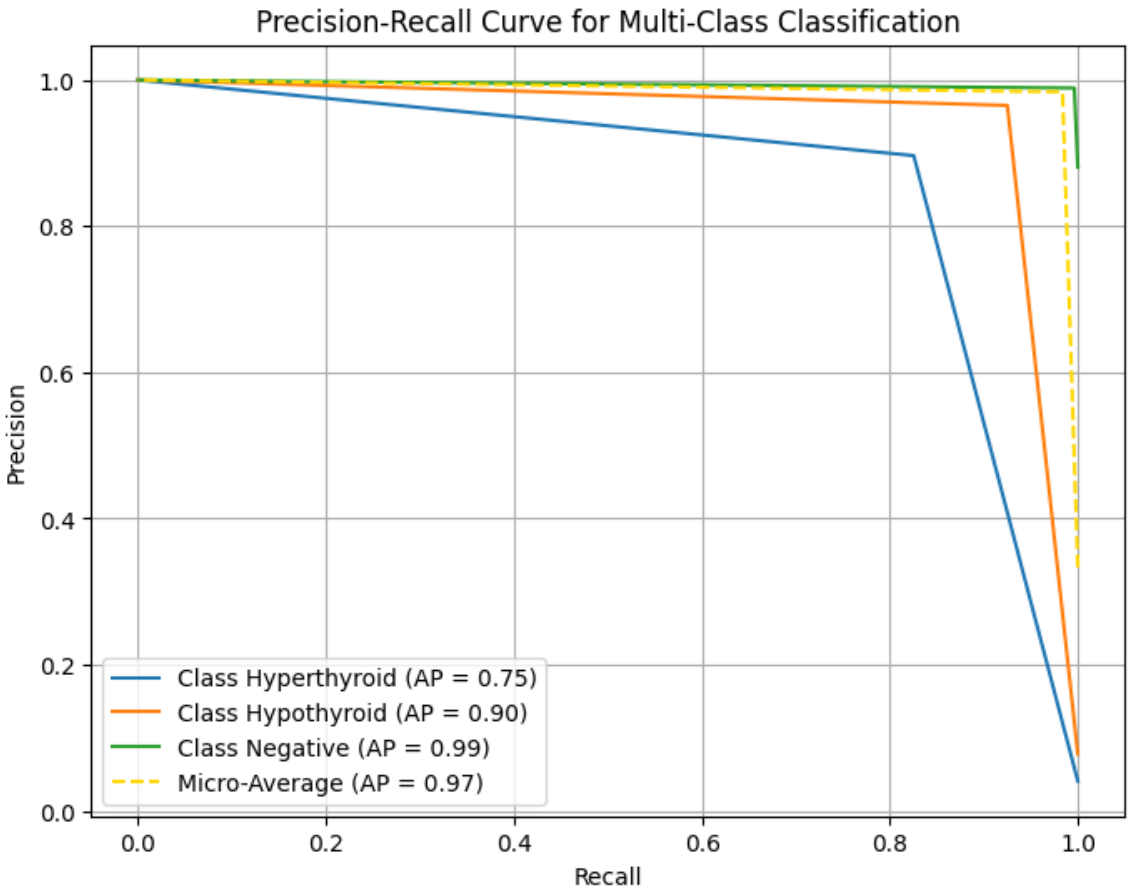


Fig 4.4: Precision vs Recall curve of Dicision Tree

4.3.3 Logistic Regression

In this paper, we examined the efficacy of Logistic Regression (LR) in classification of thyroid diseases (hyperthyroid, hypothyroid, and normal). Data was pre-processed by using standard scaling on all continuous variables, including TSH, T3, TT4, T4U, FTI and TBG important in thyroid diseases diagnosis.

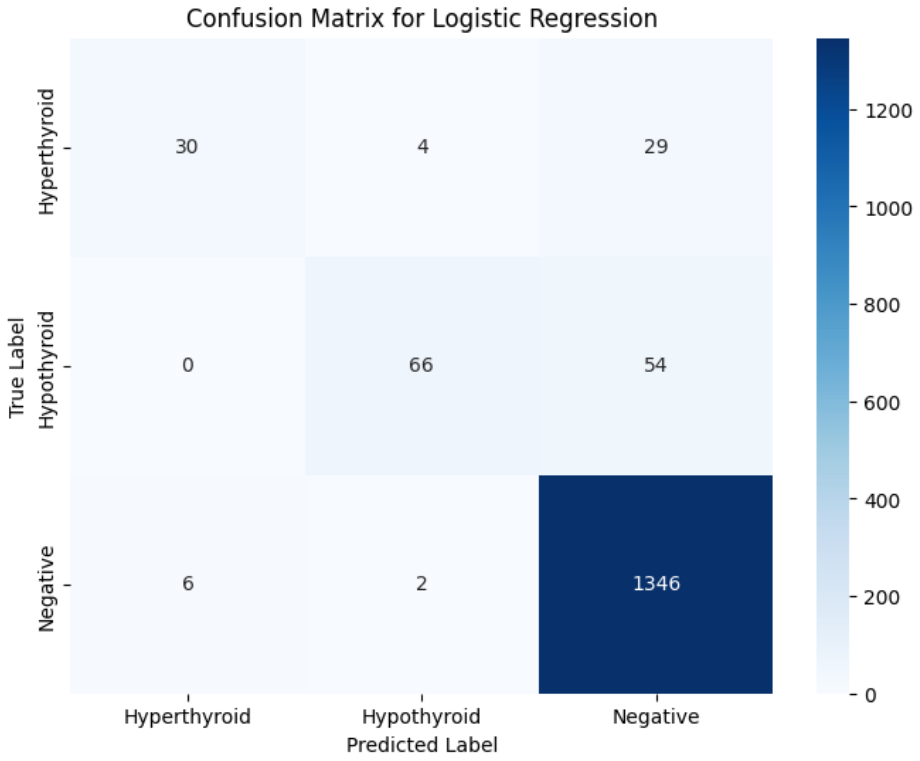


Fig 4.5: Confusion matrix of Logistic Regression

Transformed data was then used to train Logistic Regression that obtained an accuracy of 93.8%. A line of performance measures was then used to test the model such as the confusion matrix, the accuracy, and the average precision scores.

