Thyroid Disease Prediction Using Machine Learning Algorithm

MUKABBIR HOSSAIN

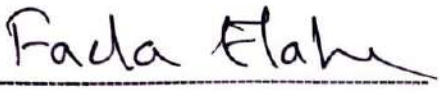
Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

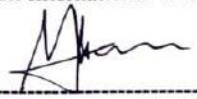
This thesis titled on "Thyroid Disease Prediction Using Machine Learning Algorithm", submitted by **Mukabbir Hossain (ID: 221-35-974)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS



Chairman

Dr. Fazla Ealhe
Assistant Professor & Associate Head
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



Internal Examiner 1

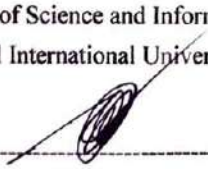
Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Dr. Shabnom Mustary
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Md. Rajib Mia
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University



External Examiner

Mohammod Abul Kashem, PhD
Professor
Department of Computer Science and Engineering
DUET, Bangladesh

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	MUKABBIR HOSSAIN
Date of Birth	:	30.12.2002
Title	:	Thyroid Disease Prediction Using Machine Learning Algorithm
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☑	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

Mukabbir

(Student's Signature)

221-35-974
Date: 23.12.2025

(Supervisor's Signature)

Dr. Marzia Ahmed
Date: 23.12.2025

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science of Science.

(Supervisor's Signature)

Full Name : Dr. Marzia Ahmed

Position : Assistant Professor, Department of Software Engineering, Daffodil International University

Date : 23.12.2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Mukabbir

(Student's Signature)

Full Name : MUKABBIR HOSSAIN

ID Number : 221-35-974

Date : 23.12.2025

Thyroid Disease Prediction Using Machine Learning Algorithm

MUKABBIR HOSSAIN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Software Engineering)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I would like to express my appreciation to my supervisor, Dr. Marzia Ahmed Ma'am I have received endless support and she provided me with advice invaluable to use in this research work, and even made constructive advice, during the research process. It is also my pleasure to the various open-source communities and data-suppliers who have enabled me to make this research effort possible. Finally, my family is my greatest asset to whom I am grateful as they have been encouraging and patient in this monumental undertaking. This research would not have been known to the light of day without their moral support and understanding.

ABSTRACT

Thyroid diseases are those disorders which are not easily detected because of their nondescript initial symptoms and complicated diagnosis. This study presents a systematic study that involves machine learning models to forecast thyroid disease at their early stages. A thyroid dataset was acquired on the UCI Machine Learning Repository UCI Machine Learning Repository from Kaggle was utilized, it's containing 9172 patient records with 31 features and a binary target indicating the presence or absence of disease. Only 11 clinical features, along with 2 categorical features and 1 binary feature were taken to predict the thyroid disease. The data set had a significant class imbalance, so we use the Synthetic Minority Over-sampling Technique (SMOTE) was applied to ensure robust training. Seven different machine learning classifiers were trained and tested. Model performance was evaluated on a stratified hold-out test set using 1,000-iteration Non-Parametric bootstrap internal validation to obtain robust estimates and 95% confidence intervals for accuracy, sensitivity, specificity, precision, F1-score, and AUC. The results indicate that the Random Forest classifier provides superior sensitivity of 96.5% that making it a reliable tool for early screening.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction of Thyroid Diseases

Thyroid itself is a small and butterfly-shaped gland nestled low at the neck, but the influence of its hormones is hardly small (Alawiyah, 2024). By controlling metabolism, and functions of temperature and the cardiovascular system, the thyroid keeps the rhythm of life beneath the surface. And when that rhythm goes away, the symptoms are often so slight fatigue, fluctuations of seven, and changes of mood that they might easily be attributed to the stresses of the day or other factors (Chaubey, 2021). This makes early detection a challenge. The evaluation of delayed detection is a serious concern, since the risk of complications and financial burdens of ongoing and repeated treatments is very costly. The experience of the medical community within a given paradigm might itself serve to introduce obstructions into the diagnostic process, volume of referrals, the time required to see the results of tests, and the complexity of reviewing clinical data that then needs to be interpreted within a larger framework (Chaganti R, 2022) (Teja, 2024). But the type of pre-screening assistance could make such a process much smoother. This paper inquires could contemporary ML techniques could serve as such an early detection system with sufficient accuracy and reliability to notify high-risk patients before deadly symptoms are visible, and transparency that will convince medical professionals of the worth of the alerts. (Sennan, 2022) The research objective now is to generate and evaluate a prediction model that can lead us toward patients with likely thyroid illness, so long as the deadliest error in the early screening process failing to identify a true positive, is kept well under check. This study aims to prove this hypothesis on an aggregated Thyroid Disease dataset received from the UCI Machine Learning Repository through a Kaggle dataset (Alawiyah, 2024) (Riley R. D., 2023). The final cleaned dataset used within this paper consists of 9172 entries with 14 features, 13 of which are predictive features, and 1 is a target feature with a binary value indicating whether a patient has a thyroid disease or not. The features introduced include a series of demographics and standard blood tests such as ‘TSH,’ ‘T3,’ ‘TT4,’ ‘T4U,’ and ‘FTI,’

which can be used collectively for an understanding of a patient's thyroid function (Akter, 2024) (Chaganti R, 2022). The dataset is significantly biased towards healthy individuals only, with a very small proportion classified as those with a 'thyroid disease.' This directly impacts the approach taken within this paper on the models designed to solve it. We used SMOTE to balance the training data set (Chawla, 2002) (J. Premalatha, 2024) to ensure the models learn the pattern, and in the test set's those work accurately.

.

1.2 Background

The thyroid gland is a small endocrine organ with a very prominent role in everyday physiology. The T3 and T4 hormones and, through the pituitary feedback signal TSH (Thyroid-Stimulating Hormone), the control of the thyroid gland itself control metabolism, heat production, the cardiac rate and rhythm, and, indirectly, mood and energy levels. "Patients with thyroid dysfunction often appear nonspecific: they may appear tired, anxious, be gaining or losing weight, and experience heat and cold intolerance. (Akter, 2024)"

Thyroid conditions are relatively common, very significant, and surprisingly difficult to recognize early. The involvement of the gland's hormones with metabolism, the cardiovascular system, the control of body temperature, and the regulation of mood makes it extremely difficult to detect such conditions, which often manifest themselves through common symptoms of fatigue and even changes in weight and/or anxiety. This obscuring of symptoms contributes to the challenges of early detection of the condition and therefore increases the risk of complications and additional healthcare costs spent on multiple tests and visits (Teja, 2024). The increased flow and turns add to the bottleneck when we need fast decisions. And the types of biological data that matter most for this task aren't all that esoteric. TSH is the lab "front-door" as small TSH changes are very telling of the extent that the pituitary glands has to push your thyroid to keep hormones in balance. The TSH is usually elevated with overt hypothyroidism, and suppressed with overt hyperthyroidism (Sennan, 2022). And then size, that's the other kind of data that's usually meaningful, are medical history flags and medications and then background information. A good screener will both (1) focus on TSH and its

properties, and (2) not let you trip over the similarly informative but less central data. The problem for the largest part of medical applications, and the one being tackled in a practical sense by this thesis, is detecting early warning signs (such as possible thyroid disorders) from all available data. This thesis addresses this problem thought a data-based and organized approach to disease whose goal is for these thyroid logically profiled outputs to matter.

1.3 Problem Statement

Thyroid problems are common, but they hardly ever walk in with a label. They come in feeling tired, a little anxious, gaining or losing weight, too hot or cold and all of that can be caused by dozens of things. Clinicians do the best they can with history, examination, and lab tests like TSH, but busy clinics and long queues mean it is often some time before the right people get the right attention (Teja, 2024). In that gap, the patients can be left waiting and the system expends energy on low-risk cases while the high-risk ones are not flagged early enough. This thesis focuses to find out the thyroid diseases that can predict before doctor's diagnosis and it's pretty useful I am not challenging the diagnosis but I think it will boost our doctor's diagnosis system pretty easier. The data that we rely upon is the same data that the clinics are already gathering: basic demographic information, medical history indicators, and standard lab work particularly TSH, which is often the first indication of a change in thyroid functioning (Chaubey, 2021). The dataset is well-ranked for a machine learning model because the information is structured, because there are identifying patterns that a person would understand, and because there are interactions that are recognizable to algorithms and not so much to the end user's eye.

