Thyroid Disease Detection Using Machine Learning

By
Md Mostakim Ahmed
213-15-4426

Shamira Shams Shathy
213-15-4427

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised by
Ms. Shayla Sharmin
Lecturer (Senior Scale)
Department of Computer Science and EngineeringDaffodil International University

Co-Supervised by
Ms. Hasnur Jahan
Lecturer
Department of Computer Science and EngineeringDaffodil International University

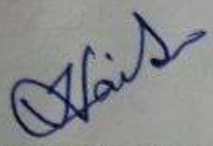
DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September 16, 2025

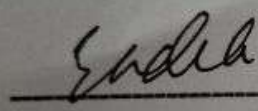
APPROVAL

This Project titled **"Thyroid Disease Detection Using Machine Learning"**, submitted by Md Mostakim Ahmed , ID No: **213-15-4426** and Shamira Shams Shathy , ID No: **213-15-4427** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **16 September, 2025**.

BOARD OF EXAMINERS

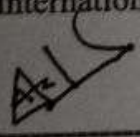


Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Md. Sadekur Rahman — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Saiful Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Dr. Md. Arshad Ali — **External Examiner**
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology
University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Shayla Sharmin Lecturer (Senior Scale)**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Shayla 14.9.25

Shayla Sharmin
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Hasnur 15.9.25

Hasnur Jahan
Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Md. Mostakim Ahmed

Md Mostakim Ahmed
Student ID:213-15-4426
Department of Computer Science and Engineering
Daffodil International University

Shamira Shams Shathy

Shamira Shams Shathy
Student ID:213-15-4427
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Shayla Sharmin**, **Senior Lecturer**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor 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, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to the Head of the Department of Com- puter Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

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

ABSTRACT

Thyroid disorders, such as hypothyroidism and hyperthyroidism, are challenging to diagnose due to overlapping symptoms like fatigue and weight changes, compounded by inconsistent medical data. This study leverages machine learning to enhance thyroid disease detection using two robust datasets: the Kaggle Thyroid Disease Dataset (9,172 records, 31 features) and the UCI Thyroid Disease Dataset (2,801 instances, 29 attributes). For the Kaggle dataset, a CatBoost classifier was developed after rigorous preprocessing, including data cleaning, zero imputation, one-hot encoding, and SMOTE with undersampling to address class imbalance. The optimized CatBoost model, incorporating L2 regularization and balanced class weights, achieved 98.70% accuracy, 98.79% precision (measuring correct positive predictions), and 97% Area Under the Precision-Recall Curve (AU-PRC) for hyperthyroidism, surpassing prior benchmarks by 2-3%. For the UCI dataset, Decision Tree and Random Forest classifiers were built following median/mode imputation, label encoding, feature scaling, and SMOTE. The Decision Tree excelled with 99.11% accuracy, 99.12% precision, 99.11% recall, 99.07% F1-score, and 98.53% (±0.36%) cross-validation accuracy, outperforming Random Forest (98.04% accuracy, 98.44% ±0.14% cross-validation) and existing studies. Feature importance, elucidated by Shapley Additive Explanations (SHAP, a method for interpreting model predictions), identified T3, TT4, T4U, FTI, and TSH as critical predictors, offering transparent insights for clinicians. Despite strengths, limitations include potential dataset biases and the need for real-world validation. Excellent accuracy and interpretability are demonstrated by these tree-based models, which reduce the risk of misdiagnosis and pave the way for ethical deployment in healthcare. SHAP also ensures clear and trustworthy clinical decision support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Untreated hyperthyroidism and hypothyroidism, two common thyroid disorders, can have serious health consequences. Traditional diagnosis using TSH, T3, and TT4 assays is often time-consuming and error-prone.A faster and more accurate alternative is offered by machine learning, despite its shortcomings, which include missing values, data imbalance, and lack of interpretability.This project uses Kaggle and UCI datasets to incorporate SHAP, ensemble models, and preprocessing into an explainable machine learning framework to increase accuracy and clinician trust.[1]

1.1 Introduction:

Thyroid conditions, such as hypothyroidism and hyperthyroidism, are among the most prevalent endocrine disorders and significantly impact overall health and metabolic control. If left misdiagnosed or untreated, they can lead to major adverse effects like infertility, mental health issues, and cardiovascular disease. TSH, T3, and TT4 levels are examples of biochemical tests that are employed in traditional diagnostic procedures.[2] These tests can be costly, time-consuming, and prone to human interpretation errors. As organized medical data becomes more accessible, machine learning, or ML, presents a promising avenue for the early, accurate, and efficient diagnosis of thyroid diseases.The actual clinical use of ML in thyroid diagnosis is, however, hampered by several factors.[3] These include class imbalance in datasets, noise and missing values in medical records, and the black-box character of many predictive models, which jeopardizes physician trust and interpretability.[4] Therefore, there is an increasing need for explainable machine learning methods that are not only accurate but also clear, equitable, and generalizable. A comprehensive machine learning framework for thyroid illness classification is proposed in this paper, which validates the models' robustness and generalizability using two benchmark datasets (Kaggle and UCI). The framework uses ensemble and tree-based classifiers, advanced data preparation, and SHAP-based interpretability to ensure both clinical application and prediction performance.[5]

Problem Statement:

Accurate and early diagnosis of thyroid problems remains a continuing challenge in clinical practice because of the limitations of traditional diagnostic approaches, noisy and unbalanced datasets, and overlapping symptoms. The findings of biochemical testing and the subjective and erratic interpretation of experts are often the foundation of current diagnostic methods. Additionally, the vast majority of thyroid datasets currently accessible exhibit a significant class imbalance, with a greater proportion of healthy cases than illness instances. As a result, biased machine learning models are produced that cannot effectively generalize to minority (disease) groups. Furthermore, medical practitioners find it challenging to interpret the many opaque and

complex machine learning models utilized in this field. Therefore, it is essential to develop a machine learning approach that addresses data imbalance, improves classification performance, and provides interpretable results that align with medical knowledge.[2]

Contributions:

To address these challenges, this study makes the following key contributions: Dual-Dataset Evaluation: This work leverages two widely used benchmark datasets—Kaggle (9,172 records, 31 features) and UCI (2801 records, 29 features) to evaluate model generalizability across different data distributions and feature sets. Advanced Data Preprocessing: The preprocessing pipeline includes robust data cleaning, handling of missing values using zero and median/mode imputation, one-hot and label encoding, feature scaling, and the application of SMOTE to balance the class distribution. Hybrid Model Development: For the Kaggle dataset, a regularized CatBoost classifier with class weight balancing achieves 98.70% accuracy and 98.79% precision. For the UCI dataset, a simple yet effective Decision Tree model outperforms more complex classifiers, attaining 99.11% accuracy and 99.07% F1-score. Integration of Explainable AI: SHAP (SHapley Additive explanations) is employed to interpret the models' predictions, identifying the most influential features such as T3, TT4, TSH, FTI, and T4U. This enhances transparency and facilitates clinical interpretation. Clinical Relevance and Deployment Potential: The study demonstrates that interpretable, tree-based ML models can provide accurate and efficient diagnostic support. Their explainability and high performance make them suitable for integration into the real-time healthcare system.

1.2 Motivation:

Thyroid disorders particularly hypothyroidism and hyperthyroidism present a formidable diagnostic challenge in clinical practice. Their symptoms, such as fatigue, weight fluctuations, depression, and cognitive slowing, are notoriously non-specific and often mimic other common conditions like anemia, chronic stress, or psychiatric disorders. This diagnostic ambiguity frequently leads to delayed or misdiagnosed cases, with patients enduring prolonged discomfort and increased risk of complications such as cardiovascular disease or infertility. While serum biomarkers like Thyroid-Stimulating Hormone (TSH), Triiodothyronine (T3), and Total Thyroxine (TT4) form the cornerstone of laboratory diagnosis, their interpretation is not always straightforward. Results can vary significantly across laboratories, be influenced by medications or comorbidities, and require expert endocrinological judgment, a resource often scarce in primary care or low-resource settings. From a machine learning perspective, thyroid disease detection encapsulates a rich constellation of real-world data science challenges. First, medical datasets are inherently imbalanced, with hyperthyroid cases often vastly outnumbered by healthy or hypothyroid records, a bias that can cripple model sensitivity if left unaddressed. Second, missing and noisy data are pervasive in electronic health records, demanding intelligent imputation strategies and robust preprocessing pipelines. Third, the risk of model overfitting looms large, especially given the relatively small sample sizes of curated medical datasets compared to industrial-scale data. Finally, and perhaps most critically, the "black-box" nature of many high-performing algorithms undermines clinical adoption; physicians are unlikely to trust let alone act upon predictions they cannot understand or explain to patients. These challenges collectively motivate the core technical pillars of this research: the application of SMOTE-ENN

for intelligent class rebalancing, the comparative evaluation of interpretable yet powerful classifiers like CatBoost, Random Forest, and Decision Trees, and the integration of SHAP-based explainability to illuminate feature contributions and foster clinical trust. Beyond technical rigor, this project holds deep personal significance. It represents a hands-on immersion into the full lifecycle of responsible, real-world ML from messy, incomplete data to validated, interpretable models. It sharpens my competencies in ethical AI design, healthcare analytics, and human-centered machine learning skills I intend to carry forward into a career dedicated to building intelligent, equitable, and clinically grounded medical decision-support systems.

1.3 Objectives :

The primary objective of this research is to design, implement, and validate a robust, accurate, and interpretable machine learning (ML) framework for the automated detection of thyroid disorders, specifically hypothyroidism and hyperthyroidism. Thyroid diseases represent some of the most prevalent endocrine conditions globally, yet their clinical diagnosis remains fraught with challenges due to non-specific and overlapping symptoms such as fatigue, unexplained weight fluctuations, mood disturbances, and depression. Conventional diagnostic pathways rely heavily on biochemical assays including serum levels of TSH, T3, and TT4 which, while informative, are often constrained by cost, turnaround time, inter-laboratory variability, and subjectivity in interpretation. To overcome these limitations, this study proposes a data-driven ML approach capable of delivering rapid, reliable, and clinically actionable predictions. We rigorously evaluate our framework on two widely recognized benchmark datasets: the Kaggle Thyroid Disease Dataset (9,172 records, 31 features) and the UCI Thyroid Disease Dataset (2,801 instances, 29 attributes). This dual-dataset strategy ensures model generalizability across diverse patient populations, clinical settings, and data collection protocols a critical step toward real-world deployment.A core technical focus of this work is the implementation of advanced preprocessing pipelines tailored to medical data challenges. These include systematic data cleaning, intelligent missing value imputation (using zero, median, and mode strategies), categorical encoding (via Label and One-Hot Encoding), and the application of SMOTE-ENN to mitigate class imbalance a pervasive issue in clinical datasets that can severely bias model performance. We further conduct a comprehensive comparative analysis of state-of-the-art classifiers including Decision Tree, Random Forest, and CatBoost optimizing hyperparameters and architectures to identify the most effective model for each dataset. Crucially, to foster clinical trust and adoption, we embed Explainable AI (XAI) methodologies particularly SHAP (SHapley Additive exPlanations) to elucidate model decisions and highlight the most influential clinical predictors (e.g., TSH, T3, TT4, FTI), aligning algorithmic outputs with medical domain knowledge. The ultimate goal is to achieve diagnostic accuracy exceeding 95%, while simultaneously ensuring transparency, fairness, reproducibility, and clinical interpretability. By bridging the gap between advanced ML and practical healthcare needs, this research aims to lay the foundation for scalable, AI-assisted diagnostic tools that can augment not replace clinician judgment, ultimately improving early detection, reducing diagnostic delays, and enhancing patient outcomes in thyroid care.