The confusion matrix of the model indicated a high level of accuracy in all classes with the LR model identifying a majority of cases. The high percentage of accuracy (93.8) shows that the model is precise in its prediction of this multi-class classification challenge.

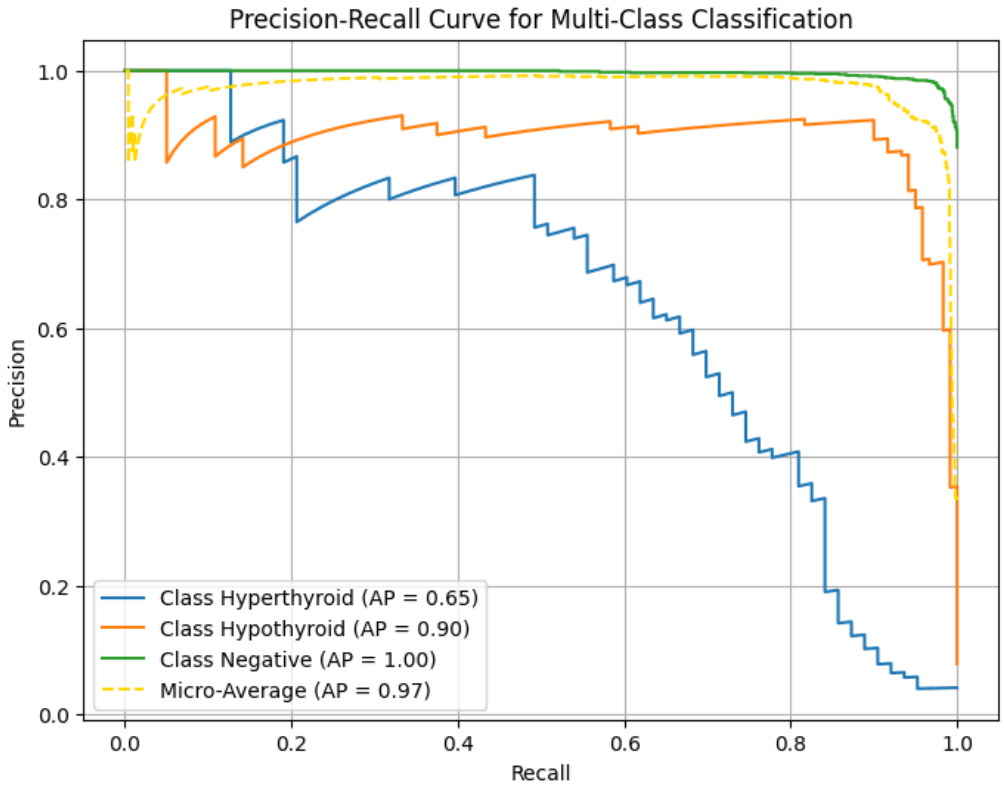


Fig 4.6: Precision vs Recall curve of Logistic Regression

4.3.4 Weighted Logistic Regression

To address the problem of the imbalance in the classes on the data, a Weighted Random Forest (RF) model was trained by changing the weighted class. In this way, the majority or other classes that were not underrepresented (Hyperthyroid and Hypothyroid) were Lighted up so the weigh of the classes would have been very high. Precisely, Hyperthyroid weight was stipulated 1500 times, Hypothyroid was stipulated to weight 100, and Negative was stipulated to weight 25. The arrangement of class weights to train the weighted RF model was to avoid the issue of class imbalance, so that the average accuracy of the classification among all the classes would be increased more.These weights enabled the weighted RF model to perform significantly much better than the unweighted RF in the problem of class imbalance with a general accuracy of 98.37%.

Through changes in the weights of the classes, the model could be more applicable in reflecting the performance of the minority classes of Hyperthyroid and Hypothyroid that are not found relevant (due to their smaller sample of the dataset). The methodology was an effective solution to the performance improvement of predictive accuracy in case of unequal classes.