1.4 Motivation

Thyroid diseases are common but announcing their presence infrequently. Symptoms that are generally nonspecific at the outset fatigue, weight change and an unusual sensitivity to cold or heat frequently have patients bouncing from appointment to appointment before anyone thinks of thyroid dysfunction. Meantime, clinics are full and resources are limited. Every week, doctors sift through a ton of low-risk cases in order to find a handful who really truly need immediate attention. That time lag comes with a

cost: Untreated thyroid problems can silently chip away at quality of life, exacerbate other health conditions and lead to repeat testing and visits that clog up both patients and systems. Machine learning fits these requirements since the information expressed by thyroid scanning is tabled data with clear and defined levels of interaction. With robust process flows including signing into data, training models to know the issue is an imbalanced one and assessing performance of models which are suitable for purpose, dull patient information becomes a credible triage signal. The immediate implications of adopting a model of this kind are clear: patients at high risk get swift attention, whereas those at low risk do not need to be followed unnecessarily. That could mean a week or two shaved off at the front of the care chain for patients. This adds clarity and avoids confusion in the future.

1.5 Limitations

Our research has real-world limitations. We are not intending to cover all the possible methodologies of thyroid condition detection, nor its potential model of detection. Instead, we have strived for a practical list of methods that is suitable for the data at our disposal and the objectives of early screening. Likewise, although we explore a number of algorithms through the machine learning, this is not a comprehensive list of every model and set of parameters. Second, accuracy is always limited: there are no perfect models and can make no guarantees with regard to perfect predictive power for all individuals and clinics.

CHAPTER 2

LITERATURE REVIEWS

2.1 Literature Reviews

A Recent approaches based on machine learning have been demonstrated as effective tools for the prediction of thyroid diseases, but challenges still arise in dealing with imbalanced datasets, quantifying uncertainty and prioritizing sensitivity for the purpose of clinical screening applied several classifiers on thyroid datasets: they demonstrated high diagnostic accuracy of the machine learning system, but also their concerns on class imbalance effect in detecting minority-class (Chaubey, 2021). (Sennan, 2022) Sankar et al adopted XGBoost for better accuracy than other models in thyroid disease diagnosis based on non linear relationships between laboratory and clinical features . (Akter, 2024) Akter et al built on this work by incorporating clustering-based balancing with ten different models such as ANN, CatBoost, XGBoost, Random Forest, and LightGBM with not only focusing accuracy but also interpretability along with deep feature importance analysis of imbalanced thyroid data. A few works have taken more general systemic comparison of thyroid prediction algorithms. (J. Premalatha, 2024) Premalatha et al. demonstrated that Decision Tree models after Random Forest based pre-processing reach the highest accuracy (98.54%), compared to Logistic Regression, SVM, and Ada Boost, but no formal interval estimates are provided for performance measures. (Chaganti R, 2022) Chaganti et al. presented a selective feature approach (Forward and Backward elimination, Extra Trees) to identify significant predictors like FTI, TSH-measurement f lags, and referral source with close to perfect performance on all thyroid disease categories (Patel, 2024). In addition to initial diagnosis, (Aversano, 2021) Aversano et al. showed the treatment prediction using machine learning (on clinical and therapeutic variables-data-driven methods can be used for therapy planning, besides diagnosis (Aversano, 2021). (Dhyan Chandra Yadav, 2020)Yadav et al. employed decision tree ensemble learning to predict thyroid conditions, striking the tradeoff between model predictivity and explicability with explicit rule paths that can be explained based on

clinical reasoning. Recent work has unveiled innovative ensemble and deep learning techniques for thyroid disease. (Ji, The novel self-stack ensemble model for thyroid disease prediction, 2024) Ji proposed SSC, a self-stack ensemble model that exploits a random tree structured stacking structure to ameliorate the deficiencies of traditional ensembles and significantly enhanced the generalization performance on thyroid prediction tasks. (Alawiyah, 2024) Alawiyah et al. applied XGBoost to predict thyroid cancer recurrence on clinicopathological features, including tumor size, lymph node metastasis, and vascular invasion, with an accuracy of 97.74% and F1-score at 95.94%, showing the competence of gradient boosting for complicated oncologic endpoints. (Teja, 2024) Teja et al. present a recent survey of machine learning techniques for thyroid disease, which lists popular algorithms and feature sets but do not introduce any new validation framework. (Ji, The novel self-stack ensemble model for thyroid disease prediction, 2024) Jha et al. addressed the prediction accuracy improvement for thyroid disorders in an imbalanced setting by employing an ensemble technique, and showed a significant increase in both sensitivity as well as specificity, thereby strengthening the need to balance minority class data for public health gain. Narrative and methodological reviews have highlighted more general trends with respect to this area and prompted the development of more rigorous assessment tools. Martin et al. Martin et al examined the role of artificial intelligence in thyroid ology, noting that ensemble decomposable tree-based models and deep neural network structures are becoming progressively critical for early diagnosis of thyroid disorders using automatically generated models, especially when applied with structured biochemical and imaging data (Chaganti R, 2022) (Shrestha, 2025). (Fasira, 2025) Fasira et al. reviewed deep learning techniques for the diagnosis of thyroid diseases, in which they report convolutional and recurrent architectures as promising models, but not with externally validated or explainable performance on most studies published. Reiterated that clinical prediction models, including those derived using machine learning, need internal validation commonly performed through bootstrap re sampling- to overcome optimistic performance estimates and to provide trustworthy confidence intervals on metrics such as discrimination and calibration; (Riley R. D., 2023). (Fornaciari, 2022) Fornaciari et al. apply 1000-iteration bootstrap procedures to fixed evaluation sets as a means of inferential statistics for model performance, thereby demonstrating how bootstrapped-based confidence intervals can bring stability to conclusions about classifier quality in applied domains.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter provides a detailed account of the systematic methods used to develop, evaluate, and interpret machine learning models for the early diagnosis of thyroid disease. The experimental methodology is devised to be systematic in a way that not only does the resultant prediction model have statistical stability, but also can provide reliable predictions for clinical use, especially when missing such identification errors becomes too expensive. The work is divided into seven integrated stages, as outlined in the research framework (J. Premalatha, 2024) (Akter, 2024). This working strategy involves obtaining a dataset, cleaning it, and comparing several classifiers in order to produce the best results by handling the data. The first stage of the research is to perform robust data cleaning activities in order to guarantee both the integrity and fairness during model training. This encompassed important Data Pre-processing steps such as addressing missing values with specific imputation methods, transforming all categorical and non-numeric characteristics by encoding them, and scaling numerical ones (Alawiyah, 2024). Such standardization is essential to avoid an artificial domination of the learning by one feature. The second, equally important one was Dealing with Class Imbalance. Since the dataset is highly imbalanced, with a large number of samples from healthy compared to diseased patients, the re-sampling by SMOTE was carefully turned on only in the training set (Aversano, 2021). This method effectively synthesized minority class samples, which has reduced of the training set in the region, and avoid models defaulting to majority classes so that we can make the model achieve the high capture ratio needed for clinical screening. After preparing the data, the research moved to Model Selection and Evaluation. A varied package of seven machine learning classifiers was trained and compared from heavy ensemble methodologies, such as Random Forest, XGBoost, K-Nearest Neighbors (KNN), or Logistic Regression. All the resulting models were evaluated on a completely unseen,

unmodified test set with Non-Parametric Bootstrap Internal Validation and 95% Confidence Interval (Fornaciari, 2022). In line with our key research aim, emphasizing patient safety through the prevention of missed diagnoses the performance analysis mainly focused on Recall (Sensitivity): whether disease cases were detected. This analysis helped to guarantee that the model selected after analyses had a reasonable trade-off between sensitivity and overall accuracy.