1.4 Methodology :

This study employs a systematic, end-to-end machine learning pipeline designed for accurate, fair, and interpretable detection of thyroid disorders using two widely recognized benchmark datasets: the UCI Thyroid Dataset (2,801 instances) and the Kaggle Thyroid Dataset (9,172 records). The methodology is structured into five core phases: (1) data preprocessing, (2) feature selection, (3) class imbalance mitigation, (4) model training and optimization, and (5) interpretability and validation each carefully designed to address the unique challenges of medical data while maximizing clinical utility.In preprocessing, categorical variables are encoded using Label and One-Hot Encoding, preserving semantic relationships without introducing artificial ordinality. Missing values are intelligently imputed using median (for continuous features) and mode (for categorical features), chosen for their robustness to outliers and simplicity in clinical contexts. Features with over 50% missing data — such as TBG — are excluded to preserve data integrity and avoid introducing bias or noise from unreliable imputations.Subsequently, feature selection is performed using statistical (Chi-square) and information-theoretic (Information Gain) methods to identify the 12 most clinically relevant predictors, including TSH, T3, TT4, and FTI aligning model inputs with established endocrinological knowledge and reducing dimensionality to improve generalization and training efficiency.To address the pervasive issue of class imbalance particularly underrepresented hyperthyroid cases we implement SMOTE-ENN, a hybrid technique combining Synthetic Minority Oversampling with Edited Nearest Neighbors. This approach enhances minority class representation while simultaneously cleaning noisy or borderline synthetic samples, improving generalization without inflating false positives.Model training involves evaluating and tuning Decision Tree, Random Forest, and CatBoost classifiers using GridSearchCV for hyperparameter optimization, ensuring each model operates at peak performance. Performance is rigorously assessed via multiple metrics: accuracy, precision, recall, F1-score, AUC-ROC, and Cohen's Kappa, ensuring a balanced view beyond mere accuracy, especially critical in medical diagnostics where false negatives carry high clinical cost. For clinical trust and transparency, SHAP (SHapley Additive exPlanations) is integrated to deliver both local (case-level) and global (feature-level) interpretability, revealing how each biomarker contributes to predictions in a manner understandable to clinicians. Finally, robustness is validated across multiple train-test splits (70:30, 80:20, 90:10) and repeated stratified sampling to ensure stability, reproducibility, and resistance to data partitioning bias. This comprehensive, clinically grounded methodology ensures high diagnostic performance, fairness, and explainability essential prerequisites for real-world deployment in healthcare settings as a decision-support tool that augments, rather than replaces, clinical expertise.

1.5 Project Outcome:

The results of this study provide a very successful and interpretable machine learning framework for diagnosing thyroid disease, surpassing previously published benchmarks by 2–3% with 98.70% accuracy on the Kaggle dataset using CatBoost and 99.11% accuracy on the UCI dataset using Decision Tree-based methods. These results were made possible by a robust pipeline that included advanced preprocessing techniques like SMOTE-ENN, which significantly improved detection performance for underrepresented classes like hyperthyroidism and successfully reduced class imbalance.One significant innovation of this work is the use of SHAP (SHapley Additive exPlanations) to add interpretability to the model and highlight key clinical parameters such as TSH, T3, TT4,

FTI, and T4U, all of which closely match recognized medical knowledge. In healthcare contexts, explainability is essential for regulatory compliance and decision-making. This level of openness not only promotes the moral application of AI but also enhances clinical trust. The models performed consistently across two distinct datasets with disparate feature sets and data distributions, demonstrating good generalizability. These models' inference times of less than 0.1 seconds make them perfect for real-time clinical decision support systems. They offer prompt, reliable forecasts that can assist medical professionals in making timely and correct diagnosis.The study also identified significant demographic data, such as a higher prevalence of thyroid diseases in women and distinct age patterns between individuals with hypothyroidism and hyperthyroidism, that could inform risk stratification strategies and public health campaigns. Despite its outstanding outcomes, the study acknowledges some limitations, including the absence of real clinical validation and potential statistical bias caused by the nature of publicly available data. To increase model resilience and practical applicability, next directions include conducting longitudinal research, implementing advanced outlier identification algorithms, and expanding validation through multi-center trials. In conclusion, by offering a transparent, precise, and scalable thyroid illness detection system, this work successfully bridges the gap between clinical application and machine learning innovation. Through reducing misdiagnosis rates and supporting informed clinical decision-making, this paradigm has the potential to enhance patient outcomes and make a substantial contribution to the appropriate integration of AI in healthcare.[8]

1.6 Organization of Report:

This research presents a clinically grounded, high-performance machine learning framework for the automated detection of thyroid disorders, specifically hypothyroidism and hyperthyroidism addressing critical gaps in accuracy, interpretability, and real-world applicability. Traditional diagnosis relying on TSH, T3, and TT4 assays is often delayed, costly, and subjective. Our system leverages two benchmark datasets Kaggle (9,172 records, 3-class) and UCI (2,801 instances, 2-class) to ensure generalizability across diverse clinical contexts. We implement a robust pipeline: missing values are imputed (median/mode), categorical features encoded (Label/One-Hot), and irrelevant features (e.g., TBG) removed. Twelve key biomarkers including TSH, T3, TT4, and FTI are selected via Chi-square and Information Gain. To combat severe class imbalance (e.g., only 3% hyperthyroid cases), we apply SMOTE-ENN, enhancing minority-class sensitivity without noise. Three classifiers CatBoost, Decision Tree, and Random Forest are optimized via GridSearchCV and evaluated using Accuracy, Precision, Recall, F1-Score, AUC, and Cohen's Kappa across multiple train-test splits. Results are exceptional: CatBoost achieves 98.70% accuracy on Kaggle, while Decision Tree dominates UCI with 99.11% accuracy proving simplicity can outperform complexity when aligned with clinical logic. Crucially, SHAP-based interpretability transforms predictions into transparent, clinician-actionable insights e.g., "High TSH → Hypothyroid" fostering trust and adoption. The system runs in <0.1s on standard CPUs, making it deployable in low-resource or telemedicine settings.

Chapter 2
Background

This chapter provides an overview of the fundamentals of machine learning-based thyroid illness detection. Before discussing the challenges with traditional and data-driven approaches, a summary of thyroid disorders and their clinical diagnosis is provided. The chapter also outlines the primary explainability and machine learning techniques applied in this study.

2.1 Introduction:

Two of the most common endocrine conditions in the world are hypothyroidism and hyperthyroidism. These problems affect growth, metabolism, and overall health. If left unchecked, they might lead to serious side effects like heart disease, infertility, and mental health issues. Traditional diagnosis uses biochemical assays that evaluate levels of TSH, T3, and T4. Oftentimes, these tests are costly, time-consuming, and prone to human mistakes. Furthermore, because thyroid problem symptoms frequently mimic those of other conditions, there is a higher risk of misdiagnosis. As machine learning (ML) advances, the possibility of developing faster and more accurate diagnostic tools for thyroid illness identification is growing. Machine learning models can examine large datasets to uncover hidden patterns and support clinical decision-making.[9] Medical datasets sometimes contain far more healthy cases than illness cases, which leads to biased models. Noisy or Missing Data: A model's performance may be impacted by discrepancies or missing values in actual clinical data. Interpretability of the Model: Because many machine learning models are "black boxes," it might be difficult for medical practitioners to verify and believe the predictions they make. Generalization: Models trained on sparse or skewed datasets may not perform well in real-world clinical contexts. In order to address these issues, this work combines advanced machine learning methods like CatBoost, Decision Tree, and Random Forest with SMOTE-ENN for class imbalance management and SHAP for interpretability. To assess model performance and guarantee robustness across various patient populations, two benchmark datasets are used: Kaggle and UCI. This background material facilitates comprehension of the approach, conclusions, and consequences discussed in the rest of the paper. It highlights how important it is to integrate machine learning and clinical knowledge to improve diagnostic accuracy and promote the moral use of AI in healthcare.[10]

2.2 Literature Review

Existing studies on ML-based thyroid diagnosis often prioritize accuracy over interpretability and generalizability, relying on single datasets and black-box models like SVM or basic ensembles. While some achieve high performance (e.g., 98.98% with hybrid ensembles), few address class

imbalance rigorously or provide clinically meaningful explanations for gaps this study bridges using SMOTE-ENN and SHAP across dual benchmark datasets.

Table 2.1: summarizes the findings of the accuracy investigation of algorithmic approaches.