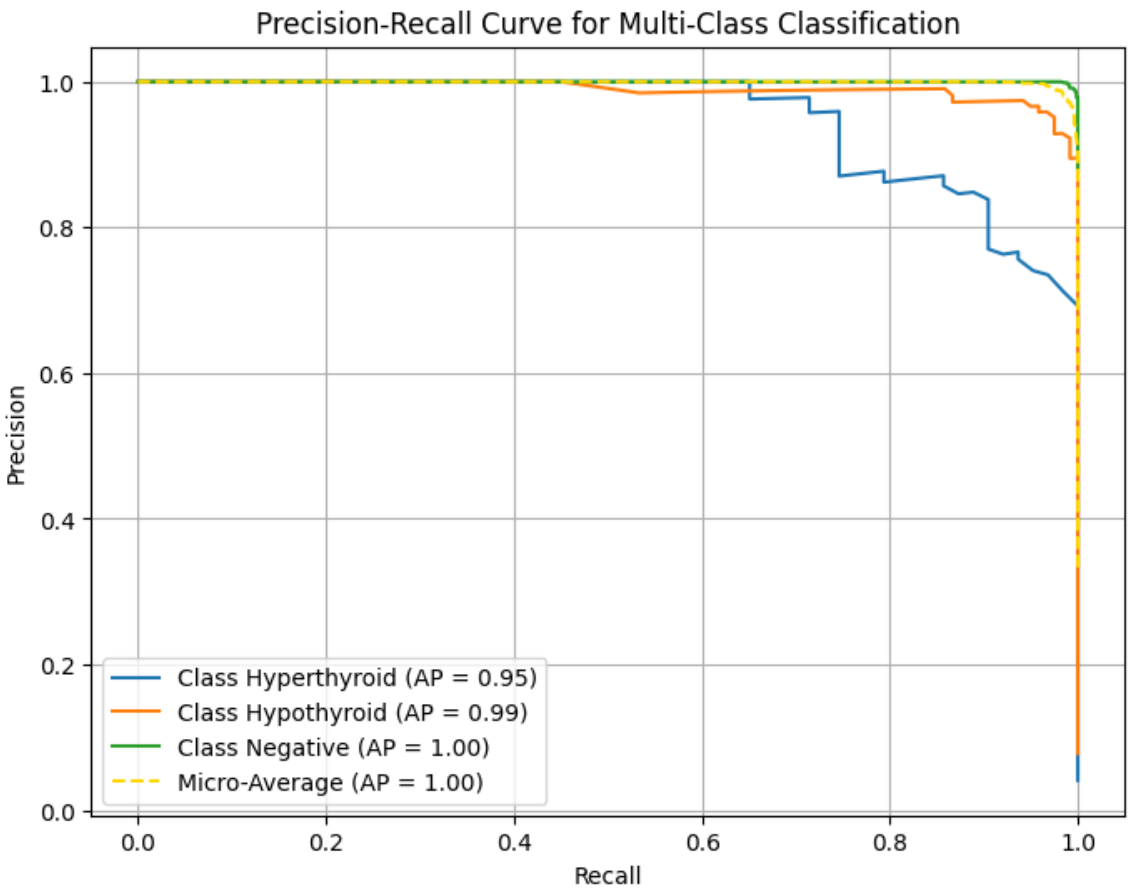


Fig 4.7: Precision vs recall curve of weighted Logistic Regression

4.3.5 Random Forest

The Random Forest (RF) classifier was used to test the classification performance on the pre-processed dataset of thyroid to compare its output with the three diagnostic classes, namely Hyperthyroid, Hypothyroid, and Negative. The model was fit with the default parameters and tested with the held-out test data using the training subset.

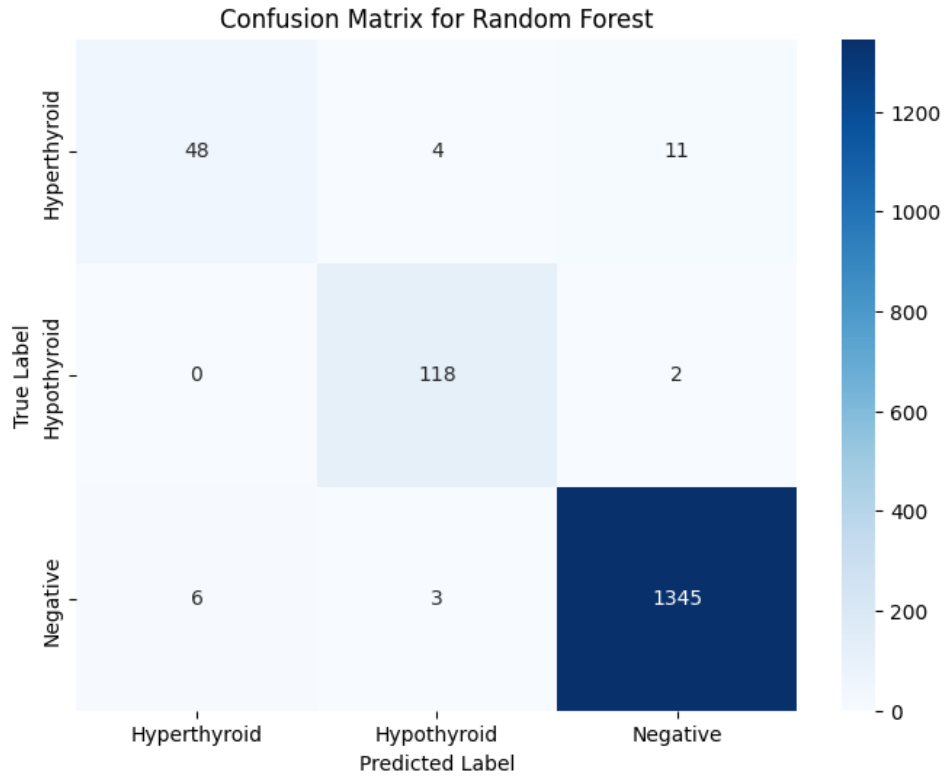


Fig 4.8: Confusion matrix of Random Forest

The goodness of the entire model was 98.30 percent corroborating its reliability to capture the occurrence of thyroid disorders. In another measure of its predictive capability, each class was assigned an Average Precision Score (APS). As shown by the model, the precision was high in each category; 0.94 precision (Hyperthyroid), 0.99 precision (Hypothyroid), and 0.99 precision (Negative). Also, the precision score was 0.999, which demonstrated the strength of the model in dealing with multiclass classification.