3.2 Proposed Framework

This methodology has been geared towards ensuring that the model is both repeatable and transparent to a considerable degree. The steps of the methodology are considered to be inter-linked and comprise seven steps: data acquisition, data preparation, train-test data splitting, class balancing, model training, Non-Parametric bootstrap internal 1000X validation on test data set (Fornaciari, 2022), along with 95% confidence interval for performance analysis. The first and foremost goal of the methodology used in this paper is to extract the best-performing algorithm of for the classification of thyroid diseases that focuses on sensitivity (Recall). Here is the proposed framework diagram. Here is the figure 1 where it's shows how the procedure works in details.

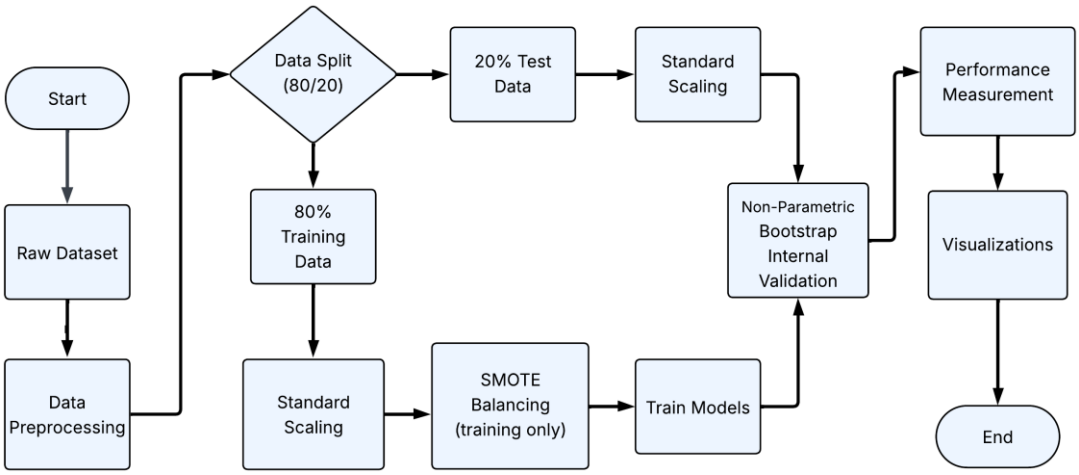


Figure 1 Thyroid Disease Prediction Methodology Flowchart

3.3 Dataset Description

We used the Thyroid Disease data from UCI's Machine Learning Archive, it's well known for testing thyroid predictions. The dataset used in this study is a consolidated version of thyroid disease data from the UCI Machine Learning Repository, which was distributed in curated form through an open-source Kaggle resource. After cleaning and selecting the clinical feature only, the final dataset used in this paper contains 9172 patient records and 14 columns, of which 13 are input features, and 1 is the binary target variable (target). The target indicates whether a patient is diagnosed as having thyroid disease (1) or not (0). These features broadly fall under three categories, as high lighted in Table 1.

Table 1 FEATURE CATEGORIZATION AND DESCRIPTION

Category	Features	Count
Demographic	age, sex	2
Medical History	lithium, goitre, tumor, psych , hypopituitary	5
Laboratory Tests	tsh, t3, tt4, t4u, fti,tbg	6
Total Count		13

3.4 Data Pre-processing

Since it's a real-world data from Kaggle in this case, measurements in a medical study, is being used, there is a need to pre-process it, which may involve dealing with missing values, categorical data, and possibly handling problems of class imbalance. In order to make sure that each of the 14 features equally makes a contribution to the prediction made by the model, a data scaling method was used in the pre-processing stage. The ranges of features like 'TSH' and 'Age' are very different, and so the Min-Max Normalization was applied to the values and made them fall within a common range of 0 to 1. This makes the model non-biased to features whose numerical values are large. Each step of the pipeline began by specifying the target variable for thyroid diseases. Hers is the formula of Min-Max normalization.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \textbf{.....(1)}$$

Where,

X_{norm} is the normalized value

X is the original feature value

X_{min} and X_{max} is the minimum and maximum value of the corresponding feature

Variables that acted as identifiers, like patient numbers, were eliminated since they did not provide any data related to patient health and may otherwise cause bias during the learning phase. Variables that remained relevant served as demographic data, past health indicators, or lab results for functions of the thyroid (Chaganti R, 2022).

2. Handling Missing and Uninformative Features :

Each step of the pipeline began by specifying the target variable for thyroid diseases. Variables that acted as identifiers, like patient numbers, were eliminated since they did not provide any data related to patient health and may otherwise cause bias during the learning phase. Variables that remained relevant served as demographic data, past health indicators, or lab results for functions of the thyroid.

3. Encoding and Target Handling :

Categorical variables like gender and other binary variables were encoded into a numeric format using label encoding. This made it possible for the models to process the categorical data without losing the original label (Benabid, 2024). The target variable was encoded into a binary numeric format. The target column encoded the class label as:

0.0 → No thyroid disease

1.0 → Thyroid disease present

This binary format immediately made the data compatible with the classification algorithms.

4. **Data Train and Test Split:** To perform fair evaluations of model performance, the dataset was split using an 80/20 stratified split, preserving the original class proportions in both sets. This resulted in:

TRAIN SET: 7337 samples × 13 features

TEST SET: 1835 samples × 13 features.

For categories such as sex or past health status we have turned text categories into numbers through Label Encoding. Similarly, we indicated in the outcome column what one's health condition is: 0 meant no disease while 1 equalled some illness. Translating labels like these into numbers maintains the information, but an easier "language" for machines to understand (Riley R. D., 2024) (Shrestha, 2025).

5. **Feature scaling or normalization:** The value of data numbers may be very different according to what is being measured. For example, age ranges from 1, 4 to about 100 and TSH takes values smaller than 10. If these data are used 'as is', models such as KNN discriminate against large numbers. Therefore, we maintained mathematical rules of tradition and used a Standard Scaler, which is

$$z = \frac{x-\mu}{\sigma} \quad \text{.....(2)}$$

where z

x is what we refer to as z after making it standard

x represents the beginning value, and μ represents the average

μ represents the average value of the characteristic, and σ

σ is the measure of how spread out the values are. What this achieves is that it centres all features on zero, and yet they remain standardized on one, so they line up well. The big thing is, we are training the scaling tool only on the training piece, then doing the same thing on the testing piece. Here is the snapshot of Feature scaling was implemented using Standard Scaler as shown in Code Snippet 1

```
# Code Snippet 1: Feature Scaling Implementation
from sklearn.preprocessing import StandardScaler

# Initialize scaler
scaler = StandardScaler()

# Fit on training data only
X_train_scaled = scaler.fit_transform(X_train)

# Transform test data using training statistics
X_test_scaled = scaler.transform(X_test)
```

Figure 2 Code Snip of Feature Scaling

3.5 Data Distributions and Testing

For an objective comparison of model performance, a ‘test set’ was held back unseen throughout the entire learning process. A total of 9172 records were used in full, divided into:

Training set: 7,337 examples (80%)

Test dataset: 1,835 examples (20%)

The train and test data split is stratified so that the ratio of healthy cases to sick cases is roughly 68.6% vs. 31.4% in both sets. This is especially necessary in cases of an extremely unbalanced dataset, because otherwise, sick cases may be relegated entirely to one split set. A constant random seed with value 42 has been used for the purpose of reproducibility. All other processing steps which may pose a potential threat of information disclosure, including scaling and SMOTE, have only used the training dataset, with the unseen dataset being a realistic clinical scenario.