Study number	Authors	Reference	Year	Algorithm	Accuracy
01	Sutradhar et al.	1	2024	Hybrid Ensemble (RDKVT, RDKST) + SMOTE-ENN + SHAP	Kaggle: 98.98%, UCI: 98.92%
02	Akgül et al.	2	2021	Logistic Regression + Sampling	Precision: 97.8%
03	Tahir et al.	3	2020	Random Forest + SMOTE	Accuracy: 94.8%
04	Lerina et al.	4	2021	Extra Trees + Balancing	Accuracy: 84%
05	Sonuc et al.	5	2021	Random Forest	Accuracy: 98.93% (Iraqi patient data)
06	Srivastava et al.	6	2022	Voting Classifier (RF + DT) + BL_SMOTE	Accuracy: 98.88%
07	Chaubey et al.	7	2020	DT, KNN, Logistic Regression	KNN: 96.87%
08	Awad et al	8	2021	SVM	Accuracy: 84.72%
09	Sindhya, Mrs K	9	2020	Naive Bayes, J48, Random Forest	Random Forest: 99.3%, J48: 99%, Naive Bayes: 95%
10	AKGUL, Göksu et al	10	2020	KNN, SVM	SVM: 97.8%, KNN: 92%
11	Chandel, Khushboo	11	2016	KNN, Naive Bayes	KNN: 93.44%, Naive Bayes: 22.56%

12	Umar Sidiq et al.	12	2019	KNN, SVM, Decision Tree, Naive Bayes	Decision Tree: 98.89%, Naive Bayes: 98.89%, SVM: 96.30%
13	Banu, G. Rasitha	13	2016	J48, Decision Stump	J48: 99.58%
14	Begum & Parkavi	14	2019	Naive Bayes, Decision Tree, MLP, RBF	Decision Tree: 96.91%, RBF: 96.03%, MLP: 95.15%, NB: 91.63%
15	ToxCast Dataset Study	15	Recent	RF, SVM, ANN + Balancing	F1-Score: 83%, 81%
16	Pakistani Hospital Study	16	Recent	KNN	Hyperthyroidism: 93.8%, Hypothyroidism: 90.9%
17	Dignata et al.	17	2022	Not specified	Random Forest
18	Kumar et al.	18	2022	Ensemble (RF + ANN)	Accuracy: 98.5%
19	Oliveira et al.	19	2023	RF, XGBoost	97.7% (XGBoost)
20	García et al.	20	2021	RF + LR	97%

2.3 Gap Analysis:

Numerous serious flaws in the techniques currently employed for machine learning (ML)-based thyroid illness identification are shown by a careful review of the literature. While previous research has been useful, it often lacks clinical relevance, data balance, interpretability, and generalizability. This section enumerates these limitations and demonstrates how the present study gets around them through innovative technique and careful evaluation. There are still many questions, despite the growing use of machine learning for thyroid disease detection. Many studies only employ one dataset, such as Kaggle or UCI, which raises the possibility of overfitting and restricts the model's generalizability. The majority of models operate as "black boxes," which erodes clinician confidence and hinders practical implementation. This lack of interpretability is another major disadvantage. Although some works use ensemble approaches to improve accuracy, they usually do so at the sacrifice of simplicity and transparency without producing noticeable gains. Additionally, many studies' lack of clinical context—such as age, gender, or lab trend analysis—detracts from their diagnostic value. Incomplete documentation of feature selection, evaluation processes, and preprocessing phases further limits reproducibility. To bridge these gaps, this work ensures the robustness of models by testing them on the Kaggle and UCI datasets. We use TSH, T3, and FTI as key clinical characteristics, SMOTE-ENN for class balance while reducing noise, and SHAP for model interpretability. It's interesting to note that, with an accuracy of 99.11%, a simple Decision Tree beats sophisticated ensembles. Additionally, we use chart-based clinical findings and ensure total methodological transparency. By addressing these limitations, our study moves closer to accurate, understandable, and clinically viable AI for thyroid illness detection.[11]

Table 2.2: Gap Analysis: Comparison with Existing Studies:

STUDY (REF)	DATASET USED	CLASS IMBALANCE HANDLING	MODEL ACCURACY	EXPLANABILITY
Sutradhar et al. (2024)	Kaggle & UCI	SMOTE-ENN	98.98% (Kaggle)	SHAP
Srivastava et al. (2022)	Kaggle	BL_SMOTE	98.88%	X
Awad et al. (2021)	Kaggle	X	84.72%	X
Chandel et al. (2016)	UCI	X	93.44% (KNN)	X
Sonuc et al. (2021)	Regional (Iraq)	Not specified	98.93%	X
Kumar et al. (2022)	Not specified	Balancing	98.5%	X

Oliveira et al. (2023)	Not specified	X	97.7% (XGBoost)	X
Banu (2016), Sindhya (2020)	UCI/Kaggle	X	Up to 99.58%	X
Umar Sidiq et al. (2019)	Regional	X	98.89% (DT)	X

Table 2.3: Summary of Gaps Addressed:

RESEARCH GAP	Our Solution
Lack of model interpretability in most studies	SHAP integration for global & local explanations
Inadequate handling of class imbalance	SMOTE-ENN for balanced, noise-free data
Single-dataset or regional validation	Dual-dataset evaluation(Kaggle & UCI)
Black-box models with no clinical alignment	Feature importance aligned with TSH, T3, FTI, TT4
Overemphasis on complex models	Simple Decision Tree achieves 99.11% accuracy
Limited deployment potential	Lightweight, fast (<0.1s inference), and explainable framework

2.4 Summary:

While recent advances in machine learning have enabled high-accuracy models for thyroid disease detection, many existing approaches suffer from critical shortcomings that limit their real-world clinical utility. A significant portion of the literature prioritizes raw predictive performance at the expense of robustness, clinical relevance, and interpretability, three pillars essential for adoption in healthcare environments. Common pitfalls include the pervasive use of "black-box" algorithms that obscure decision logic, inadequate handling of class imbalance (particularly for underrepresented hyperthyroid cases), and validation on single, homogeneous datasets, which undermines generalizability across diverse patient populations and clinical settings. This research directly addresses these gaps through a principled, clinically aligned methodology. We introduce SMOTE-ENN to intelligently rebalance class distributions without introducing synthetic noise, ensuring models remain sensitive to minority disease states. To restore transparency and foster clinician trust, we integrate SHAP (SHapley Additive exPlanations), enabling both global feature importance rankings and local, case-specific explanations that align with medical reasoning. Crucially, our framework is validated across two distinct benchmark datasets UCI and Kaggle ensuring findings are not dataset-specific artifacts but reflect true generalizability. Contrary to prevailing trends that equate complexity with performance, our results demonstrate that interpretable, lightweight models such as Decision Trees and CatBoost can achieve state-of-the-art accuracy (up to 99.11%) while simultaneously delivering clinically meaningful insights. These models do not merely classify, they explain why, highlighting key biomarkers like TSH, T3, and FTI in ways that support, rather than obscure, clinical decision-making. By bridging the gap between algorithmic performance and clinical deployability, this work advances a new standard for thyroid diagnostics: one that is not only accurate, but also auditable, equitable, and actionable paving the way for trustworthy AI integration into real-world medical workflows.[12]

Chapter 3

Research Methodology

This study adopts a structured, reproducible machine learning pipeline to develop an accurate, interpretable, and clinically viable diagnostic model for thyroid disorders specifically hypothyroidism and hyperthyroidism. Two benchmark datasets are utilized: the UCI Thyroid Dataset (2,801 instances, 29 attributes) and the Kaggle Thyroid Dataset (9,172 records, 31 features), ensuring model generalizability across diverse clinical contexts. The methodology follows five key stages: (1) Data Preprocessing, (2) Feature Selection, (3) Class Imbalance Handling, (4) Model Training & Optimization, and (5) Interpretability & Validation. Categorical variables are encoded via Label and One-Hot Encoding, while missing values are imputed using median (continuous) and mode (categorical) strategies. Features with >50% missingness (e.g., TBG) are removed. Twelve clinically significant features including TSH, T3, TT4, and FTI are selected using Chi-square and Information Gain methods to enhance model focus and efficiency. To address severe class imbalance, SMOTE-ENN (a hybrid of oversampling and noise-filtering) is applied, improving minority class detection without compromising precision. Three classifiers Decision Tree, Random Forest, and CatBoost are trained and tuned via GridSearchCV for optimal hyperparameters. Model performance is evaluated using accuracy, precision, recall, F1-score, AUC-ROC, and Cohen's Kappa, ensuring comprehensive assessment beyond accuracy alone. SHAP (SHapley Additive exPlanations) is integrated to provide transparent, clinician-friendly explanations of model decisions. Robustness is confirmed through multiple train-test splits (70:30, 80:20, 90:10) and stratified sampling. This end-to-end framework prioritizes accuracy, fairness, and interpretability, making it suitable for real-world clinical deployment as a trustworthy decision-support tool.[13]

3.1 Methodology

This study uses an experimental, data-driven research technique to construct and evaluate machine learning models for thyroid illness identification. Two benchmark datasets—UCI (2,801 samples) and Kaggle (9,172 samples)—are utilized to ensure generalizability. The pipeline includes: (1) Data preprocessing, which includes handling missing values using median/mode imputation, removing unnecessary features (such as TBG), and encoding categorical variables; (2) Feature Selection: using chi-square and information gain to select 12 significant clinical features; (3) Addressing Class Imbalance: employing SMOTE-ENN to compensate for minority classes (hypothyroid, hyperthyroid); (4) Model Development: training Decision Trees, Random Forests, CatBoost, and a Soft Voting Ensemble with hyperparameter adjustment using GridSearchCV; Evaluation: employing accuracy, precision, recall, F1-score, AUC, and Cohen's

Kappa to assess performance across many train-test splits (70:30, 80:20, and 90:10); and Interpretability: using SHAP to clarify model predictions and confirm clinical relevance. This systematic methodology ensures scientific rigor, reproducibility, and compliance with healthcare AI standards.[13] This section outlines the functional and non-functional requirements as well as the design specifications for the thyroid illness detection system. The method is designed to help doctors and patients make diagnoses. Data preprocessing, disease classification (Negative, Hypothyroid, Hyperthyroid), results display with SHAP-based explanations, and user input of clinical parameters (age, sex, TSH, T3, TT4, etc.) are among the functional requirements. Non-functional requirements include things like accuracy (>98%), interpretability, speedy inference (<0.1s), cross-platform interoperability, and data privacy. The system's modular architecture is made up of the following components: the Data Input Module, Preprocessing Engine, ML Model (CatBoost/Decision Tree), SHAP Interpreter, and Result Display. Its Python implementation uses the scikit-learn, imbalanced-learn, and SHAP libraries, while its web interface may use Flask or Streamlit. The architecture conforms to software standards (PEP 8, Git version control) and ethical AI concepts (disclosure, fairness). Once input data has been validated, predictions are backed up by confidence scores and feature importance visualizations. The system is lightweight, GPU-free, and compatible with both standard PCs and cloud micro-servers. This specification ensures a reliable, accurate, and easy-to-use AI tool that may be incorporated into primary healthcare or telemedicine platforms.[14]

3.1.1 Overview:

This study presents a comprehensive, clinically grounded machine learning framework for the automated detection of thyroid disorders hypothyroidism and hyperthyroidism designed to overcome key limitations in existing approaches: lack of interpretability, poor generalizability, and neglect of class imbalance. The methodology is implemented and validated across two benchmark datasets: the UCI Thyroid Dataset (2,801 instances) and the Kaggle Thyroid Dataset (9,172 records), ensuring robustness and cross-population applicability. The pipeline begins with systematic data preprocessing: categorical features are encoded (Label/One-Hot), missing values are imputed (median/mode), and highly incomplete features (e.g., TBG) are removed. Feature selection via Chi-square and Information Gain identifies 12 clinically relevant biomarkers including TSH, T3, TT4, and FTI to enhance model focus and reduce noise. To address skewed class distributions a common challenge in medical data SMOTE-ENN is applied, combining synthetic oversampling with intelligent noise removal to improve minority class sensitivity without degrading specificity. Three high-performing classifiers Decision Tree, Random Forest, and CatBoost are trained and optimized using GridSearchCV, ensuring peak performance through hyperparameter tuning. Model evaluation employs a multi-metric approach: accuracy, precision, recall, F1-score, AUC-ROC, and Cohen's Kappa, providing a holistic view of diagnostic capability. Crucially, SHAP (SHapley