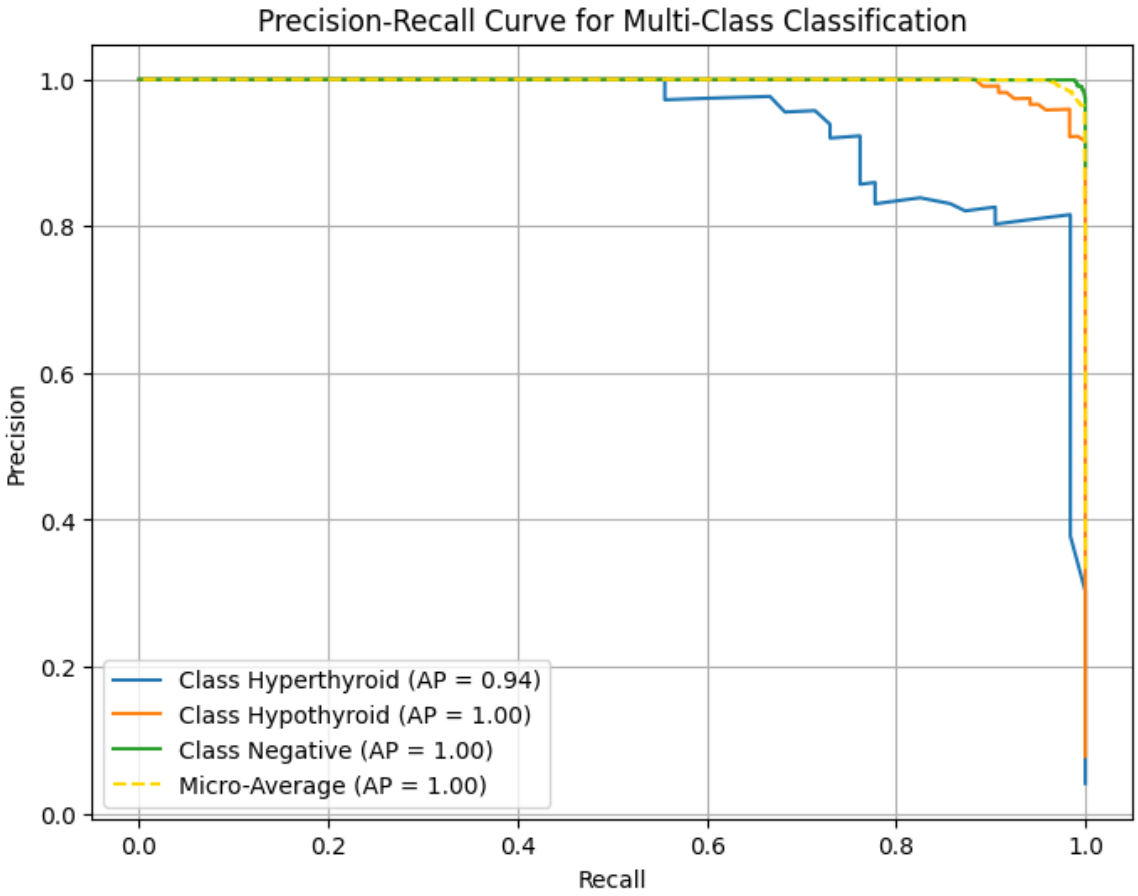


Fig 4.9: Precision vs recall curve of Random Forest classification

4.3.6 Gradient Boosting

The Gradient Boosting classifier also generated the confusion matrix on thyroid disease dataset as given below:

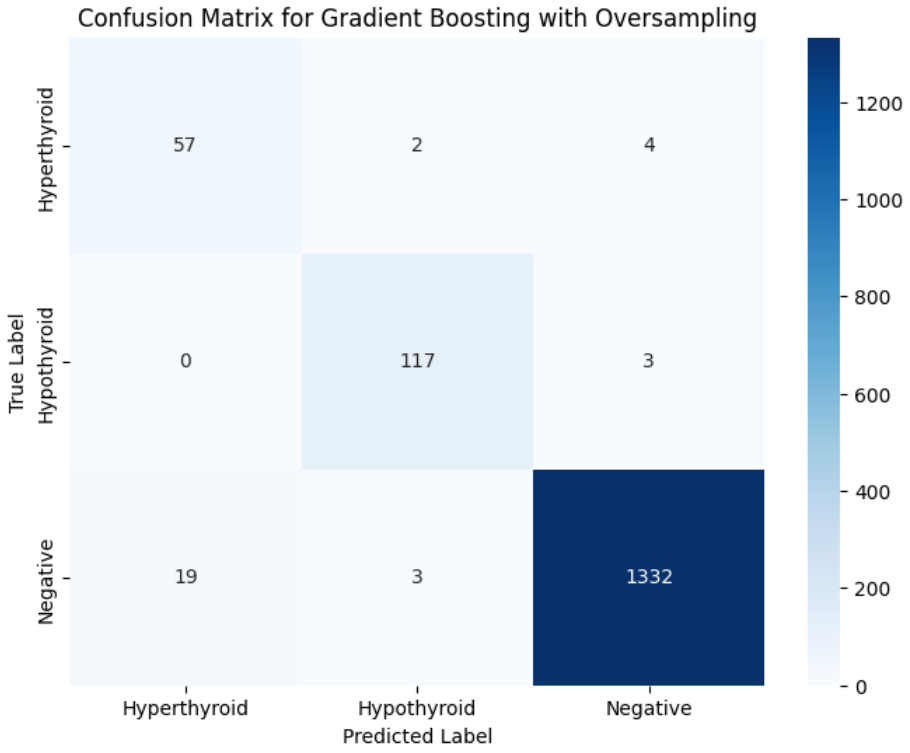


Fig 4.10: Confusion matrix of Gradient Boosting

This finding shows good detection of the Hypothyroid and Negative cases with only little misclassification detected across categories. The accuracy of the model was 97.98 percent. The average precision is 0.908, 0.995 and 0.999 by classes Hyperthyroid / Hypothyroid / Negative, and 0.996 is the micro-average precision. These results indicate that Gradient Boosting is quite competitive, especially in reliability performance in Hypothyroid and Negative class. Similar to the other models, however, it experienced minor shortcomings when fitting the Hyperthyroid cases because of their lower prevalence in the dataset.

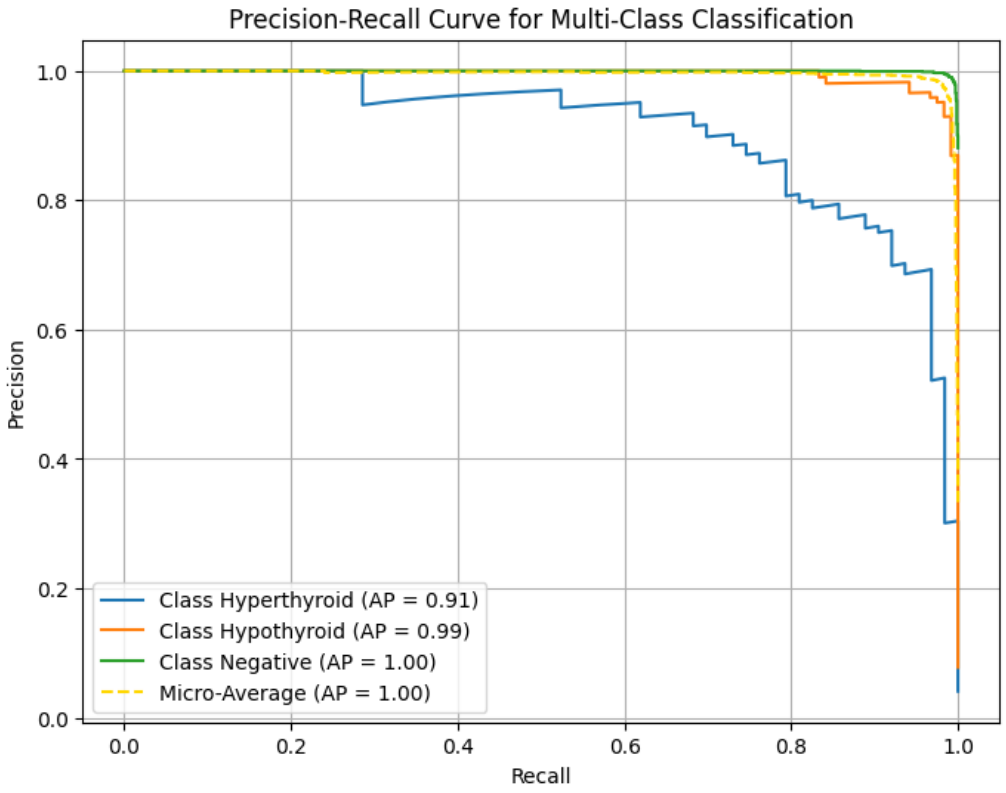


Fig 4.11: Precision vs Recall curve of Gradient Boosting

4.3.7 CatBoost Classifier

The CatBoost classifier was tested with both under-sampling and over-sampling strategies to address class imbalance within the dataset.