3.6 Class Balancing

The problem is that the original dataset was markedly imbalanced, with only 31.4 % of the dataset being positive cases (patients with thyroid disease). This means a very simple model would work well with a high accuracy value by simply guessing "no disease" for most people. This may look beneficial with a high accuracy value, but it is a problem because many actual disease cases would be overlooked. This problem can be addressed by applying the 'Synthetic Minority Over-sampling Technique', also known as 'SMOTE', only on the 'training dataset' (Chawla, 2002). Unlike other techniques used for handling the minority problem, unlike simply replicating similar cases for the positive class (the disease), 'SMOTE' generates new cases for the minority class by connecting each minority example with its 'nearest neighbours' in the 'feature space' with a line segment, thus "filling in" areas around actual sick individuals with 'synthetic' examples. Here is the equation for the SMOTE

$$x_{new} = x_i + \lambda(x_{nn} - x_i), \quad \lambda \in (0,1) \text{.....} \textbf{(3)}$$

Where:

x_i = minority class sample

x_nn = nearest neighbour

λ = random number between 0 and 1

The procedure followed in this study was: Begin with the original set of 7337 examples used for training. Use SMOTE with the parameter random_state=42 for creating a balanced dataset. I have left the test dataset intact, ensuring it reflects the true class

distribution. During the process, these models view sufficient examples with a positive outcome so they can establish relevant patterns for a thyroid problem, and during testing, performance is evaluated on a realistic, but imbalanced data distribution that provides a fair approximation of what the performance would be during a realistic clinical situation. Here is the code snipe of SMOTE

```
from imblearn.over_sampling import SMOTE

# Initialize SMOTE with fixed random seed
smote = SMOTE(random_state=42)

# Apply to training data only
X_train_balanced, y_train_balanced = smote.fit_resample(
    X_train_scaled,
    y_train
)

# Verify balance
print(f"Before SMOTE: Healthy={sum(y_train==0)}, Disease={sum(y_train==1)}")
print(f"After SMOTE: Healthy={sum(y_train_balanced==0)}, Disease={sum(y_train_balanced==1)}")
```

Figure 3 Code snip of SMOTE

3.7 Models Introductions

For a thorough understanding of the performance of different learning paradigms on the predictive analysis of the thyroid disease, seven different supervised classifiers from varying algorithmic families were chosen:

Random Forest: Random Forest is a combination of decision trees induced from Non-parametric bootstrap samples of the dataset, with a random subset of features used at each node. This introduces decorrelation among the trees and helps with generalization. For this study, a Random Forest with 100 trees was employed. The ultimate decision is made by a majority vote among all the trees. Random Forest can be used to compute the importance of features, which is useful for clinical interpretation.

XGBoost (Extreme Gradient Boost): The decision trees in XGBoost are built sequentially, with each new tree trying to optimize the errors of the previous ensemble solution. It uses a form of gradient boost with regularization terms L1, L2, shrinkage, and tree pruning, which helps in preventing overfitting and makes it a very commonly

used technique in medical prediction problems involving tabular data. For this analysis, the settings are a learning rate of 0.05 and a max depth of 3 .

Gradient Boosting: Gradient Boosting is a similar algorithm to XGBoost, but is coded within scikit-learn with a much simpler and less heavily optimization-tuned engine. Usually, Gradient Boosting is a robust algorithm on structured data, although slower than, or even regularized by comparison with, XGBoost (Shiuh, 2023) (Obaido, 2025).

Decision Tree: The decision tree classifier is a decision tree where the decision is represented in the form of a tree split on the features, with the leaves labelling the classes. It is very visual and interpretable but tends to overfit if grown unwarily with unrestricted growth. A constraint on the growth of the tree by the imposition of a limit on depth, such as max depth = 15, and default random state is used in this research.

Logistic Regression: The aim of Logistic Regression is to model the log odds of the positive class with a linear combination of features. It is a surprisingly effective method on tabular data despite its simplicity, and provides coefficients with intuitive interpretations regarding the characteristics' strength and direction.

KNN: K-Nearest Neighbors (KNN) is a non-parametric and simple algorithm that classifies the instances in the feature space by the majority-class of the k closest similar instances in the feature space.. Most of these k neighbours vote to designate the type of test sample. We used KNN with k=5 neighbours in this study in order to strike a balance between bias and variance so that the algorithm would be sensitive to small scale patterns in the thyroid disease data, but would still be reasonably efficient in terms of computational overhead.

Naive Bayes: Naive Bayes is a probabilistic classifier which follows the Bayes probabilistic theory . Naive Bayes has 5 been observed to work best on high-dimensional data and is computationally inexpensive both to train and predict. This probabilistic method offers a new outlook on the tree-based and distance-based frameworks and will be useful in delivering comparative analysis on the model performance on the thyroid disease classification problem.

Here is the code snip of the model's

```
models = {
    'Random Forest': RandomForestClassifier(n_estimators=100, random_state=42, n_jobs=-1, max_depth=15),
   'XGBoost': XGBClassifier(
        n_estimators=100,
        learning_rate=0.05,
        max_depth=3,
        eval_metric='logloss',
        verbosity=0,
        random_state=42
    ),

    'Gradient Boosting': GradientBoostingClassifier(
        n_estimators=100,
        learning_rate=0.05,
        max_depth=3,
        random_state=42
    ),
    'Decision Tree': DecisionTreeClassifier(max_depth=15, random_state=42),
    'Logistic Regression': LogisticRegression(max_iter=1000, random_state=42),
    'KNN': KNeighborsClassifier(n_neighbors=5),
    'Naive Bayes': GaussianNB()
```

Figure 4 Code snip of 7 Models

3.8 Performance Measurement

We checked each model for four performance score of the model kinds

Accuracy: Accuracy indicates how many of the patient were , regardless of type. You figure it out by:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{…….. …..(4)}$$

Precision: Precision is the percentage of the positive and correct guesses.

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad \text{…… ……(5)}$$

The high accuracy will reduce the frequency of false alerts that enable the model to rectify in detecting a disease in most instances. It matters due to the fact that the wrong warnings may stress individuals and precondition the needless additional testing.

Recall (Sensitivity) : Shows how many true sick people model found it's allows model to find out it's potentiality.

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} \quad \textbf{.....(6)}$$

F1-Score : F1-Score really helps when you're just as concerned about missed cases as well as wrong alarms. F1-score is used to compare Recall and Precision to achieve improved result.

$$\textbf{F1-Score} = \frac{\textbf{2}\times\textbf{Precision}\times\textbf{Recall}}{\textbf{Precision+Recall}} \quad \textbf{.....(7)}$$

3.9 Model Performance Measurement

Non-Parametric Bootstrap internal validation with 1000 iterations was carried out on this fixed test data set for each of the models. This entailed a resampling test used for each iteration of bootstrapping to obtain a new test data set, with replacement of the same number of instances as were in the original test data set it effectively replicating the thought process of selecting a pool of patients slightly different from the original one but still drawn from the same source or distribution. A subset of the pre-computed test predictions was then selected for each of the resampled test data sets. To calculate the following metrics: accuracy, sensitivity or recall in the diseased class, specificity or true negative rate for healthy instances, and area under the Receiver Operating Characteristic Curve. This was carried out for each metric for each of the models over the 1000 bootstrap replications, which led to an empirical distribution of each metric for each of the models (Fornaciari, 2022). The mean of the distributions became the point estimate of performance, with the standard deviation taking the role of a stability measure, indicating the degree to which the metric varied for slight variations in the test data.

Here is the equation of the Mean standard Deviation of Non-Parametric Bootstrap

$$\boldsymbol{SD_{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta_b^*} - \overline{\theta^*}\right)^2} \text{.....(8)}$$

Where,

B Is the total number of Bootstrap resample

θ_b^* is the individual performance metric for each iteration.

$\overline{\theta^*}$ Is the mean

Here is the Bootstrap Mean Formula

$$\boldsymbol{\overline{\theta^*} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta_b^*}} \text{ ..... (9)}$$

Where,

B is the Total number of iteration

$\overline{\theta^*}$ is the Bootstrap Mean

$\widehat{\theta_b^*}$ is the Individual Performance Metric

Non-parametric 95% confidence intervals were then calculated using the 2.5th and 97.5th percentiles of the distributions, with no assumptions about the distributions of the metrics.

Here is the Confidence Interval's Equation

$$\boldsymbol{CI_{1-\alpha} = \left[\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*\right]} \text{ ..... (10)}$$

In this study we have set α = 0.05 to obtain a 95% Confidence Level. Using the percentile method on the 1,000 bootstrap resamples, the interval is determined using the percentile of the 2.5th and 97.5th percentiles. Here is the Equation after this settings.

$$\boldsymbol{CI_{95\%} = [\theta_{0.025}^*, \theta_{0.975}^*]} \text{ ..... (11)}$$

Finally, these bootstrap-based results were then utilized for the comparison and ranking of the seven classifiers, as well as for the production of a series of publication-ready graphical plots. Receiver Operating Characteristic plots were then created for each of the models

Here is the AUC-ROC Curve Equation

$$\mathbf{AUC} = \frac{\sum_{i=1}^{n_p} \sum_{j=1}^{n_n} I(P_i > P_j)}{n_p \cdot n_n} \text{.....(12)}$$

Where,

$\mathbf{n_p}$ Predicted number of positive instances (patients with thyroid disease).