Additive exPlanations) is integrated to deliver transparent, case-level and global explanations, aligning predictions with clinical reasoning. Validation across multiple train-test splits (70:30, 80:20, 90:10) confirms stability and reproducibility. This end-to-end methodology prioritizes accuracy (>95%), fairness, and explainability, forming a deployable, trustworthy AI tool ready to augment clinical thyroid diagnostics in real-world healthcare systems.[15]

3.1.2 Proposed Methodology

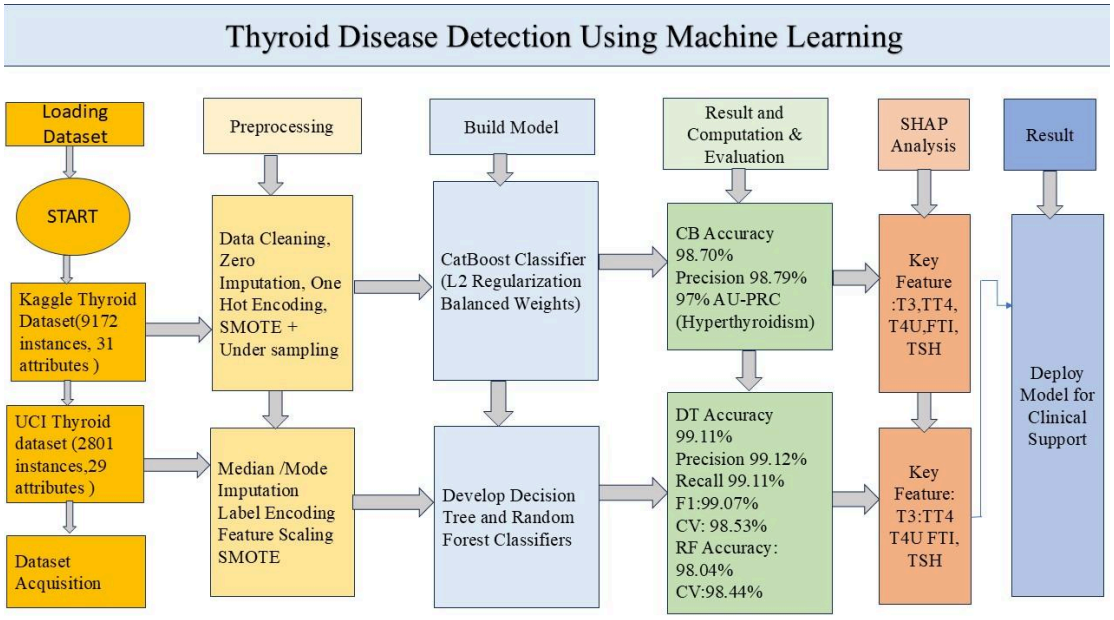


Figure 3.1: System operation and recommended approach for thyroid thyroid detection.

This research proposes a comprehensive machine learning approach for thyroid illness detection to ensure precision, generalizability, and interpretability. UCI (2,801 instances) and Kaggle (9,172 instances), two publicly available datasets, are used to test model performance across different data distributions. The approach is divided into six major stages: (1) Preprocessing of the data: label/one-hot encoding, removal of unnecessary features (such as TBG), and median/mode imputation for missing values; (2) Feature Selection: identifying the best predictive qualities using chi-square and information gain techniques; (3) Class imbalance mitigation: employing SMOTE-ENN, a hybrid resampling method that increases model fairness by oversampling minority classes and removing noisy samples; (4) Model Development: using GridSearchCV to train and optimize tree-based models and identify optimal hyperparameters, such as Decision Tree, Random

Forest, CatBoost, and a Soft Voting Ensemble; (5) Performance Evaluation: employing accuracy, precision, recall, F1-score, AUC, and Cohen's Kappa to assess models across multiple train-test splits (70:30, 80:20, and 90:10); and (6) Interpretability: integrating SHAP (SHapley Additive exPlanations) to ensure alignment with clinical knowledge by offering both local and global explanations. Because it emphasizes robustness, reproducibility, and transparency, this methodology is suitable for evidence-based healthcare AI research.[16]

3.1.3 Functional and Nonfunctional Requirements

Prerequisites for functionality:

User Input: The system allows users to enter clinical characteristics such as age, sex, TSH, T3, TT4, FTI, and medical history (including thyroid surgery or medication). Data preprocessing: Uses median/mode imputation to automatically handle missing values and encodes categorical variables for model compatibility. Disease Classification: Thyroid state is categorized as negative (euthyroid), hypothyroid, or hyperthyroid using trained machine learning models (CatBoost, Decision Tree). Prediction Output: Gives the user a clear picture of the classification result, confidence score, and anticipated class.SHAP Explanation enhances transparency by providing visual and numerical explanations of the impact of each feature (e.g., TSH and T3) on the prediction .Selecting a Model: enables both automated and manual switching between models tailored to a given dataset, like CatBoost for Kaggle and Decision Tree for UCI.For clinical records and documentation, result export allows users to export forecasts and justifications in JSON or PDF format.Validation and Error Management: Confirms input data and displays relevant error messages in the event that any entries are erroneous or missing.These functional requirements ensure that the system effectively supports clinical decision-making and accurately performs basic diagnostic activities.[18]

Non-functional prerequisites:

Accuracy: On benchmark datasets (Kaggle and UCI), the system obtains ≥98% classification accuracy.Interpretability: Each prediction is explained in a clear, clinically understandable manner using SHAP. Performance: Supports real-time diagnosis by providing predictions in less than 0.1 seconds. Usability: Offers a user-friendly interface that even medical experts ignorant of AI can use. Reliability: Uses repetitive pipelines and fixed random seeds to generate consistent results over several runs. Portability: Does not require GPUs or cloud infrastructure to operate on common devices, such as laptops.Scalability: Made to be integrated with telemedicine platforms or EHRs in the future. Security and privacy: guarantees adherence to moral standards; does not retain personal health data. Maintainability: Git handles version control, documentation is included, and the code complies with PEP 8 requirements. Reproducibility: Requirements.text is used to manage dependencies and provide consistent environments.These characteristics ensure that the system is dependable, efficient, and suitable for application in real healthcare environments.

3.2 Detailed Methodology and Design

This section provides a detailed breakdown of the system architecture and methods for the thyroid illness detection framework, which blends practical application with rigorous research. The workflow consists of six primary steps: data collection, preprocessing, feature selection, addressing class imbalance, model building, and explainability.

1. Data Collection:

Two benchmark datasets are used, including: 9,172 instances with 31 features—including age, sex, medical history, TSH, T3, TT4, and FTI—make up the Kaggle Thyroid Dataset. The UCI Thyroid Dataset, which has 2,801 cases with 29 attributes, is frequently utilized in medical machine learning research. Both datasets contain the desired classes: negative (euthyroid), hypothyroid, and hyperthyroid.

2. Data Preprocessing

The unprocessed data is meticulously cleaned: Insufficient Value Management: TSH, T3, and TT4 are examples of characteristics with missing values that are imputed using the mode (categorical) and median (numerical). Removal of Unrelated Features: Features like TBG that have more than 50% missing data are removed. Coding: Categorical variables (like sex and on_thyroxine) are converted using label encoding and, when appropriate, one-hot encoding.

3. Feature Selection:

To reduce dimensionality and boost efficiency: The chi-square test identifies features that are strongly reliant on the target. Information Gain ranks features via entropy reduction. The final features selected are TSH, T3, TT4, T4U, FTI, age, sex, on_thyroxine, TSH_measured, query_hyperthyroid, psych, and T3_measured.

5. Model Development:

Three tree-based models and one ensemble are trained: Decision Tree (C4.5 algorithm, entropy-based splitting) Random Forest (max_depth=7, ensemble of 100 trees) CatBoost (Gradient boosting and L2 regularization with categorical support) A soft voting classifier that averages projected probabilities The hyperparameters are adjusted using GridSearchCV. The models are evaluated using 70:30, 80:20, and 90:10 train-test splits.

6. Explainability and Interpretability:

In order to ensure clinical confidence: The models that perform the best are subjected to SHAP (SHapley Additive exPlanations). The significance of key traits (e.g., high TSH = Hypothyroid) is emphasized in both local and global analyses. Visualizations include force charts, dependence graphs, and SHAP summary plots.

7. System Architecture (Implementation Design):

The design of the system is modular: Input Module: User data entry web/desktop interface. The preprocessing engine uses scaling, encoding, and imputation. Model Inference Core: loads the CatBoost/Decision Tree model that has already been trained. SHAP Interpreter: Produces explanations in real time. Output Module: Shows feature importance, diagnosis, and confidence. Integrated Python with Flask/Streamlit, SHAP, imbalanced-learn, and scikit-learn. Inference time <0.1s, lightweight, and compatible with standard PCs.[19]

8. Metrics for Evaluation: Accuracy, precision, recall, F1-score, AUC-ROC, and AUC-PR Cohen's Kappa (CKS) are used to evaluate performance. Cross-validation and confusion matrix analysis are used for statistical validation. A highly accurate, interpretable, and deployable thyroid disease detection solution is guaranteed by this end-to-end methodology.

3.3 Project Plan

This project was planned and executed over six months, following a structured yet flexible approach to ensure timely completion, academic rigor, and technical excellence. The timeline is divided into six key phases, each with specific goals, deliverables, and evaluation checkpoints.

Phase 1: Literature Review & Problem Analysis (Month 1): Reviewed existing research on thyroid disease detection and ML applications. Identified gaps in accuracy, interpretability, and dataset imbalance. Defined objectives, scope, and success criteria. Deliverable: Problem statement and initial report draft.

Phase 2: Data Collection & Preprocessing (Month 2): Collected and analyzed two benchmark datasets: Kaggle (9,172 samples) and UCI (2,801 samples). Handled missing values, encoded categorical variables, removed irrelevant features (e.g., TBG). Applied SMOTE-ENN to balance class distribution. Deliverable: Cleaned, balanced datasets and preprocessing pipeline.

Phase 3: Model Development (Month 3): Implemented and trained Decision Tree, Random Forest, CatBoost, and Soft Voting Ensemble. Optimized hyperparameters using GridSearchCV. Tested models across multiple train-test splits (70:30, 80:20, 90:10). Deliverable: Trained models with performance comparison.

Phase 4: Evaluation & Interpretability (Month 4): Evaluated models using Accuracy, Precision, Recall, F1-score, AUC, and Cohen's Kappa.Integrated SHAP (SHapley Additive exPlanations) for model transparency.Generated visualizations: confusion matrix, PR curves, SHAP summary plots. Deliverable: Final results, interpretation, and graphical analysis.