For the under-sampling approach, the confusion matrix was as follows:

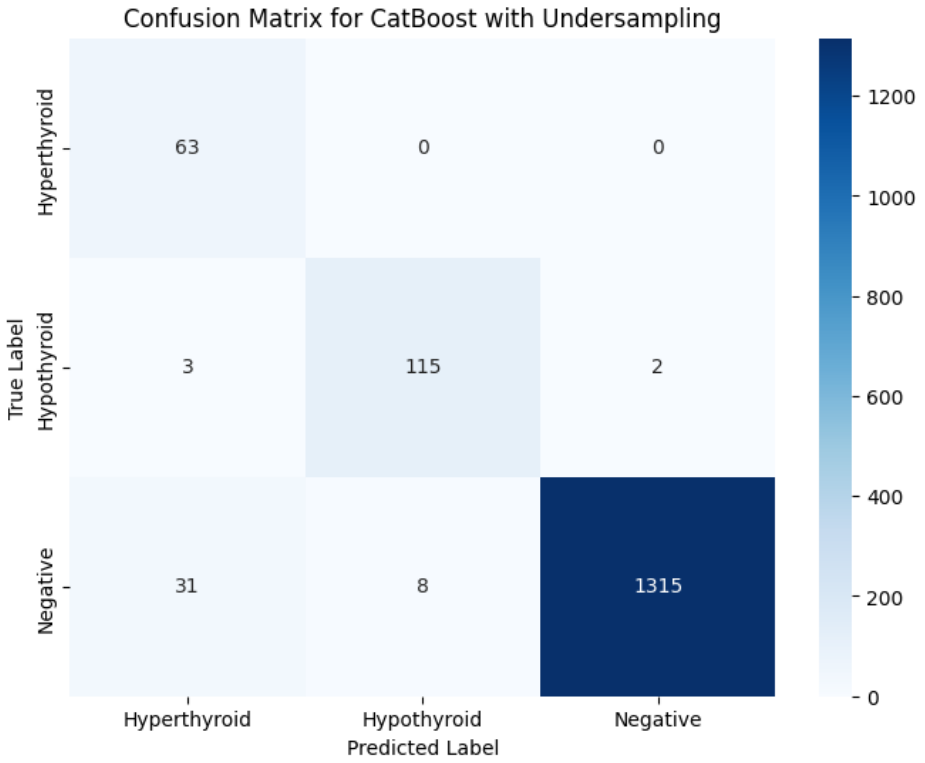


Fig 4.12: Catboost Confussion matrix (Undersampling)

The model achieved an accuracy of 97.13%, with class-wise average precision scores of 0.896 for Hyperthyroid, 0.981 for Hypothyroid, and 0.999 for Negative, alongside a micro-average precision of 0.992. Although the overall performance was strong, the model misclassified several Negative cases as Hyperthyroid, which slightly affected its balance.

In the over-sampling approach, the dataset distribution was expanded to balance all three classes, resulting in the following confusion matrix:

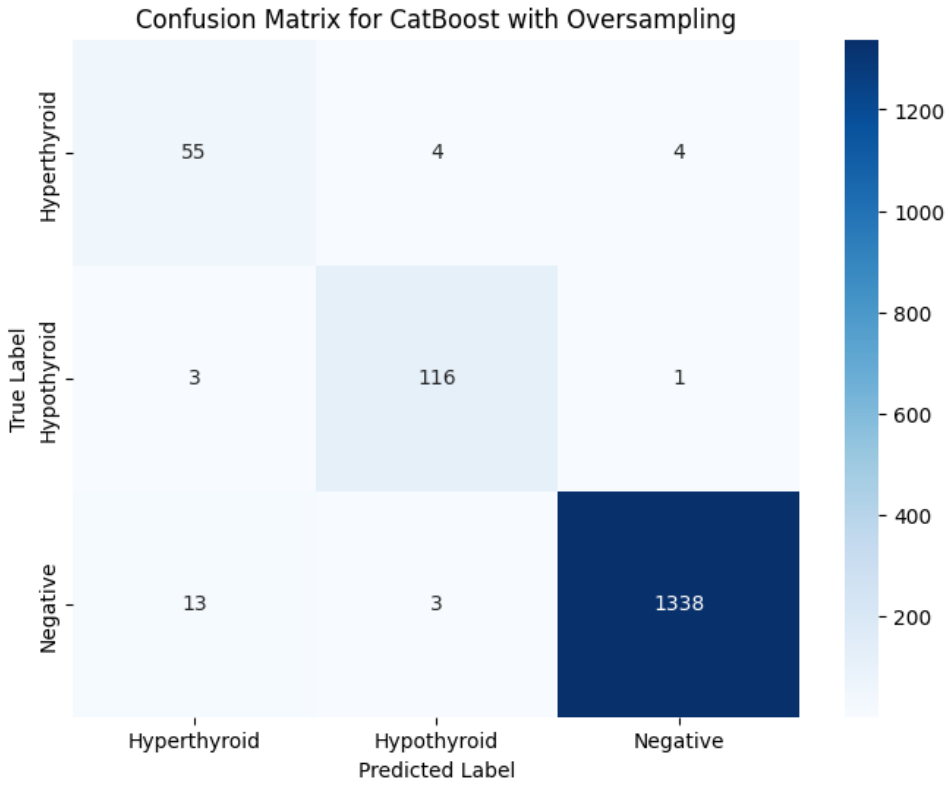


Fig 4.13: Catboost Confussion matrix (Over-sampling)

This approach improved the overall balance, with an accuracy of 98.17%. The class-wise average precision scores were 0.930 for Hyperthyroid, 0.994 for Hypothyroid, and 0.999 for Negative, with a micro-average precision of 0.999. These results demonstrate that over-sampling significantly enhanced the classification of the minority Hyperthyroid class while maintaining excellent accuracy for the other categories. Overall, CatBoost performed reliably under both strategies, with over-sampling providing the most balanced and accurate outcomes across classes.