$\mathbf{n_n}$ Predicted number of negative instances (healthy patients).

$\mathbf{P_i}$: predicted probability of a positive instance.

$\mathbf{P_i}$: predicted probability of a negative instance.

I : indicator 1 when positive, 0 when negative.

The confusion matrices is for the best-performing classifiers were produced, and bar charts with error bars were then utilized for illustrating the results for both accuracy and sensitivity, along with their 95% confidence intervals and standard deviations. Additionally, a medical-metrics heat map was then used for summarizing the bootstrapped mean values of sensitivity, specificity, accuracy, and AUC for each of the models developed.

3.10 Implementation Environment

I ran everything through Google Collab using online programming with Python 3.12. and because it's an online platform, there's no requirement of installing multiple instances of the coding execution. This makes it possible for each of us to analysed every move with one step disassembly.

A SMOTE analysis was added; there may even be an application of sorts of probability calibration. As regards the machine learning and data pre-processing tasks undertaken here, the version of the scikit-learn library used was v1.3.0. However, assuming that we were aiming at the boosting trees on top of that as well, we could have used the XGBoost v2.0.0 library instead. On the other hand, the handling of an out-of-balance condition during the balance the data with SMOTE step of our process with the imbalanced-learn v0.11.0 library took care of the imbalance of the classes. As we were working with the numbers behind the scenes during the OLAP step of our process, the numbers were being processed by OLAP. Finally, we used the seaborn v0.12.0 library during our process of preparing the chart.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Overview

This Chapter discusses the experimental outcomes of seven machine learning models that attempted to predict the thyroid disease with routine clinical and laboratory parameters. Experiments were performed on the validated dataset of thyroid with features including 9,172 patient records 13 features for input (cleaned up features removed) A binary target A binary (target):

0 → No thyroid disease

1 → thyroid disease present

The principal problem with this dataset is that it is seriously imbalanced, with only one disease class accounting for about 31.4%. A naive solution to this problem would be a classifier that always predicted “no disease,” which would already be about 98% accurate but would miss all the actual cases of people with thyroid disease. The purpose of this chapter is therefore not only to document accuracy but to assess how well the classifiers can pick out the positive cases. For handling the class imbalance problem, the training dataset is balanced with the help of the SMOTE technique, whereas the test dataset is left unchanged. The next step is the construction of seven models, then model performance was evaluated using a 1000 iteration Non-Parametric bootstrap internal validation on test sets (Fornaciari, 2022) (Martin, 2023). Here, the test set was resampled with replacement to achieve the mean performance and 95% confidence intervals for all metrics.

4.1.1 Multi-Metric Performance Analysis Using Non-Parametric Bootstrap Internal Validation

The accuracy only gives a snapshot of a model's performance rather than showing a model's decision-making process, especially in a medical setting where a false negative can be much worse than a false positive due to a potential undiagnosed illness. Because of this, the performance of the seven machine learning models on the test dataset was further assessed with precision, recall, and F1 scores, with the help of a single combined heatmap.

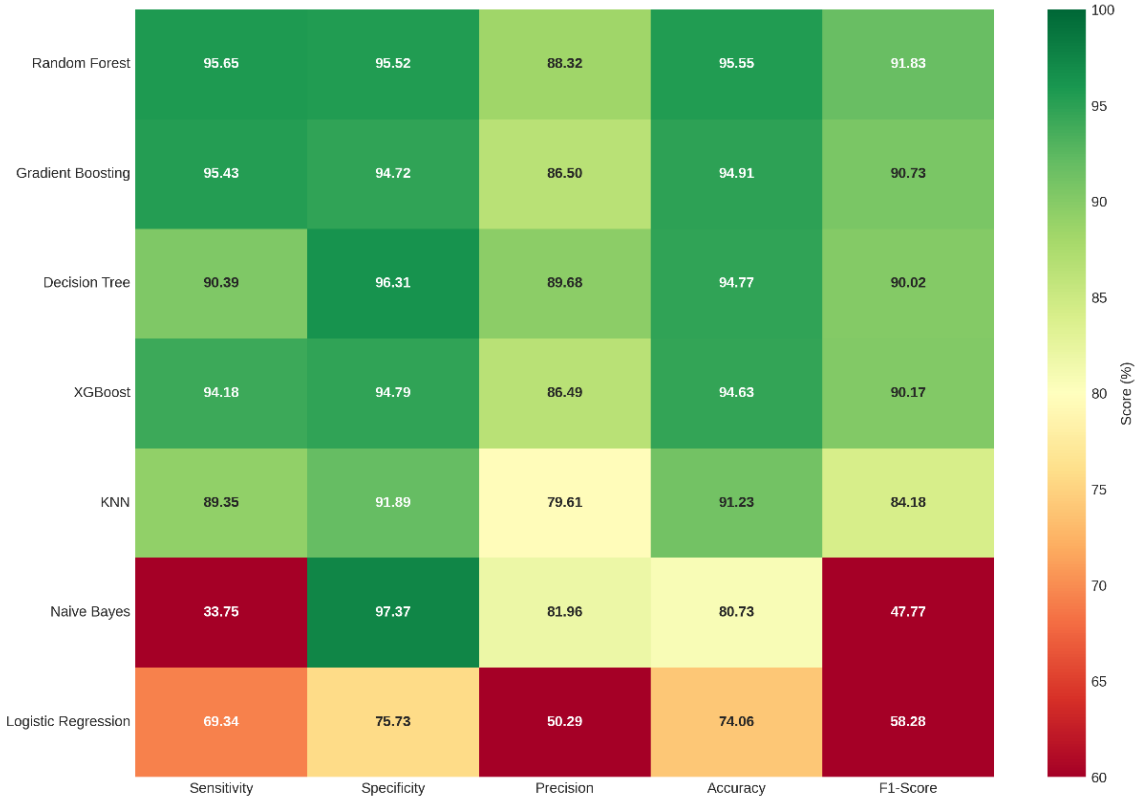


Figure 5 Heatmap of Non-Parametric Bootstrap Internal Validation

The performances on the test dataset were evaluated through the 1,000 iteration test set Non-Parametric bootstrap internal validation approach. The performances were measured in terms of the mean. The best results were obtained by the Random Forest method, followed by high sensitivities of 95.65%; high specificities of 95.52%; high values for precision of 88.32%; high values for accuracy of 95.55%; and high F1 scores of 91.83% for the Gradient Boosting sensitivities is 95.43%; specificities of 94.74%; for

precision of 86.50%; high values for accuracy of 94.91%; and high F1 scores of 90.73%, XGBoost sensitivities is 94.18%; specificities of 94.74%; for precision of 86.49%; high values for accuracy of 94.63%; and high F1 scores of 90.17%, and Decision Tree algorithm sensitivities is 90.39%; specificities of 96.31%; for precision of 89.68%; high values for accuracy of 94.77%; and F1 scores of 90.02% this model performed slightly lower. The remaining classifiers, the Decision Tree and KNN, resulted in moderate performances, while the Logistic Regression and Naive Bayes classifiers resulted in poor performances, where Logistic Regression resulted very low.

4.1.2 ROC-AUC Curve Analysis

We use ROC-AUC curve to learn the model's performance at all performance settings of the model. Here is result of ROC-AUC Curve.