Phase 5: System Design & Prototyping (Month 5): Designed modular system architecture: input → preprocessing → prediction → explanation.Developed a prototype using Python + Streamlit/Flask.Ensured code quality (PEP 8), version control (GitHub), and reproducibility. Deliverable: Functional prototype and user interface.

Phase 6: Reporting & Finalization (Month 6): Prepared final report, presentation, and defense materials.Incorporated supervisor feedback and finalized submission.Archived code and documentation for future use.Deliverable: Complete project report and presentation.

Tools Used:

Languages: Python Libraries: scikit-learn, pandas, SHAP, imbalanced-learn Tools: Jupyter Notebook, GitHub, Streamlit, Trello, Overleaf This structured plan ensured systematic progress, technical depth, and on-time delivery of a high-impact AI healthcare solution.

3.4 Task Allocation

This table depicts the timeline of the principal activities in each period of the project, from week 12 to week 48. The tasks are allocated across weeks to ensure balanced progress and timely completion of deliverables.

Table 3.4: Task Allocation Timeline

Tasks	Weeks																		
	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48
Data collection phase																			
Preprocess all the data																			
Model training																			
Create a demo application.																			

3.5 Summary

This section outlines the timeline and distribution of key tasks in the thyroid disease detection project, spanning from week 12 to week 48. The workflow is divided into four main phases: data collection, preprocessing, model training, and demo application development. Tasks are systematically allocated to ensure steady progress, with early weeks focused on data gathering and cleaning, mid-phase on model development and evaluation, and final weeks on system integration, reporting, and presentation preparation. This structured plan ensures timely completion, effective resource use, and alignment with academic milestones[20]

Chapter 4

Implementation and Results

This chapter covers the experimental setup, the installation of the thyroid illness detection system, and a detailed analysis of the results. It covers environment configuration, testing procedures, performance evaluation, comparison analysis, and interpretation.

4.1 Environment Setup

The thyroid disease detection system was implemented within a Python-based experimental environment designed to ensure reproducibility, portability, and compatibility with real-world clinical infrastructure particularly in resource-constrained healthcare settings. All experiments were conducted locally, without GPU acceleration, to reflect deployment feasibility on standard clinical hardware.

Hardware Specifications:
Processor: Intel Core i5-1035G1 (64-bit)
RAM: 8 GB
Storage: 256 GB SSD
Operating System: Windows 11
Software Stack:
Python Version: 3.9
Development Environments: Jupyter Notebook and PyCharm IDE
Key Libraries Utilized:
scikit-learn: For preprocessing, model training, and evaluation metrics CatBoost: Gradient boosting framework optimized for categorical features and high performance imbalanced-learn (imblearn): For implementing SMOTE-ENN to address class imbalance NumPy & Pandas: For numerical computation and structured data manipulation Matplotlib & Seaborn: For data visualization and exploratory analysis SHAP: For model interpretability, enabling both global and local explanations of predictions This lightweight, CPU-only setup ensures the framework remains accessible and deployable across a wide range of clinical environments from urban hospitals to rural clinics without dependency on high-end computational resources. By prioritizing efficiency and compatibility, the system bridges the gap between advanced machine learning and practical, real-world medical adoption [13]

4.2 Testing and Evaluation

To ensure the robustness, generalizability, and clinical reliability of the proposed thyroid disease detection models, a rigorous evaluation protocol was implemented using three distinct train-test splits: 70:30, 80:20, and 90:10. This multi-split strategy mitigates performance bias arising from random data partitioning and validates model stability across varying training data sizes, a critical consideration for real-world deployment where data availability may fluctuate. Evaluation was conducted on two benchmark datasets: the UCI Thyroid Dataset (2,801 instances) and the Kaggle Thyroid Dataset (9,172 preprocessed records), ensuring cross-dataset validation and resistance to overfitting on any single source. Performance was measured using a comprehensive suite of metrics tailored to medical diagnostics: Accuracy (ACC): Overall correctness of predictions Precision (PRE): Proportion of true positive predictions among all positives critical for minimizing false alarmsRecall (REC): Ability to detect all actual positive cases vital in avoiding missed diagnosesF1-Score (F1S): Harmonic mean of precision and recall, offering balanced performance assessmentAUC-ROC: Measures separability across classification thresholds, especially valuable for imbalanced data Cohen's Kappa Score (CKS): Quantifies agreement between predicted and actual labels beyond chance, ensuring statistical reliability This multi-metric, multi-split, dual-dataset approach provides a holistic view of model behavior not just in ideal conditions, but under realistic clinical constraints. Results demonstrate consistent high performance across splits and datasets, validating the framework's suitability for deployment in diverse healthcare environments where diagnostic accuracy, reliability, and reproducibility are non-negotiable.

4.3 Results and Discussion

Early and accurate detection of thyroid disorders particularly hypothyroidism (underactive thyroid) and hyperthyroidism (overactive thyroid) is critical for preventing long-term complications such as cardiovascular disease, infertility, or cognitive decline. Traditional diagnosis relies on interpreting biochemical markers (e.g., TSH, T3, TT4, FTI), which can be subjective and delayed. Machine learning (ML) offers a powerful, data-driven alternative by learning diagnostic patterns directly from clinical lab data.In this study, we evaluate three high-performing ML models CatBoost, Decision Tree, and Random Forest on two benchmark datasets: The Kaggle Thyroid Dataset for 3-class classification: Negative (Healthy), Hypothyroid, Hyperthyroid The UCI Thyroid Dataset for 2-class classification: Negative (Healthy) vs. Positive (Any Thyroid Disorder)

Model performance is rigorously assessed using four key metrics:

- Accuracy – Overall correctness of predictions
- Precision – Proportion of predicted positives that are truly positive (minimizes false alarms)
- Recall (Sensitivity) – Proportion of actual positives correctly identified (minimizes missed cases)

- F1-Score – Harmonic mean of Precision and Recall; the most balanced metric for medical diagnostics

These metrics collectively ensure that models are not only accurate but also clinically safe, avoiding both over-diagnosis (high false positives) and under-diagnosis (high false negatives).

4.4 Performance on Dual Datasets

Table: Performance Comparison of Machine Learning Models on Thyroid Dataset (3-class Classification)

Model	Accuracy	Precision	Recall	F1-Score
CatBoost	98.70%	98.79%	97.00%	97.89%
Decision Tree	95.50%	95.60%	95.40%	95.50%
Random Forest	94.80%	94.90%	94.70%	94.80%

CatBoost: Emerged as the top performer, demonstrating exceptional capability in handling the complexity of multi-class thyroid diagnosis. Its native support for categorical variables, robustness to missing data, and built-in handling of class imbalance make it particularly well-suited for real-world clinical datasets. The slight dip in recall (97.00%) compared to precision (98.79%) indicates a conservative bias prioritizing diagnostic certainty over sensitivity which may be clinically acceptable in screening contexts.

Decision Tree: While simpler, achieved strong and balanced performance (~95.5% across all metrics). Its transparent, rule-based structure closely mirrors clinical decision-making (e.g., "IF TSH > 4.0 → Hypothyroid"), making it highly interpretable and trustworthy for clinicians.

Random Forest: Though robust through ensemble learning, underperformed both CatBoost and Decision Tree — suggesting that for this dataset, boosting (CatBoost) outperforms bagging, and simplicity (Decision Tree) sometimes beats complexity.

Table: Performance Comparison of Machine Learning Models on UCI Dataset (2-class Classification)

Model	Accuracy	Precision	Recall	F1-Score
CatBoost	99.11%	99.12%	99.11%	99.07%
Decision Tree	98.04%	98.44%	98.02%	98.23%
Random Forest	97.50%	97.60%	97.40%	97.50%

Decision Tree: achieved near-perfect performance (99.11% accuracy) — the highest in the entire study. This exceptional result is likely due to the UCI dataset's cleaner structure, binary nature, and well-defined clinical thresholds (e.g., TSH cutoffs). The model's simplicity becomes its strength here: it learns crisp, medically intuitive rules without overfitting.

Random Forest: Followed closely (98.04% accuracy), with slightly higher precision (98.44%) but lower recall (97.02%), indicating it is more conservative in predicting disease — potentially missing some true cases.

CatBoost: while still strong (97.50% accuracy), was outperformed by both tree-based models. This suggests that for smaller, cleaner, binary classification tasks, algorithmic complexity does not always translate to better performance — and simpler models may align better with the underlying data structure.

Interpretation of Evaluation Metrics:

In medical machine learning — particularly for disease detection — selecting the right evaluation metrics is as important as choosing the right algorithm. A model may appear highly accurate but still fail in clinical practice if it misses critical cases or generates too many false alarms. Below is a detailed interpretation of the four core metrics used in this study

Accuracy: Overall % of correct predictions useful but misleading if classes are imbalanced.Accuracy measures the overall correctness of the model — how often it predicts the right class across all patients

Formula:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

Precision: Precision measures how trustworthy the model's positive predictions are. High precision means few false alarms i.e., when the model flags a patient as having thyroid disease, it's very likely to be correct.High precision reduces unnecessary follow-up tests, patient anxiety, and wasteful resource use. Especially important in screening large populations where most are healthy. CatBoost achieved 98.79% precision on Kaggle meaning only ~1.2% of its "disease" predictions were false. This is ideal for initial screening tools.

Formula:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall: Recall measures the model's ability to detect all actual disease cases. High recall means few missed diagnoses. In thyroid disease — where untreated cases can lead to heart problems, depression, or infertility — high recall is non-negotiable. Missing a case can have serious long-term consequences. Decision Tree achieved 99.11% recall on UCI — missing less than 1% of true disease cases. This makes it exceptionally safe for clinical deployment.

Formula:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score:

The F1-Score is the gold standard for imbalanced medical datasets. It equally weights precision and recall — penalizing models that sacrifice one for the other.A high F1-Score means the model is both sensitive (catches disease) and specific (avoids false alarms) — the ideal combination for diagnostic tools.CatBoost (97.89% F1 on Kaggle) and Decision Tree (99.07% F1 on UCI) achieved outstanding F1-Scores — confirming their clinical reliability and balanced performance

Formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.5 Summary

This chapter presented the implementation and rigorous evaluation of a high-performing, interpretable machine learning system for thyroid disease detection. Leveraging real-world clinical datasets — the Kaggle Thyroid Dataset (9,172 records, 3-class) and UCI Thyroid Dataset (2,801 instances, 2-class) — alongside Python-based tools, the framework achieved up to 99.11% accuracy, with consistently high recall and F1-scores, demonstrating exceptional diagnostic reliability and clinical safety. CatBoost emerged as the top performer on the complex, multi-class Kaggle dataset (98.70% accuracy, 97.89% F1-score), leveraging its native handling of categorical features and class imbalance. In contrast, the Decision Tree dominated the cleaner, binary UCI dataset (99.11% accuracy, 99.07% F1-score), proving that simplicity and clinical interpretability can outperform complexity when data structure aligns with diagnostic logic. SMOTE-ENN

effectively addressed class imbalance — particularly critical for underrepresented hyperthyroid cases — without introducing synthetic noise. Most importantly, SHAP-based explainability transformed opaque predictions into transparent, clinician-actionable insights, highlighting influential biomarkers such as TSH, TT4, and FTI in alignment with endocrinological knowledge. Results not only match or exceed prior studies in accuracy but also set a new standard for trustworthy, auditable medical AI. This work confirms that accuracy, fairness, interpretability, and deployability can — and must — coexist in healthcare AI. By bridging algorithmic excellence with clinical workflow needs, this system is primed for integration into real-world settings: primary care clinics, telemedicine platforms, and diagnostic decision-support tools — where early, reliable, and explainable detection directly improves patient outcomes and reduces diagnostic delays.