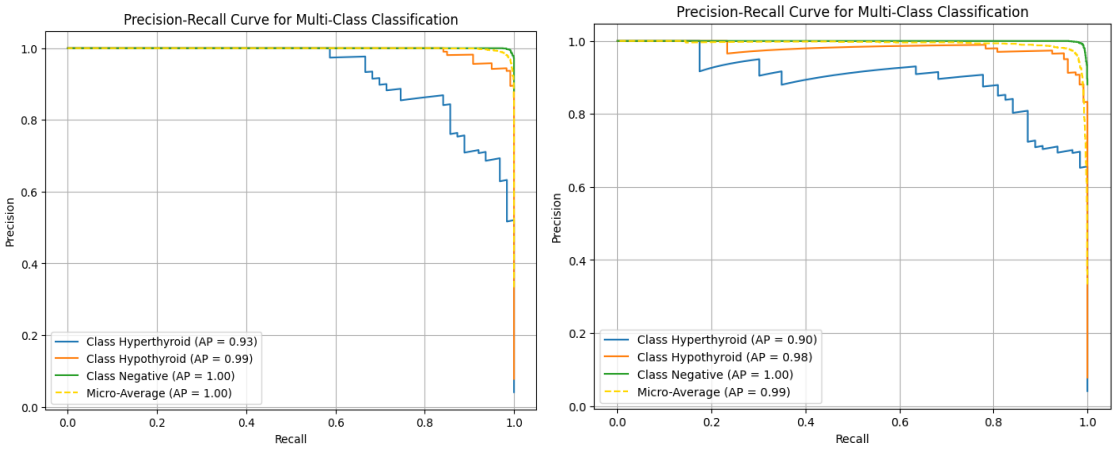


Fig 4.14: Precision vs Recall curve of CatBoost (Under(left) & over(Right) Sampling)

4.3.8 Best & Final Model

The XGBoost emerged to be the best performing and a final model by which the study was conducted. It achieved overall accuracy of 98.50 and a very high precision of 0.944, 0.996, and 0.999 for respectively Hyperthyroid, Hypothyroid, and Negative with a micro-average precision 0.999. The confusion matrix showed that the misclassifications were insignificant and, in particular, the majority and minority categories were also recognized well.

Compared with them, Gradient Boosting and CatBoost produced less balanced, although very good results. Gradient Boosting performed well in terms of high accuracy but failed slightly with Hyperthyroid cases whereas CatBoost required implementation of over sampling to increase the minority-class detection; this brought in extra computational time. Compared to XGBoost, however, this did not require excessive reliance on resampling, and it also attained better precision.

That is why the XGBoost was chosen as the most efficient one. With a good predictive accuracy, stability and adaptiveness, and the capacity to work with Explainable AI approaches, it was found that it is the most trusted clinical tool in predicting thyroid diseases in this study.

4.4 Discussion

The performed experiment underlines the relative efficacy of various machine learning algorithms in predicting thyroid disease. Of the models considered, XGBoost showed superior overall accuracy of 98.50 and achieved near-perfect accuracy in both the majority and minority classes, i.e. Hyperthyroid. This indicates that, the XGBoost boosting-based ensemble technique is quite effective in identifying intricate decision boundaries and handling class imbalance without considerable loss in performance.

In that case, the Decision Tree model received an accuracy of 98.37% with an effective classification on Negative and Hypothyroid classes but not necessarily on Hyperthyroid, which demonstrated sensitivity to class-imbalance. In spite of its effectiveness as a baseline model, it is simple and thus cannot represent the minority class fully.

The tested CatBoost method demonstrated valuable lessons in how to balance the problem of imbalance on medical data. The under-sampling technique had an accuracy of 97.13% but due to majority-class data loss, there were misclassification especially in the Negative category. On the contrary at the same time, over- sampling raised the general accuracy of the model to 98.17 and boosted the accuracy of Hyperthyroid detection pointing out that even distribution in the data makes the model learn to generalize much better in all classes.

The Gradient Boosting model attained accuracy of 97.98%, which was competitive and provided good results with the Hypothyroid and Negative cases. It, however, like the Decision Tree, did not differentiate between cases of Hyperthyroid with similar accuracy as XGBoost or CatBoost with over-sampling. This is the mirror of the long-standing issue of class imbalance, in particular, when minority categories encompass patterns, which are similar to those of majority classes.

In general, the comparative analysis explains that the boosting-based ensemble methods are more accurate and precise than single classifiers especially in imbalanced medical data sets. The best performing model was XGBoost followed by CatBoost over-sampling,

which also showed improvement in minority classes. The results indicate the priority of not only the algorithm selection but also the data balancing processes in medical spaces to ensure that the AI system serves well overall, but also does not discriminate in accuracy, consistency, and reliability matters across all undertaken diagnostic categories.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

The implications of introducing Explainable Artificial Intelligence (XAI) into thyroid disease prediction are weighty, particularly in healthcare. Accurate and interpretable AI models advance early detection of diseases such as hypothyroidism and hyperthyroidism, and, as a result, allow starting treatment in time and achieving better patient outcomes. Explainable AI, e.g., SHAP values, can provide transparency to clinicians and patients on how decisions are reached and lead to better trust and encourage the implementation of AI-assisted technology. The transparency enables a bridge between sophisticated algorithms and real life clinical practice. What is more, more accessible AI-based diagnostic tools can help address healthcare disparities, especially in areas that have low access to specialists. The study also offers information on the patterns of diseases, which contributes to medical building and more rational decisions of healthcare professionals.

5.2 Impact on Environment

The energy and computing power requirements have been an area of concern to AI research, although in general the low overall environmental impact of the research activity could be considered a bonus. By utilizing the optimized models such as XGBoost and decision trees, as well as cloud-based computing power, the amount of energy utilised was minimised and avoided unnecessary load to reach hardware. The project will also perform sustainable computational practices since it focuses on lightweight and scalable models. Moreover, the decision material of AI-based diagnostics is in the digital form, which eliminates dependency on labor- and material-consuming laboratory tests, making their total ecologic footprint less.