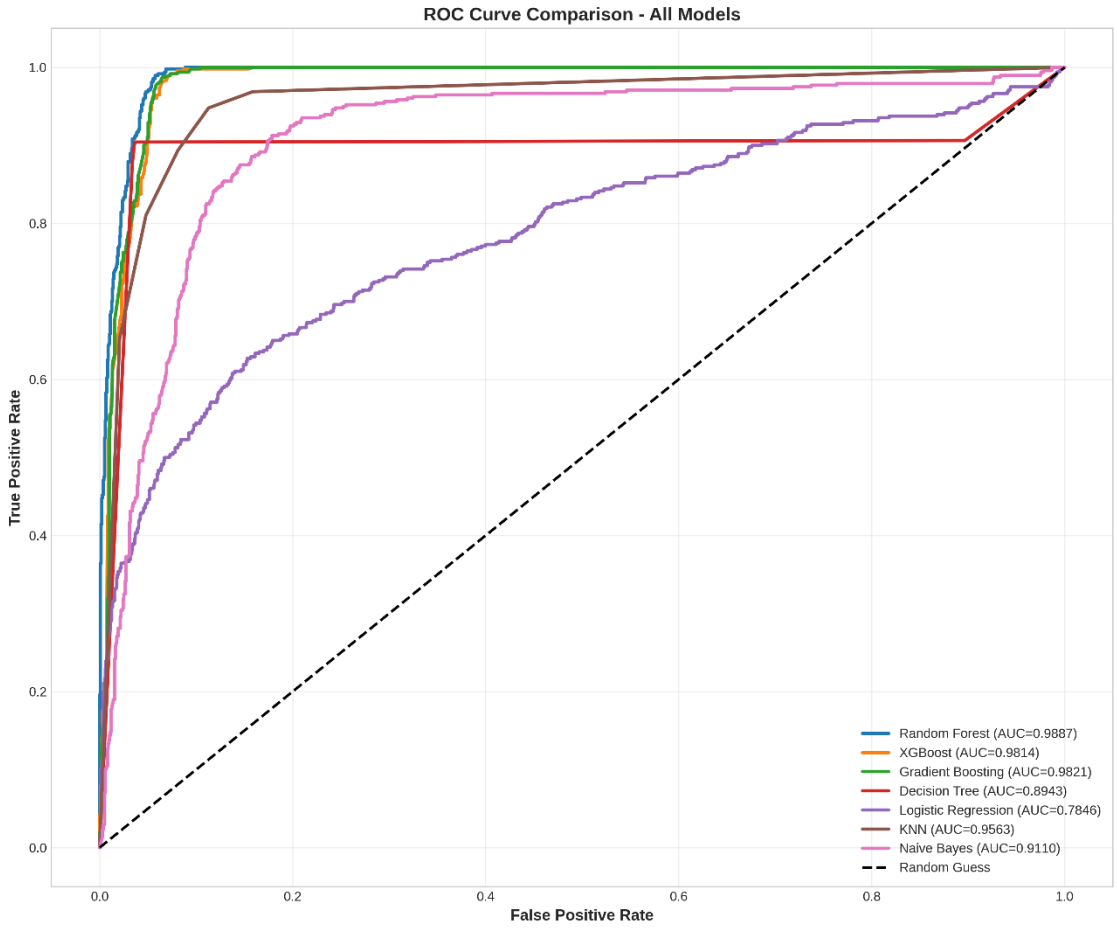


Figure 6 ROC-AUC Curve Analysis

The use of tree models was also justified by the further results of ROC curves as compared to other models. Random Forest provided a measure of about 0.98 on AUC measures, then XGBoost and Gradient Boosting gave slightly less measures of 0.98 . The ROC curve of these models had the top right corner of the graph just a bit close but the curves of the Logistic Regression and Naive bayes were much further. This rendered them weak in their discriminative powers against the other models. To confirm the random forest as the screening model, the confusion matrix of the random forest model demonstrated that there was a considerable number of both true and false samples that had very low amounts of false negatives and false positives.

4.1.3 Performance Evaluation with 95% CI

In order to better understand the performance of each model beyond just high accuracy, we conducted a model sensitivity summary of 7 model with 95% Confidence Interval of Non-Parametric Bootstrap because In terms of clinical safety, it is imperative that the models offer a high level of sensitivity due to the fact that the missing diagnosis represents false negatives or the missed instances of thyroid disease.

Table 2 PERFORMANCE EVALUATION WITH 95% CI

Model	Mean Sensitivity	Stability (± Std %)	95% CI Range
Random Forest	95.65	± 0.96	[93.65, 97.39]
Gradient Boosting	95.43	± 0.94	[93.51, 97.16]
XGBoost	94.18	± 1.04	[91.97, 96.11]
Decision Tree	90.39	± 1.36	[87.58, 92.89]
KNN	89.35	± 1.38	[86.77, 92.02]
Logistic Regression	69.34	± 2.12	[65.06, 73.42]
Naive Bayes	33.75	± 2.24	[29.20, 37.95]

The Table 2 illustrates the mean bootstrap estimate of sensitivity (or recall for the positive class) and 95% confidence intervals on 1,000 resampling's of the testing dataset

for each of the seven classifiers. Also shown are the standard deviations of the resampling results. The two most sensitive classifiers were those utilizing forests, namely Random Forest and Gradient Boosting, with the highest mean values for sensitivity at 95.65% and 95.43%, and corresponding 95% confidence intervals of (93.65-97.39%) and (93.51-97.16%), and smallest standard deviations of (±0.96%) and (±0.94%), respectively, which objectively indicated a high ability of these classifiers to identify almost all truly diseased subjects in the resampled datasets. XGBoost ranked third with a mean sensitivity of 94.18% and 95% CI (91.97-96.11%) and a comparable standard deviation of (±1.04%), followed by Decision Tree and KNN classifiers with apparently lower mean values for sensitivity at 90.17% and 89.35%, with similar 95% CIs. Notably, logistic regression had a significantly lower mean sensitivity at 69.34% and 95% CI (65.06-73.42%), while Naive Bayes had a mean sensitivity of only 29.20% CI and 37.95% As a bootstrap test reveals, the sensitivity levels for the Random Forest and Gradient Boosting models are far higher and remain beyond 93% levels even within the lower confidence limits of the 95% confidence interval. Thus, it can safely be assumed that there is a low probability associated with the missed diagnosis by these approaches. In contrast, the sensitivity levels are found to be low with large confidence limits, remaining far below the 90% desired level considered acceptable within clinical practice for the other models, indicating unacceptably large false negatives. From a technical perspective, the small confidence intervals and standard deviations produced for the ensemble techniques imply that their estimates for sensitivity are not sensitively 7 reliant on a given beneficial split for training and testing, as a stable result can be derived based on multiple resampling's of the testing dataset. This further underpins the validity of the Random Forest and Gradient Boosting algorithms and their employment as recommended models in this context. In contrast, the large confidence intervals for the less robust models demonstrate that a bootstrap technique for internal validation helps identify possible instability that could be obscured by a single estimate value for sensitivity.

4.1.4 Confusion Matrix Analysis

The final Random Forest classifier achieved strong discriminative performance on the single-run test. Here is the result of the confusion matrix of the random forest. Here is the result of confusion matrix of the Random Forest.