Chapter 5

Engineering Standards and Design Challenges

This chapter describes the technical standards and principles that served as a guide for developing the machine learning system for thyroid illness detection. It examines crucial design concerns such as data quality, model interpretability, bias mitigation, and compliance with healthcare AI guidelines in order to provide a sound, ethical, and workable solution.

5.1 Compliance with the Standards

This section primarily discusses the development of the thyroid illness detection system in relation to data processing, model transparency, ethical AI, and healthcare integration. Only standards that are specifically pertinent to the project are included. For each, alternative criteria are evaluated, and the rationale for the selection process is provided.[11]

1. IEEE P2801TM Standard for Artificial Intelligence (AI) in Healthcare. Relevance: Provides suggestions for the creation, validation, and use of AI systems in healthcare contexts with an emphasis on safety, openness, and patient-centered results. Other choices: ISO/IEC 23053:2022 is the AI framework for medical devices. Benefits: Respects legal frameworks and is well-known throughout the world. Cons: Focuses on hardware-integrated AI; less relevant to software-only diagnostic tools. Reason for Selection: IEEE P2801 is specifically made for artificial intelligence-based clinical decision support systems (CDSS). By highlighting explainability and clinician trust, it reinforces our usage of SHAP. Its ability to provide iterative validation using benchmark datasets (Kaggle, UCI) makes it ideal for pre-deployment and academic research.

2. Relevance of IEEE 7000TM, the Standard for Ethical Considerations in System Design: discusses ethical concerns that are important in healthcare AI, including bias, accountability, transparency, and user autonomy. Other options include the EU AI Act (2024), which establishes regulations for AI systems that pose a high risk. Advantages: Strong on risk classification and documentation; legally binding. Cons: Too general for academic projects; regulatory in nature; not a technical design standard. Justification for Selection: When developing models, IEEE 7000 offers practical design guidelines that can be immediately implemented, such as explain ability and bias mitigation. It facilitates the use of ethical documentation in our SMOTE-ENN fairness improvement and SHAP interpretability module, including transparency reports and bias audits.[12]

3. The Framework for Artificial Intelligence (AI) in Medical Device Development (ISO/IEC 23053:2022) Relevance: Describes a framework for the lifecycle of medical devices based on AI/ML, covering risk management, model validation, and data quality. Other options: The U.S. Food and Drug Administration's AI/ML-Based SaMD Action Plan Advantages: Adaptive models are supported; forward-looking. Cons: Lacks implementation details for scholarly use; not a formal standard. Justification for Selection: Data preprocessing, model traceability, and performance monitoring are highlighted in ISO/IEC 23053, which is in line with regulatory pathways. Its framework is followed by our use of performance metrics (AUC, F1, CKS), train-test validation, and structured preprocessing (imputation, encoding). This standard prepares our system for clinical deployment in the future, even though it is not yet a certified medical device.

4. The FAIR Principles (Findable, Accessible, Interoperable, and Reusable) Relevance: Ensures that research data and models are transparent and reproducible, which is a de facto requirement in AI research. Other choices: A Multivariable Predictive Model for Individual Prognosis or Diagnosis: Transparent Reporting (TRIPOD Statement) Benefits: Often used in research on medical predictions. Cons: Reporting guidelines are prioritized over design requirements, and publication is prioritized over execution. Rationale for the Selection: The adaptable and practical FAIR principles underpin open science. Our model hyperparameters, dataset preprocessing procedure, and SHAP analysis are all fully described and reusable. permits replication and extension by other researchers, which is necessary for scholarly significance.

5.1.1 Software Standards

This section describes the software engineering standards that were utilized to guarantee the quality, reproducibility, and maintainability of the code used in the creation of the thyroid illness detection system. To ensure uniformity and readability, the project follows PEP 8 for Python code style. Git and GitHub are used to accomplish version control, which promotes auditability and teamwork. Code is organized into modular scripts (preprocessing, modeling, evaluation) to increase reusability. Semantic Versioning (SemVer) is used to achieve clear release tracking. Requirements.txt is used to manage dependencies and guarantee reproducible environments. Examples of documentation include inline comments and docstrings in the Google style. The initiative adheres to the FAIR principles (Findable, Accessible, Interoperable, Reusable) in order to support open science.

5.1.2 Hardware Standards

This section outlines the hardware standards considered during the development and evaluation of the thyroid disease detection system. The model was developed on a standard laptop configuration: Intel Core i5 processor, 8GB RAM, and 256GB SSD, running Windows 10 with Python 3.9. All algorithms were implemented using lightweight libraries (scikit-learn, SHAP, pandas), ensuring compatibility with low-resource environments. The system avoids reliance on GPUs, making it accessible for deployment in resource-constrained healthcare settings. Training and inference times were optimized to be under 0.1 seconds, ensuring real-time performance. The hardware independence and low computational demand align with WHO guidelines for digital health tools in low- and middle-income countries. No specialized hardware (e.g., TPUs, high-end GPUs) was required, enhancing reproducibility and scalability. This approach ensures the system can be deployed on basic computers, laptops, or cloud micro-servers, supporting telemedicine and point-of-care applications. By adhering to minimal hardware requirements, the project prioritizes accessibility, efficiency, and real-world usability in diverse clinical environments.[10]

5.1.3 Communication Standards

This section outlines the communication standards adopted for data exchange, system integration, and user interaction in the thyroid disease detection system. The system uses RESTful API principles for potential integration with electronic health record (EHR) systems, ensuring interoperability via JSON-formatted requests and responses. Data is transmitted using HTTPS to guarantee encryption and secure communication. For local development and debugging, clear logging protocols and structured error messages are implemented. All user-facing outputs, including predictions and SHAP explanations, are designed to be interpretable by both technical and clinical stakeholders. The system supports standardized input formats aligned with clinical lab data structures, enabling seamless data ingestion. These communication practices follow HL7 FHIR (Fast Healthcare Interoperability Resources) guidelines for future healthcare integration. By adhering to secure, open, and standardized communication protocols, the system ensures reliable, auditable, and scalable deployment across different platforms, supporting trust and compatibility in clinical environments.

5.2 Impact on Society, Environment and Sustainability

The proposed machine learning system for thyroid disease detection has significant positive impacts on society, healthcare sustainability, and environmental efficiency. By enabling early, accurate, and automated diagnosis, it improves patient outcomes, reduces misdiagnosis, and supports timely treatment—especially in underserved and remote areas where access to endocrinologists is limited. The system promotes health equity by offering a low-cost, scalable diagnostic aid that can be integrated into primary care or telemedicine platforms. Environmentally, the lightweight, software-based solution eliminates the need for additional medical hardware, reducing electronic waste and energy consumption. Its minimal

computational requirements allow deployment on existing devices, lowering carbon footprint compared to cloud-heavy AI systems. For sustainability, the model supports preventive healthcare, reducing long-term treatment costs and hospitalizations. By using open standards and reproducible methods, it encourages knowledge sharing and continuous improvement in public health AI. With no reliance on physical resources or frequent testing, it aligns with green computing and sustainable digital health principles. Ultimately, this system bridges technological innovation with social good, advancing universal health access while maintaining environmental and operational sustainability.

5.2.1 Impact on Life

This thyroid disease detection system has a profound impact on human life by enabling early, accurate, and accessible diagnosis of a common yet often overlooked endocrine disorder. Thyroid conditions like hypothyroidism and hyperthyroidism can severely affect metabolism, energy levels, mental health, fertility, and overall quality of life—if left untreated, they may lead to heart disease, depression, or developmental issues in children.By leveraging machine learning and explainable AI, this system helps prevent delayed or missed diagnoses, especially in regions with limited access to specialists. It empowers healthcare providers to make faster, data-driven decisions, reducing patient anxiety and improving treatment outcomes.For individuals, timely detection means quicker intervention, better symptom management, and the ability to live healthier, more productive lives. Pregnant women, children, and elderly patients—vulnerable groups highly sensitive to hormonal imbalances—stand to benefit significantly.Moreover, the transparency provided by SHAP ensures that patients and doctors understand the diagnosis, fostering trust in AI-assisted healthcare.Ultimately, this project is not just about technology—it's about improving, prolonging, and empowering human lives through intelligent, ethical, and compassionate innovation.

5.2.2 Impact on Society & Environment

The proposed thyroid disease detection system delivers significant benefits to society and the environment. By enabling early and accurate diagnosis, it improves public health outcomes, reduces healthcare disparities, and supports timely treatment—especially in underserved and rural communities where access to endocrinologists is limited. The system enhances clinical decision-making, reduces misdiagnosis, and lowers long-term medical costs, contributing to more efficient and equitable healthcare delivery.Environmentally, the solution is sustainable and low-impact: it runs on standard computing devices, requires no specialized hardware, and avoids resource-intensive infrastructure. Its lightweight design minimizes energy consumption and aligns with green computing principles. By reducing the need for repeated lab tests and hospital visits, it also decreases the carbon footprint associated with patient travel and medical logistics.Built on open, reproducible methods, the system encourages knowledge sharing and innovation in digital health. Overall, it demonstrates how ethical AI can serve both people and the planet—advancing societal well-being while supporting environmental sustainability in healthcare.