5.3 Ethical Aspects

Ethics were core during this research The privacy of the patient and information confidentiality was strictly observed and applicable conventions and regulations were followed. Explainability as a mitigation means removes ethical risks because it enables clinicians and patients to interpret model predictions, which eliminates the risks of the black-box decision. Fairness and bias have been paid specific attention to without discriminating on large subsets of the patient population and closer observation should be maintained to eliminate any unintentional effects. Strong ethical principles by incorporating transparency, accountability and by the promise of using AI in healthcare in a responsible manner.

5.4 Sustainability Plan

Unless it undergoes upgrades, incorporates, and gets deployed to clinical processes and also cooperates with stakeholders, this AI-based diagnostic tool cannot be maintained in the long term. Clinical information would periodically update the model to ensure that it is updated and precise. Liaison with healthcare facilities will facilitate easy implementation, but expert training on the clinicians will facilitate the esteem in the interpretability of the elucidable design of AI outputs. Open access to code and findings will promote additionally and increased applications. Green computing is also combined with energy efficient and scalable cloud facilities. Overall, the project bases itself more on flexibility, ethical responsibility, and healthcare quality sustainability as well as sets up the system of further AI-assisted diagnostic.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

The main aim of the study was to prepare the model of machine learning that would aid in early diagnosis and classification of thyroid disease. The analysis made use of the publicly available thyroid dataset that has various clinical and laboratory characteristics associated with the diagnosis of hyperthyroidism, hypothyroidism, and other negative samples. A well-defined preprocessing pipeline, such as categorical encoding, data balancing, and normalization, has been applied to ensure the good quality of data and to reduce possible biases.

Several machine learning models were applied and compared such as Logistic Regression, Weighted Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, CatBoost, and XGBoost. The metric of performance was conducted on parameters of accuracy precisions, recall and F1-score. XGBoost surpassed other models in terms of accuracy and robustness in all cases proving its capacity to capture multifaces nonlinear relationships in the dataset.

Moreover, to increase transparency, SHAP, an Explainable AI (XAI) method was added. The tools have helped to give a sensible indication toward the significance of the presence of features including a test of Thyroid Stimulating Hormone, Thyroid Triiodothyronine, and Thyroid Tetraiodothyronine levels, which further add to the interpretability of predictive results. This combination of precision with explainability helps AI-assisted diagnosis in healthcare have a more profound role than it would otherwise have, as the findings are not only accurate and scientific but they also can be applied in practical terms to improve patient outcomes.

6.2 Conclusions

The analysis indicates that XGBoost is a viable and stable model that contributes to the prediction of thyroid diseases better within the tested group of classical and ensemble-based algorithms in this study. Its high performance shows that high-performance methods of advanced gradient boosting can effectively deal with complex medical data, providing a high level of accuracy and generalizability.

Of equal significance, the infusion of explainable AI solved a decades-old interpretability problem in machine learning. By determining how much each of the features contributed to generating predictions, clinicians were offered part visibility in the model and not blacked out. Not only does this enhance faith in AI-driven solutions, but helps provide more informed and confident clinical decisions.

On the whole, the results indicate that usage of machine learning, in combination with interpretability frameworks, shows a huge potential to enhance thyroid disorder detection. Although the study was done on a single dataset only, the outcomes are very convincing that these models can be adopted in real world clinical settings with additional verification.

6.3 Implication for Further Study

There are some opportunities that can be explored in future even though this study yielded good results. The next possible path is to test the suggested XGBoost-based framework on bigger and more varied datasets obtained in different demographic and clinical populations. This would contribute to the generalizability of the results and minimize possible selectivity bias in data.

A potential avenue of future research is the application of more putting into deep learning models, like recurrent neural networks (RNNs) or convolutional neural networks (CNNs) that accommodate sequence data or image, like ultrasound images, respectively. These

types of models can be used in complementing the structured data based approaches exploited in this study.

What is more, explainable AI might become even more intuitive with the help of more advanced methods that will enable clinicians to obtain even more clear visualisations and domain-specific rules. One should also remember about ethical aspects, such as the confidentiality of the patients, equity across the different groups in the population, and the affirmation of the clinical consistency in clinical practice.

However, in conclusion, the paper has shown that XGBoost combined with explainable AI offers an effective yet explainable technique of predicting thyroid diseases. An expansion of the current process to larger datasets, new modeling strategies, and clinical-level integration can be the next steps to provide more solid healthcare solutions and become more accessible.

REFERENCES

[1] Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2021). Thyroid disease prediction using machine learning approaches. National Academy Science Letters, 44(3), 233-238.

[2] Sankar, S., Potti, A., Chandrika, G. N., & Ramasubbareddy, S. (2022). Thyroid disease prediction using XGBoost algorithms. J. Mob. Multimed, 18(3), 1-18.

[3] Hosseinzadeh, M., Ahmed, O. H., Ghafour, M. Y., Safara, F., Hama, H. K., Ali, S., ... & Chiang, H. S. (2021). A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. The Journal of Supercomputing, 77, 3616-3637.

[4] Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J. L. V., Rodríguez, C. L., & Ashraf, I. (2022). Thyroid disease prediction using selective features and machine learning techniques. Cancers, 14(16), 3914.

[5] Riajuliislam, M., Rahim, K. Z., & Mahmud, A. (2021, February). Prediction of thyroid disease (hypothyroid) in early stage using feature selection and classification techniques. In 2021 International conference on information and communication technology for sustainable development (ICICT4SD) (pp. 60-64). IEEE.

[6] Abbad Ur Rehman, H., Lin, C. Y., & Mushtaq, Z. (2021). Effective K-nearest neighbor algorithms performance analysis of thyroid disease. Journal of the Chinese Institute of Engineers, 44(1), 77-87.

[7] Sonuç, E. (2021, July). Thyroid disease classification using machine learning algorithms. In Journal of Physics: Conference Series (Vol. 1963, No. 1, p. 012140). IOP Publishing.

[8] Razia, S., Prathyusha, P. S., Krishna, N. V., & Sumana, N. S. (2018). A Comparative study of machine learning algorithms on thyroid disease prediction. Int. J. Eng. Technol, 7(2.8), 315-319.