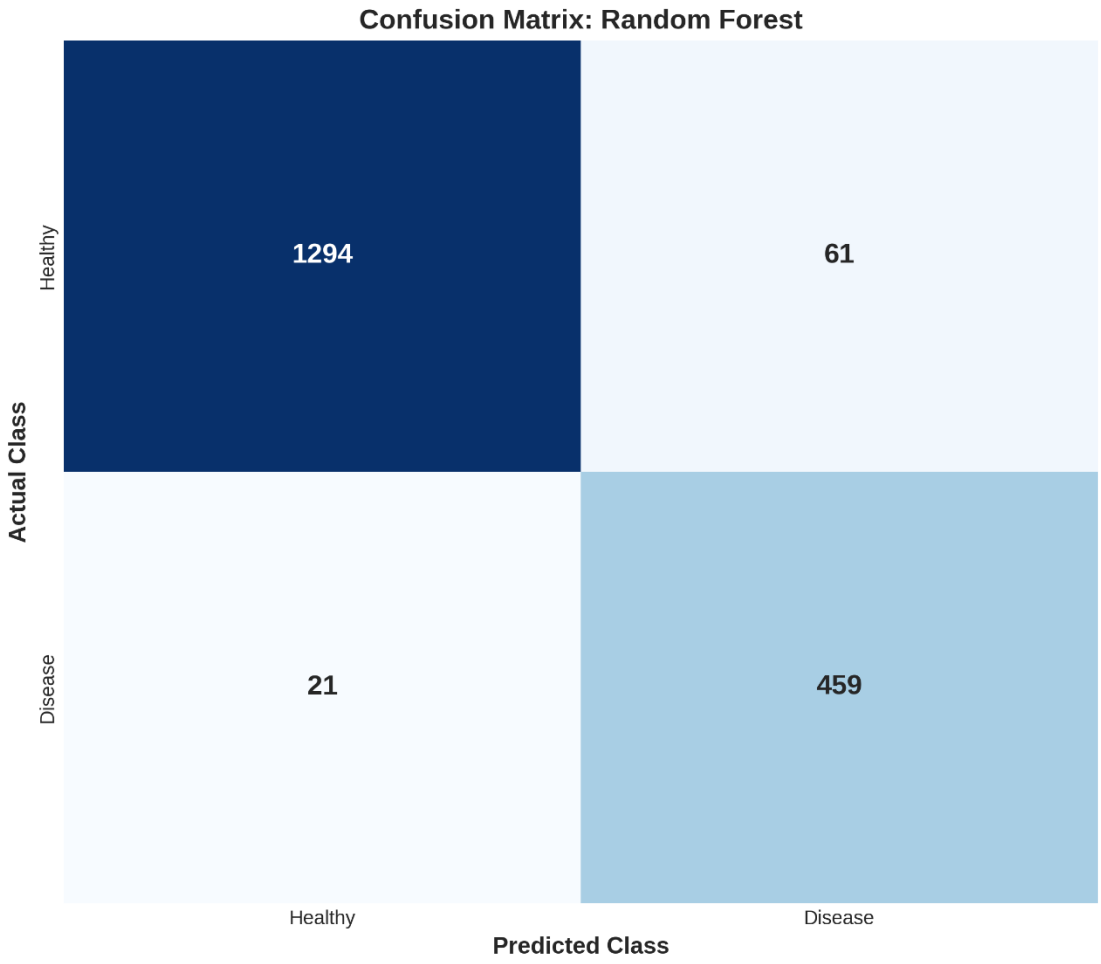


Figure 7 Confusion Matrix of Random Forest

In Figure 7. We calculated additional clinical metrics from this confusion matrix.

The confusion matrix shows the true negatives TN=1294, false positives FP=61, false negatives FN=21, true positives TP=459.

Calculation of the Random Forest's Confusion Metrics Result From Figure 7

$$\textbf{Accuracy} = \frac{\textbf{TP+TN}}{\textbf{TP+TN+FP+FN}} = \frac{\textbf{1753}}{\textbf{1835}} \approx \textbf{95.53\%} \text{ .....(13)}$$

$$\textbf{Sensitivity} = \frac{\textbf{TP}}{\textbf{TP+FN}} = \frac{\textbf{459}}{\textbf{480}} \approx \textbf{95.63\%} \text{ .....(14)}$$

$$\textbf{Specificity} = \frac{\textbf{TN}}{\textbf{TN+FP}} = \frac{\textbf{1294}}{\textbf{1355}} \approx \textbf{95.49\%} \text{ .....(15)}$$

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP+FP}} = \frac{\textbf{459}}{\textbf{520}} \approx \textbf{88.27\%} \text{ .....(16)}$$

$$\textbf{F1score} = \approx \textbf{0.918}(\textbf{91.8\%}) \text{ .....(17)}$$

These metrics shows a highly reliable diagnostic tool that could genuinely assist clinicians in thyroid disease screening. In terms of clinical relevance, very few false negatives (FN = 21) is especially relevant, as these cases represent patients with thyroid dysfunction who will incorrectly be classified as being normal. Sensitivity is 95.53% and an F1 measure of is 91.8% show that there is a good trade-off between predicting a high proportion of patients with thyroid dysfunction and keeping a low rate of false alarms, which is a key for a safe screening or triage procedure in endocrinological care. The high specificity (~95.5%) and precision (~88.3%) values often indicate that a large proportion of patients predicted to belong to the "disease" class actually do, thus avoiding unnecessary follow-up tests for them. From a technical perspective, the near symmetry between the sensitivity and the specificity of the model implies that the threshold value of 0.5 is a well-calibrated operating condition that does not strongly favour either the positive or the negative class. The confusion matrix also helps to validate the confidence intervals determined from the Non-Parametric bootstrap in a patient level, intuitive manner, where the prominent diagonal values of 1294 and 459 strongly contrast with the smaller number of patients in the off-diagonal entries, 61 and 21, in the confusion matrix, thus supporting the robust nature of the RF classifier on the measurement of thyroid disease within the current dataset.

4.1.5 Feature Analysis

It is a significant challenge to understand what features drive the predictions of a given model, especially within a medical setting where both interpretability and relevance are paramount. However, in order to realize this objective, the importance scores yielded by the Random Forest classifier, identified here as the best-performing model, were evaluated. This is because the top top features are captured in Figure 8 (Top 10 Features Random Forest), and some crucial observations can be gleaned from here.

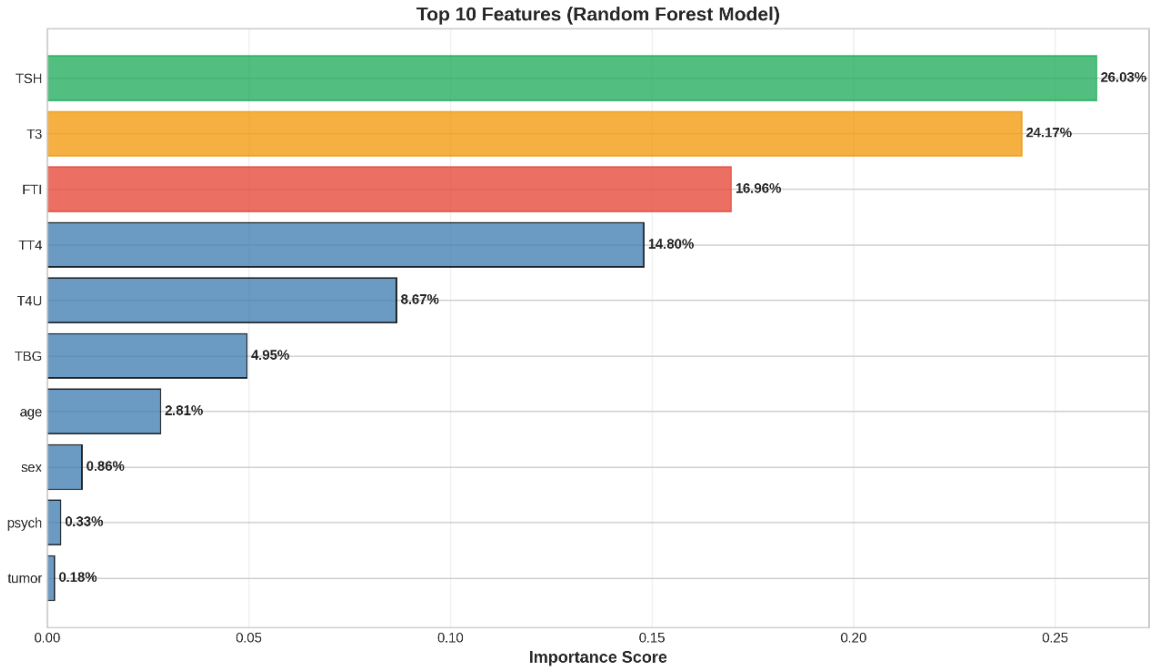


Figure 8 Important Features Random Forest

The figure 8 Strong Influence of Laboratory Measurement Flags The most prominent features were not the hormone levels, but rather the features related to whether a given hormone test had been done. These features include: TSH, tt4, FTI, TSH, T4, T3 these characteristics were given the greatest importance scores within the model. Most Important feature of random forest is THS is 26.03%, T3 is about 24.17% and FTI is 16.96% . This is because psychological violence may be a reaction stemming from a Patients who present with manifestations of thyroid disease will require a series of tests. It is satisfying to see that the 8 model has correctly identified the decision to run some

tests as an indicator of a physician's suspicion and therefore used these features as strong predictors. Moderate Contribution from Actual Hormone Values The next set of 'influential features' included the actual lab values: THS, T3, TT4, FTI. These are standard markers used for medical diagnoses. Though they were smaller in contribution than the 'measured' warnings, it is clear from their importance that the model relies on valid biochemical evidence for better predictions. This is reassuring because it shows that the model is actually taking into account both the metadata and the real physiological signs of the function level of the thyroid (Benabid, 2024).