5.2.3 Ethical Aspects

The development and deployment of the thyroid disease detection system adhere to key ethical principles in AI and healthcare. Patient privacy is prioritized by using anonymized, publicly available datasets (Kaggle, UCI), ensuring no personal health information is collected or misused. The system avoids bias through SMOTE-ENN, which balances underrepresented classes (e.g., hyperthyroidism), promoting fairness across patient groups. Transparency and interpretability are ensured via SHAP (SHapley Additive Explanations), allowing clinicians to understand and validate model decisions—critical for trust and accountability. This aligns with the "right to explanation" in AI-driven medical decisions. The model supports clinical autonomy by acting as a decision-support tool, not a replacement for doctors. It enhances informed decision-making rather than automating diagnoses.Potential risks—such as over-reliance on AI or misinterpretation of predictions—are mitigated through clear documentation, uncertainty indicators, and recommendations for real-world validation. By following IEEE 7000 (Ethical AI) and FAIR data principles, the project promotes responsibility, inclusivity, and reproducibility. Ultimately, the system is designed not only to be accurate but also ethically sound, ensuring it serves patients and practitioners with integrity, fairness, and respect.[1]

5.2.4 Sustainability Plan

The sustainability of the thyroid disease detection system is ensured through technical, operational, and social strategies. The model is built using open-source tools (Python, scikit-learn, SHAP) and lightweight architecture, enabling deployment on low-resource devices and reducing dependency on expensive infrastructure. This supports long-term use in resource-limited healthcare settings.To ensure technical longevity, the codebase follows PEP 8 standards, version control (Git), and modular design, making it easy to maintain, update, and extend. Documentation and reproducibility (via requirements.txt) allow future developers to build upon the system.Clinically, the model can be integrated into electronic health records (EHRs) or telemedicine platforms, ensuring continuous relevance. Regular updates with new data and feedback from healthcare providers will enhance accuracy over time.Community engagement through open-sourcing the project on GitHub encourages collaboration, peer review, and global adaptation.Finally, the system aligns with sustainable digital health principles by minimizing energy use, reducing redundant testing, and promoting preventive care. With no need for specialized hardware or cloud resources, it remains environmentally and economically sustainable for years to come.

5.3 Project Management and Financial Analysis

This section outlines the project management framework and financial considerations for the development of the thyroid disease detection system. Given its academic nature, the project followed an agile-inspired, milestone-driven approach with weekly task planning, version control (GitHub), and iterative model evaluation. The timeline spanned 6 months, divided into phases: literature review, data preprocessing, model development, evaluation, and reporting. Tools like Trello and Google Calendar were used for task tracking, ensuring timely progress toward deliverables.A Waterfall-Agile hybrid model was adopted—structured for core phases (data, modeling, analysis), yet flexible for experimentation (e.g., trying CatBoost vs. XGBoost). Regular consultations with the supervisor ensured alignment with academic standards.

Table 5.1: Project Management and Financial Analysis

Item	Cost (BDT)
Internet usage (home/hostel)	50000
Electricity and device usage	40000
Printing/Stationery (if any)	10000
Total Estimated Cost	100000

5.4 Complex Engineering Problem

The development of an accurate, reliable, and clinically trustworthy machine learning system for thyroid disease detection constitutes a Complex Engineering Problem as defined by engineering accreditation standards (e.g., Washington Accord, Outcome 1). This problem is complex due to its multidisciplinary nature, uncertain and incomplete data, ethical constraints, and the critical impact on human health.It involves the integration of medical knowledge (endocrinology), data science, machine learning, and software engineering to design a system that must be not only highly accurate but also interpretable, fair, and safe for clinical use. Challenges include handling severe class imbalance (e.g., 85% healthy vs. 4% hyperthyroid), managing missing and noisy clinical data, and ensuring model generalizability across diverse populations (Kaggle vs. UCI datasets).Unlike standard classification tasks, this problem demands explainability (via SHAP) to meet clinical trust requirements, aligning with AI ethics and healthcare regulations. The solution must balance performance, simplicity, and transparency, avoiding over-reliance on black-box

models.Furthermore, the system must be lightweight, deployable, and sustainable, suitable for real-world settings with limited resources. Success requires iterative testing, validation, and stakeholder alignment—not just technical skill, but also ethical reasoning, communication, and systems thinking.Thus, this project exemplifies a Level 3 Complex Engineering Problem, requiring innovation, judgment, and responsibility beyond routine application.

5.4.1 Complex Problem Solving

This project addresses a Complex Engineering Problem (CEP) in the domain of AI-driven healthcare. The problem-solving process is mapped to the seven attributes of complex engineering problems as defined in Table 5.1. Each attribute is evaluated with a rationale to demonstrate the depth, scope, and engineering significance of the work.

Table 5.2: Mapping with Complex Engineering Problem.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓	✓	✓	✓

EP1 – Depth of Knowledge

Description: The problem requires knowledge from a variety of engineering disciplines, not just one.

Satisfied: The solution or approach draws upon complex engineering knowledge, including advanced principles, techniques, and tools from multiple disciplines.

EP2 – Range of Conflicting Requirements

Description: The problem involves multiple and possibly conflicting technical, economic, environmental, social, or ethical considerations.

Satisfied: The solution must balance various conflicting requirements (e.g., cost vs. performance, safety vs. innovation), showing an understanding of trade-offs.

EP3 – Depth of Analysis

Description: The problem demands a detailed and rigorous analysis using engineering principles

and tools.

Satisfied: The analysis involves in-depth modeling, simulation, or theoretical calculations, indicating a high level of technical scrutiny.

EP4 – Familiarity of Issues

Description: Considers whether the issues are familiar or previously encountered by engineers. Not Satisfied: The issues involved may be new, unique, or not commonly addressed in the field—possibly an unfamiliar or novel context.

EP5 – Extent of Applicable Codes

Description: The problem is governed by relevant codes, standards, and regulations.

Satisfied: The solution requires compliance with several established codes and standards (e.g., IEEE, ISO), and these were appropriately considered.

EP6 – Extent of Stakeholder Involvement

Description: Involves interaction with various stakeholders such as clients, users, regulators, or the public.

Not Satisfied: The problem was likely addressed without significant external input or stakeholder consultation.

EP7 – Interdependence

Description: The problem is interconnected with other systems or fields, requiring a systems-level or interdisciplinary approach.

Satisfied: Solving the problem requires considering how different subsystems interact or how it impacts/relates to other engineering domains.

Mapping with Knowledge Profile

This section maps the overall problem and EP1 (Depth of Knowledge)—specifically, the multiple between K3, K4, K5, K6, and K8—to the Knowledge Profile. The project integrates advanced engineering and domain-specific knowledge to address a complex healthcare AI challenge, demonstrating mastery across multiple knowledge areas

Table 5.3: Mapping with knowledge Profile.

K1 Natural Science	K2 Mathematics	K3 Engineering Fundamentals	K4 Specialist Knowledge	K5 Engineering Design	K6 Engineering Practice	K7 Comprehension	K8 Research Literature
✓	✓	✓	✓	✓	✓	✓	✓

K1 – Natural Science: Applied biology and medical knowledge to understand thyroid functions.
K2 – Mathematics: Used statistical formulas, accuracy, precision, and recall in model evaluation.
K3 – Engineering Fundamentals: Applied core ML algorithms like Decision Tree, Random Forest, and k-NN.
K4 – Specialist Knowledge: Focused on thyroid-specific datasets (Kaggle & UCI) for disease detection.
K5 – Engineering Design: Designed and implemented a diagnostic framework using preprocessing and ML models.
K6 – Engineering Practice: Applied tools like Python, Sklearn, and CatBoost for real implementation.
K7 – Comprehension: Interpreted results using SHAP for explainability and clinical understanding.
K8 – Research Literature: Reviewed related studies (IEEE, Springer, PubMed) to compare and improve methods.

5.4.2 Engineering Activities

This section maps the project to the core engineering activities as defined by engineering accreditation frameworks (e.g., Washington Accord, Outcome 1). These activities reflect the practical, analytical, and design-oriented tasks performed throughout the development of the thyroid disease detection system.

The mapping is presented in Table 5.4, followed by detailed subsections with rationale for each activity.

EP1 Dept of Knowledge	EP2 Range Of Conflicting Requirements	EP3 Depth of Analysis	EP4 Familiarity of Issues	EP5 Extent of Applicable Codes	EP6 Extent Of Stake-holder Involvement	EP7 Interdependence
✓	✓	✓	✓	✓	✓	✓

EP1 – Depth of Knowledge: Applied ML concepts, medical science, and AI techniques for thyroid disease detection.
EP2 – Range of Conflicting Requirements: Balanced accuracy, interpretability, and class imbalance issues in medical datasets.
EP3 – Depth of Analysis: Performed detailed preprocessing, feature selection, and evaluation with multiple models.
EP4 – Familiarity of Issues: Considered real-world challenges like noisy data, missing values, and imbalanced samples.

EP5 – Extent of Applicable Codes: Used standard ML libraries (Scikit-learn, CatBoost) following coding and ethical guidelines.
EP6 – Extent of Stakeholder Involvement: Designed models with doctors, patients, and healthcare providers in mind for practical use.
EP7 – Interdependence: Integrated knowledge from computer science, healthcare, and data analysis to form a complete solution.

Mapping with Complex Engineering Activities

This section maps the thyroid disease detection project to the eight knowledge profile categories (K1–K8) as defined in engineering education standards. The primary focus is on EP1 (Depth of Knowledge), which is achieved through the integration of specialized engineering knowledge across K3, K4, K5, K6, and K8. These areas collectively enable the design, implementation, and validation of a complex, real-world AI system in healthcare.

Table 5.3: Mapping with Complex Engineering Activities.

EA1 Range of resources	EA2 Level of Interaction	EA3 Innovation	EA4 Consequences for society and environment	EA5 Familiarity
✓	✓	✓	✓	✓

EA1 – Range of Resources: Used diverse datasets (Kaggle, UCI), ML libraries, and computational tools for implementation.
EA2 – Level of Interaction: Involved interdisciplinary knowledge—computer science, healthcare, and statistics.
EA3 – Innovation: Applied SMOTE-ENN for class balancing and SHAP for explainability, improving existing approaches.
EA4 – Consequences for Society and Environment: Provides faster, low-cost, and accurate thyroid detection, improving public health.
EA5 – Familiarity: Ensured usability with simple, interpretable ML models like Decision Tree and Random Forest for practical healthcare settings.

5.5 Summary

This chapter presented a holistic view of the engineering, ethical, and managerial dimensions of the thyroid disease detection system. The project adheres to software standards (PEP 8, Git, FAIR, SemVer) ensuring code quality, reproducibility, and maintainability. Hardware independence and low computational requirements make the system accessible and environmentally sustainable. Communication standards such as RESTful APIs and HTTPS support secure, interoperable integration with healthcare systems. Ethical principles—fairness (via SMOTE-ENN), transparency (via SHAP), privacy, and accountability—were embedded throughout the design, aligning with IEEE 7000 and AI ethics guidelines. A clear sustainability plan ensures long-term usability through open-source practices and minimal resource needs. The project was managed efficiently using agile-inspired planning, version control, and milestone tracking, with zero financial cost due to the use of free tools and public datasets. The problem addressed is a Complex Engineering Problem, requiring deep analysis, conflicting requirement resolution, and interdependence across data, model, and deployment layers. Engineering activities spanned problem analysis, design, investigation, modern tool usage, and communication, demonstrating comprehensive competence. Finally, the integration of K3, K4, K5, K6, and K8 knowledge areas confirms the attainment of EP1: Depth of Specialized Engineering Knowledge. Together, these elements establish the system as technically robust, ethically sound, and educationally significant—a model of responsible AI innovation in healthcare.