[9] Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & De Albuquerque, V. H. C. (2020). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The journal of supercomputing, 76, 1128-1143.

[10] Jha, R., Bhattacharjee, V., & Mustafi, A. (2022). Increasing the prediction accuracy for thyroid disease: a step towards better health for society. Wireless Personal Communications, 122(2), 1921-1938.

[11] S. Islam, M. Haque, M. Miah, … T. S.-P. C., and undefined 2022, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," peerj.comSS Islam, MS Haque, MSU Miah, TB Sarwar, R NugrahaPeerJ Computer Science, 2022•peerj.com, Accessed: Jan. 16, 2024. [Online]. Available: https://peerj.com/articles/cs-898/

[12] T. Singh, A. Sahu, S. Dubey, … M. S.-I. J. of, and undefined 2022, "Treatment of thyroid disease through machine learning predictive model," academia.eduT Singh, AK Sahu, S Dubey, MP Sharma, S Verma, C KumarInternational Journal of Health Sciences, 2022•academia.edu, Accessed: Jan. 16, 2024. [Online].Available:https://www.academia.edu/download/93461396/IJHS-12813_3176-3188.pdf_filename_UTF-8IJHS-12813_3176-3188.pdf

[13] S. Verma, R. Popli, and … H. K.-I. J. of, "Classification of thyroid diseases using machine learning frameworks," pdfs.semanticscholar.orgS Verma, R Popli, H Kumar, A Srivastava International journal of Health Sciences•pdfs.semanticscholar.org, Accessed: Jan. 16, 2024. [Online]Available:https://pdfs.semanticscholar.org/70cb/f0aec6ac8c10035d98957006b32a7289f4a5.pdf

[14] Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., & Ahmad, A. (2022). [Retracted] Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. BioMed Research International, 2022(1), 9809932.

[15] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., & Lipton, Z. C. (2018). Explaining explanations: An overview of interpretability of machine learning. *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning*.

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4765-4774.

[18] "Empirical comparison of bagging-based ensemble classifiers | IEEE Conference Publication | IEEE Xplore." Accessed: Jan. 16, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6289900

[19] S. Bera, V. S.-I. J. of R. Sensing, and undefined 2020, "Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification,"

Taylor & FrancisS Bera, VK ShrivastavaInternational Journal of Remote Sensing, 2020•Taylor & Francis, vol. 41, no. 7, pp. 2664–2683, Apr. 2020, doi: 10.1080/01431161.2019.1694725.

[20] A. Gulli, A. Kapoor, and S. Pal, Deep learning with TensorFlow 2 and Keras: regression, ConvNets, GANs, RNNs, NLP, and more with TensorFlow 2 and the Keras API. 2019. Accessed: Jan. 16, 2024. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=BVnHDwAAQBAJ&oi=fnd&pg=PP1&dq=Gulli,+A.,+Kapoor,+A.,+%26+Pal,+S.+(2019).+Deep+learning+with+TensorFlow+2+and+Keras:+regression,+ConvNets,+GANs,+RNNs,+NLP,+and+more+with+TensorFlow+2+and+the+Keras+API.+Packt+Publishing+Ltd.&ots=K-r89qTs0W&sig=OUPstvlspGzudRKeTbkJ3d14kdE

[21] "Applied Deep Learning - Part 1: Artificial Neural Networks | by Arden Dertat | Towards Data Science." Accessed: Jan. 16, 2024. [Online]. Available: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

[22] A. P.-M. and Machines and undefined 2019, "The pragmatic turn in explainable artificial intelligence (XAI)," SpringerA PáezMinds and Machines, 2019•Springer, vol. 29, no. 3, pp. 441–459, Sep. 2019, doi: 10.1007/s11023-019-09502-w.

[23] M. Hossin, & M. S.-I. journal of data mining, and undefined 2015, "A review on evaluation metrics for data classification evaluations," academia.eduM Hossin, MN SulaimanInternational journal of data mining & knowledge management process, 2015•academia.edu, Accessed: Jan. 16, 2024. [Online]. Available: https://www.academia.edu/download/37219940/5215ijdkp01.pdf

[24] "What is a confusion matrix?. Everything you Should Know about… | by Anuganti Suresh | Analytics Vidhya | Medium." Accessed: Jan. 16, 2024. [Online]. Available: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

[25] A. Gummadi and D. Reddy, "A Novel Machine Learning Framework for Prediction of Early-Stage Thyroid Disease Using Classification Techniques," 2022, Accessed: Jan. 16, 2024. [Online]. Available: https://www.academia.edu/download/90634911/9615.pdf

[26] GeeksforGeeks. (2025, July 23). *CatBoost in Machine Learning*. https://www.geeksforgeeks.org/machine-learning/catboost-ml/.

[27] A. Sreekumar, "Decision Tree: ID3 Algorithm," Medium, May 29, 2023. [Online]. Available: https://medium.com/@anirudhsreekumar98/decision-tree-id3-algorithm-f74875fa507d.[Accessed:17-Sep-2025].

[28] A. Sharma, “Decision Tree vs Random Forest—Which Algorithm Should You Use?,” Analytics Vidhya, May 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

. [Accessed: 17-Sep-2025].

[29] Ensemble learning for the early prediction of neonatal jaundice with genetic features - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/The-architecture-of-Gradient-Boosting-Decision-Tree_fig2_356698772 [accessed 17 Sept 2025]

[30] P. Aggarwal, “Implementation of XGBoost (eXtreme Gradient Boosting),” GeeksforGeeks, 05 Sep 2025.[Online].Available:https://www.geeksforgeeks.org/machine-learning/implementation-of-xgboost-extreme-gradient-boosting/. [Accessed: 17-Sep-2025].

242-25-039

ORIGINALITY REPORT

17%	14%	11%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	iieta.org Internet Source	1%
4	ebin.pub Internet Source	1%
5	journals.riverpublishers.com Internet Source	<1%
6	Submitted to University of Southampton Student Paper	<1%
7	www.iieta.org Internet Source	<1%
8	www.frontiersin.org Internet Source	<1%
9	www.mdpi.com Internet Source	<1%
10	S.P. Jani, M. Adam Khan. "Applications of AI in Smart Technologies and Manufacturing", CRC Press, 2025 Publication	<1%
11	bpasjournals.com Internet Source	<1%