Minimal Influence of Demographics The Features such as: age, gender, appeared much lower in the ranking. This is a good thing from an ethical and medical stand. Demographic features are not what is influencing this predictive tool. It is a predictive tool grounded in objective medical features, rather than subjective ones.

Very Low Impact of Medical History Flags The Certain features contributed very little, including: age , sex , psych. The fact that they have low importance means that historical/socioeconomic factors were much less enlightening than objective lab tests. From a clinical standpoint, this means that lab tests indicate potential thyroid problems much better than medical history (Martin, 2023)

Overall Result : Feature Importance Analysis gives a few important observations such as The Random Forest model is much in line with reality because in reality, the main diagnostic aid available for a thyroid problem is laboratory testing. It is a reflection of the fact that the most common measurements in volve those related to the flag, because a physician would only order a specific blood chemistry screen if he were concerned about an altered function. The tool relies mostly on lab results rather than an individual's demographics and medical history, placing it alongside other fair and medical models used in diagnostics. As a whole, the analysis of features has reaffirmed the fact that the model is both statistically efficient and capable of learning meaningful patterns in a clinical setting, thus enhancing its feasibility for real-world use in a medical setup.

CHAPTER 5

CONCLUSION

5.1 Conclusion

This The present study aimed at developing and testing a supervised learning framework using a machine learning algorithm that could perform an automated screening test for thyroid disorders using common clinical and laboratory parameters. The optimized pipeline involved robust processing, SMOTE-class re-balancing, and comparative testing using seven traditional and ensemble learning algorithms with Non-Parametric bootstrap internal validation with 1,000 iterations. The Random Forest learning model showed the optimal balance between discrimination and calibration with high values of accuracy (95.55%), sensitivity (95.65%),specificity(95.52%), precision(88.32%), and F1 Score (91.83%) in all the experiments with the smallest buffer or error margin of (± 0.96) of and an acceptable proportion of false negatives.

The results of the analysis indicated that traditional linear and probabilistic models like Logistic Regression and Naive Bayes were unable to leverage non-linear patterns of interaction and relationships in the thyroid data very effectively and resulted in lower sensitivity values. In contrast, tree-based ensemble models like Random Forest and Gradient Boosting were able to make optimal use of these non-linear patterns of interaction and relationships and resulted in improved values of ROC-AUC. The result of feature importance of the Random Forest classifier indicated that the top predictors were interpretable and reflected thyroid psychophysiology.

From a medical point of view, the proposed Random Forest model for the decision support system could play an important role in helping medical practitioners to identify thyroid dysfunction cases in the early stages of development and in conditions where there is limited access to medical practitioners (Riley R. D., 2023). The high F1-score and high sensitivity of the proposed Random Forest model for the decision support system qualify it for initial screening of thyroid conditions among the population, and

should not solely depend on medical practitioners for identification and diagnosis of thyroid conditions among clients for an initial screening program. There are several limitations to this study and approach to identify thyroid conditions among clients.

5.1.1 Future Works

Although the above experiment shows a great potential of machine learning-based models for the prediction of thyroid disease, some significant aspects have remained unexplored in this area, and some related possibilities can be discussed below.

Firstly, it is significant to enhance the accuracy of the minority class, i.e., the cases containing actual instances of thyroid disease. It can be attempted by implementing cost-sensitive learning, hybrid oversampling methods such as SMOTE-Tomek link and ADASYN, and probability thresholds. Secondly, involving a larger dataset of different, recent, and diverse cases would increase the generality of the proposed concept and would be significantly beneficial for enhancing the effectiveness of the proposed approach. Lastly, a significant extension of this concept would be adding a multiclass classifier instead of working on a binary classifier, which would increase the applicability and accuracy of the proposed concept for distinguishing among hypothyroidism, hyperthyroidism, and subclinical and autoimmune cases. Development of a fully functional application related to the proposed concept would be an added advantage, and would increase its functionality and reality. Implementation of some significant techniques related to Explainable AI, such as SHAP and LIME, would be appreciable for enhancing the impact of the proposed concept among physicians and would increase the reality of the proposed concept. Lastly, involving temporal models related to a series of lab tests would increase the potentiality of the proposed concept, and would be appreciable for enhancing the impact of this concept among related authors related to medical domains, including thyroid disease predictions.

REFERENCES

Akter, S. (2024). Analysis and interpretability of machine learning models to classify thyroid disease. *Plos one*.

Alawiyah, T. (2024). The prediction of thyroid cancer recurrence with the XGBoost method: The clinicopathological feature-based approach. *Journal of Computer Networks, Architecture and High Performance Computing*.

Aversano, L. (2021). Thyroid disease treatment prediction with machine learning approaches. *Procedia Computer Science*.

Benabid, M. (2024). Data Pre-Processing Technique for Enhancing Healthcare Data Quality Using Artificial Intelligence. *International Journal of Scientific Research in Science and Technology*.

Chaganti R, R. F. (2022). Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. *Cancers (Basel)*.

Chaubey, G. (2021). Thyroid disease prediction using machine learning approaches. *National Academy Science Letters*.

Chawla, N. V. (2002). *Synthetic minority over-sampling technique.*

Dhyan Chandra Yadav, S. P. (2020). Prediction of thyroid disease using decision tree ensemble method . *Springer*.

Fasira, S. J. (2025). Harnessing deep learning for thyroid disease detection : A Systematic Review .

Fornaciari, T. (2022). *Hard and Soft Evaluation of NLP models with BOOtSTrap Inferential Statistics.* Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

J. Premalatha, R. R. (2024). Comparative Analysis of Machine Learning Algorithms For Thyroid Disease Prediction. *Available at SSRN* .

Ji, S. (2024). The novel self-stack ensemble model for thyroid disease prediction. *Plos one*.

Ji, S. (2024). The novel self-stack ensemble model for thyroid disease prediction. *Plos one*.

Martin. (2023). *Artificial intelligence in thyroidology: a narrative review of current status and future perspectives.* Endocrine Connections.

Obaido, G. (2025). Improving thyroid disorder diagnosis via innovative stacking ensemble learning model. *Digital Health*.

Patel. (2024). Thyroid disease diagnosis based on feature interpolation and machine learning. *Biomedical Engineering Letters*.

Riley, R. D. (2023). *Stability of clinical prediction models developed using statistical or machine learning methods.* Statistics in Medicine.

Riley, R. D. (2024). *Evaluation of clinical prediction models (part 1): from development to external validation.*

Sennan, S. (2022). Thyroid disease prediction using XGBoost algorithms. *Journal of Mobile Multimedia*.

Shiuh, T. (2023). Prediction of thyroid disease using machine learning algorithms with feature selection. *Journal of Theoretical and Engineering Computing*.

Shrestha, K. (2025). Predicting the recurrence of differentiated thyroid cancer using whale optimization algorithm and XGBoost. *Journal of Clinical Medicine*.

Teja, A. H. (2024). Thyroid Disease Prediction Using Machine Learning. *Macaw International Journal of Advanced Research in Computer Science and Engineering*.

221-35-974

ORIGINALITY REPORT

13%	10%	8%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	3%
2	Shankar Babu, Mahesh Babu Kota. "Synergies in Smart and Virtual Systems using computational intelligence", CRC Press, 2025 Publication	<1%
3	bmcmedinformdecismak.biomedcentral.com Internet Source	<1%
4	ijmrset.com Internet Source	<1%
5	Mubin Tamboli, Geeta S. Navale, Priya Shelke, Amol V. Dhumane. "WKOA: Wolverine-Kite Optimization Algorithm for Feature Selection and Deep High-order Attention Network with Explainable AI for Thyroid Disease Detection", Biomedical Materials & Devices, 2025 Publication	<1%
6	Yang Chen, Luchen Zhang, Qi Dong. "Using natural language processing to evaluate local conservation text: A study of 624 documents from 303 sites of the World Heritage Cities Programme", Journal of Cultural Heritage, 2024 Publication	<1%
7	Udara Yedukondalu, V Vijayasri Bolisetty. "Advancing Innovation through AI and Machine Learning Algorithms - Computational	<1%