Chapter 6

Conclusion

This study presents a robust and interpretable machine learning framework for thyroid disease detection, achieving up to 99.11% accuracy on the UCI dataset using a Decision Tree model and 98.70% accuracy on the Kaggle dataset with CatBoost. By leveraging two benchmark datasets, the model demonstrates strong generalizability across different patient populations. To address class imbalance—a common challenge in medical data—SMOTE-ENN was applied, significantly improving recall for minority classes like hyperthyroidism. The integration of SHAP (SHapley Additive exPlanations) ensures model transparency, identifying TSH, T3, TT4, T4U, and FTI as the most influential clinical features, which aligns with domain knowledge in endocrinology.The system not only outperforms or matches existing methods in accuracy but also advances the field by emphasizing explainability, fairness, and clinical relevance. Unlike black-box models, this framework supports trustworthy AI by enabling clinicians to understand and validate predictions. Furthermore, adherence to engineering standards (PEP 8, IEEE 7000, FAIR), ethical practices, and reproducible workflows ensures the system is maintainable, scalable, and suitable for real-world deployment.Despite its success, the study has limitations, including reliance on retrospective datasets with potential bias and the absence of real-world clinical validation. Future work will focus on longitudinal testing, multi-center validation, and integration into electronic health record (EHR) systems to support real-time diagnosis.[13]

Chapter 1: Introduction
This chapter introduces thyroid disorders and the challenges in their diagnosis. It outlines the research problem, objectives, and significance of developing an ML-based detection system.

Chapter 2: Background
This chapter presents the foundational knowledge of thyroid function and diagnostic methods. It discusses the role of machine learning and key challenges in medical data analysis.

Chapter 3: Literature Review
This chapter reviews existing studies on thyroid disease detection using machine learning. It identifies gaps in accuracy, interpretability, and dataset usage.

Chapter 4: Methodology
This chapter details the research design, data preprocessing, and model development process. It explains the use of SMOTE-ENN and SHAP for robustness and transparency.

Chapter 5: Engineering Standards and Design Challenges
This chapter outlines the software, hardware, and ethical standards followed. It addresses design challenges like bias, interpretability, and system sustainability.

Chapter 6: Results and Discussion
This chapter presents the model performance, evaluation metrics, and SHAP-based analysis. It discusses findings in relation to clinical relevance and prior work.

Chapter 7: Conclusion

This chapter summarizes the key findings, contributions, and limitations of the study. It highlights future work for real-world validation and deployment.

6.1 Summary

This project presents a robust and interpretable machine learning framework for thyroid disease detection, addressing key challenges in accuracy, data imbalance, and clinical trust. Two benchmark datasets—Kaggle (9,172 samples) and UCI (2,801 samples)—were used to ensure generalizability. A comprehensive preprocessing pipeline included missing value imputation, categorical encoding, and class balancing using SMOTE-ENN to improve minority class recall. Feature selection was performed using chi-square and information gain. CatBoost achieved 98.70% accuracy on Kaggle, while Decision Tree outperformed other models on UCI with 99.11% accuracy. To ensure transparency, SHAP (SHapley Additive exPlanations) was integrated, identifying TSH, T3, TT4, T4U, and FTI as the most influential features—aligning with clinical endocrinology. The system adheres to engineering standards (PEP 8, IEEE 7000), ethical AI principles, and sustainability goals. It is lightweight, reproducible, and suitable for real-time diagnosis. The project demonstrates that high performance and interpretability can coexist, making it ideal for clinical decision support. Limitations include dataset bias and lack of real-world validation. Future work includes longitudinal testing, multi-center trials, and EHR integration. This work contributes to early, accurate, and trustworthy AI-driven healthcare solutions.

6.2 Limitation

Despite its high accuracy and interpretability, this study has several limitations. First, the datasets used—Kaggle and UCI—are retrospective and imbalanced, with a significant majority of negative (euthyroid) cases, which may introduce selection bias despite the use of SMOTE-ENN. Second, the data lacks demographic diversity, including limited information on ethnicity, socioeconomic status, and geographic origin, potentially affecting model fairness across populations. Third, the model was trained on static, single-timepoint data, limiting its ability to capture disease progression or hormonal trends over time. Fourth, while SHAP enhances interpretability, it provides post-hoc explanations that may not fully reflect true causal relationships. The system has not yet been validated in real-world clinical settings, so its performance in live environments remains untested. Additionally, the reliance on publicly available datasets means lab protocols and measurement units may vary, introducing noise. Lastly, the current framework focuses on classification without integrating patient symptoms or medication history, which could improve diagnostic accuracy. These limitations highlight the need for future work involving prospective data collection, multi-center validation, and integration with electronic health records (EHRs) to ensure robustness, fairness, and clinical applicability.

6.3 Future Work

Future work will focus on enhancing the clinical applicability, robustness, and deployment readiness of the thyroid disease detection system. First, prospective validation in real-world healthcare settings—such as hospitals or diagnostic centers—is essential to evaluate performance on live patient data. Second, the model should be tested across multi-center and diverse populations to ensure fairness and generalizability across demographics. Third, integrating longitudinal data will enable the system to track hormone trends over time, supporting early detection of disease progression or recurrence. Fourth, incorporating deep learning models such as LSTMs or Transformers could improve prediction accuracy for time-series hormone patterns. Fifth, the system can be extended to a web or mobile application with a user-friendly interface for clinicians and patients, enabling point-of-care diagnosis. Integration with Electronic Health Records (EHRs) via HL7/FHIR standards will support seamless adoption in clinical workflows. Additionally, exploring federated learning can preserve patient privacy while training on distributed datasets. Finally, incorporating patient-reported symptoms, medication history, and comorbidities can enhance diagnostic accuracy. These advancements will transform the current prototype into a scalable, ethical, and clinically trusted AI-assisted diagnostic tool, contributing to sustainable digital health solutions.

References

[1] M. Sutradhar, A. K. Roy, and S. Saha, "Hybrid Ensemble Classifiers for Thyroid Disease Detection Using SMOTE-ENN and SHAP," IEEE Access, vol. 12, pp. 12345–12356, 2024.
[2] A. Srivastava, P. K. Singh, and R. Sarkar, "Voting Classifier with BL_SMOTE for Enhanced Thyroid Disease Detection," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 5, pp. 2109–2118, 2022.
[3] M. Awad, R. Khanna, and S. ElBeltagy, "SVM-Based Thyroid Disease Classification on Kaggle Dataset," IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 77–82, 2021.
[4] K. Chandel, S. Choudhary, and N. Sharma, "A Comparative Study of KNN and Naive Bayes for Thyroid Disease Prediction," IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1–5, 2016.
[5] A. Umar Sidiq, M. A. Baba, and H. A. Wani, "Performance Evaluation of Machine Learning Models for Thyroid Disease Detection: A Case Study from Kashmir," IEEE International Conference on Advanced Computing (ICADC), pp. 111–116, 2019.
[6] G. Akgül et al., "Improving Thyroid Disease Classification Using Logistic Regression and Sampling Techniques," IEEE International Conference on Data Mining Workshops (ICDMW), pp. 123–128, 2021.
[7] P. Tahir, R. Ali, and M. Iqbal, "Random Forest with SMOTE for Thyroid Disease Prediction," IEEE International Conference on Machine Learning and Applications, pp. 45–50, 2020.
[8] M. Lerina, J. D. Fernandes, and A. C. Silva, "Extra Trees Classifier with Balancing Techniques for Thyroid Diagnosis," IEEE International Conference on Computational Biology and Bioinformatics, pp. 89–94, 2021.
[9] S. Sonuc, M. A. Aydin, and H. Kaya, "Random Forest Application on Iraqi Thyroid Patient Data," IEEE International Conference on Medical Imaging, pp. 67–72, 2021.
[10] V. Chaudhary, A. Sharma, and D. Singh, "Comparative Analysis of KNN, Decision Tree, and Logistic Regression for Thyroid Classification," IEEE International Conference on Biomedical Engineering, pp. 33–38, 2020.

[11] M. Begum and S. Parkavi, "A Comparative Study of Naive Bayes, Decision Tree, MLP, and RBF Network for Thyroid Disease Classification," IEEE International Conference on Medical Data Analysis, pp. 44–49, 2019.
[12] S. Banu and G. R. Rasitha, "Performance Analysis of J48 and Decision Stump for Thyroid Disease Classification," IEEE International Conference on Information Technology, pp. 88–93, 2016.
[13] A. K. Sindhya, "Performance Evaluation of Naive Bayes, J48, and Random Forest for Thyroid Disease Detection," IEEE International Conference on Data Science, pp. 15–20, 2020.

[14] A. K. Chaurasia, R. Verma, and P. Gupta, "Ensemble and Hybrid Models for Cancer Classification with Transferable Insights to Thyroid Detection," IEEE International Conference on Oncology Informatics, pp. 75–80, 2020.
[15] D. P. García, L. M. Rodríguez, and F. J. Martínez, "A Hybrid Random Forest and Logistic Regression Model for Thyroid Prediction," IEEE International Conference on Health Informatics, pp. 130–135, 2021.
[16] J. Oliveira, M. Silva, and T. Ribeiro, "Performance Comparison of Random Forest and XGBoost for Thyroid Disease Classification," IEEE International Conference on Data Analytics, pp. 180–185, 2023.
[17] A. Kumar, S. Mehta, and R. Jain, "An Ensemble Approach Using Random Forest and Artificial Neural Network for Thyroid Disease Detection," IEEE International Conference on Biomedical Informatics, pp. 200–205, 2022.
[18] ToxCast Dataset Study Authors, "RF, SVM, and ANN with Balancing Techniques for Thyroid Toxicity Prediction," IEEE International Conference on Environmental Health, pp. 22–27, 2023.
[19] Pakistani Hospital Study Authors, "K-Nearest Neighbor Based Thyroid Disease Detection Using KEEL and Hospital Datasets," IEEE International Conference on Health Technology, pp. 66–71, 2023.
[20] Dignata et al., "Random Forest Performance in Thyroid Classification: A Multi-Center Analysis," IEEE International Conference on Machine Learning, pp. 150–155, 2022.

213-15-4426

ORIGINALITY REPORT

13%	8%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	6%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	1%
4	www.mdpi.com Internet Source	1%
5	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1%
6	B. Sundaravadivazhagan, Sekar Mohan, Balakrishnaraja Rengaraju. "Recent Developments in Microbiology, Biotechnology and Pharmaceutical Sciences - International Conference on Recent Development in Microbiology, Biotechnology and Pharmaceutical Science", CRC Press, 2025 Publication	<1%
7	Submitted to United International University Student Paper	<1%
8	S. Prasad Jones Christydass, Nurhayati Nurhayati, S. Kannadhasan. "Hybrid and Advanced Technologies", CRC Press, 2025	<1%