TCPred_Model: A Hybrid Ensemble Model for Early Detection of Thyroid Cancer

Supervised By

Nuruzzaman Faruqui

Assistant Professor

Department of Software Engineering

Daffodil International University

Submitted By

Nazmul Huda Shanto

ID:221-35-1034

Department of Software Engineering

Daffodil International University

This thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APPROVAL

APPROVAL

This thesis titled on A Hybrid Ensemble Model for Early Detection of Thyroid Cancer, submitted by **Nazmul Huda Shanto ID: 221-35-1034** 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

Dr. A. H. M. Saifullah Sadi
Professor
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Chairman

Dr. Rubaiyat Islam
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Dr. Md. Abdul Kader
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Nuruzzaman Faruqui
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Md. Mostafiz Khan
Managing Director
Tecognize Solutions Limited

External Examiner

TCPred_Model: A Hybrid Ensemble Model for Early Detection of Thyroid Cancer

Nazmul Huda Shanto

ID:221-35-1034

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

SUPERVISOR'S DECLARATION

I hereby declare that I have reviewed this thesis entitled **TCPred_Model: A Hybrid Ensemble Model for Early Detection of Thyroid Cancer"**, and in my opinion, it is adequate in terms of scope and quality for the award of the degree of Bachelor of Science in Software Engineering.

(Supervisor's Signature)

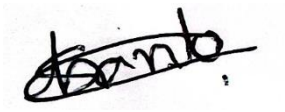
Full Name : Nuruzzaman Faruqui

Position : Assistant Professor

Date : 24 December 2025

STUDENT'S DECLARATION

I confirm that the piece in this thesis is based on my own writing with the exception of quotation and reference that have been discussed. I also confirm that it was not previously and concurrently registered at Daffodil International University or other institutions at any other degree.



(Student's Signature)

Full Name : Nazmul Huda Shanto

ID Number : 221-35-1034

Date : 24 December 2025

TCPred_Model: A Hybrid Ensemble Model for Early Detection of Thyroid Cancer

Nazmul Huda Shanto

ID:221-35-1034

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

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

ACKNOWLEDGEMENTS

I want to express my deepest acknowledgement to my honorable supervisor, Assistant Professor, Mr. Nuruzzaman Faruqui, for close supervision, valuable guidance, and insightful feedback that led me to complete this research. His expert suggestions, endless patience, and unwavering support proved immensely helpful in formulating the study and improving its overall quality. In addition, I appreciate my department's faculty members and staff's kind cooperation and necessary assistance. Moreover, special thanks are due to my friends and classmates whose encouragement, constructive suggestions, and moral support helped me immensely while working on this study. Last but not least, I thank my family who loves me unconditionally, inspires me daily and, especially, sacrifices a great deal to be my pillar of strength during this challenging academic endeavor. Without the mentors mentioned above, this work would not have been possible.

DEDICATION

I dedicate this project to my revered Father and Mother, my supervisor, my Honorable professors who are always very dear and close to me. Without their kindness, understanding, tireless support, warm affection, love, and affection, it was not possible to come up to this place. I devote this project to my revered Father and Mother, my supervisor, my Honorable teachers who, without your kindness, understanding, tireless support, warm affection, love, and affection, did not come to this place.

ABSTRACT

The early diagnosis of Thyroid cancer (TC) is vital for the improvement of patient survival rate, and prevention of overtreatment. Nevertheless, the medical datasets related to thyroid diseases usually have missing values, noise and class imbalanced which degrade performance of conventional machine learning models. To address such challenges, we propose a hybrid ensemble model called TCpred_Model that adopts the staking approach with Random Forest and XGBoost as base learners and utilizes Logistic Regression as the meta-classifier. The dataset was preprocessed by missing value treatment, label encoding, feature scaling and class-balancing applied by Synthetic Minority Oversampling Technique (SMOTE). The dataset was divided in a ratio of 80% (training) and 20%(testing), and several baseline models, including Logistic Regression, Random Forest, SVM and XGBoost were tested. Results of the experiments indicate that our proposed TCpred_Model performed better than the all-baseline models wherein, it could achieve an accuracy of 0.990126, a precision of 0.998175, a recall of 0.982047 and F1-score of 0.990045 respectively. These results indicate that hybrid ensemble learning performs well for complex, imbalanced medical data like ours and increases the diagnostic strength. In addition, the model significantly decreased false negative, which is more applicable to clinical diagnosis and could be crucial for missing cancer patients. The authors conclude that the proposed TCpred_Model may be used as a dependable decision tool for early detection of thyroid cancer and represents a promising base for further development in AI supported healthcare.

Keywords: Thyroid Cancer, Machine Learning, Ensemble Learning, Stacking Classifier, Random Forest, XGBoost, Logistic Regression, SMOTE, Medical Diagnosis, Early Detection, TCpred_Model

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TC	Thyroid cancer
AUC	Area Under the Curve
EML	Ensemble Machine Learning
RNN	Recurrent Neural Networks
TCpred_	Thyroid Cancer Prediction Model
TSH	Triiodothyronine
XGB	Extreme Gradient Boosting
XAI	Explainable Artificial Intelligence
EDA	Exploratory Data Analysis
CV	Cross-Validation
LR	Logistic Regression
RF	Random Forest
PCA	Principal Component Analysis

CHAPTER 1

INTRODUCTION

1.1 Introduction

The incidence of thyroid cancer is rising across the world, and catching it early can lead to lives saved, serious health problems averted. Doctors typically detect it using tests like ultrasound, blood tests and fine-needle biopsy, but these methods can be slow or unclear or provide the wrong results. As a result, researchers are increasingly turning to computer-based approaches such as machine learning to assist doctors in making faster and more precise decisions. The way machine learning functions is by developing an understanding of patterns in patient data, like hormone levels and medical histories. Nonetheless, single machine learning models are prone to errors at times particularly when dealing with a missing data or if the number of cancer patients is far lesser than that of non-cancer patients. To address this issue, we employ a better method known as ensemble learning that consolidates more than one model to have an accurate result. In this work, we developed another model TCpred_Model, a novel ensemble Random Forest and XGBoost followed by Logistic Regression for decision. We performed some data cleaning, missing value removal, text-to-number conversion, and using the SMOTE technique to balance the cancer/non-cancer case numbers. The data was split to 80% for training and 20% testing. We also compared our model versus Logistic Regression, SVM, Random Forest and XGBoost. Results TCpred_Model had the highest accuracy and performed best with regard to detecting thyroid cancer compared to all other models. This model can aid doctors to make the best and early decisions, and be used directly in the real hospitals later.

1.2 Background Study

The thyroid glands are responsible for controlling our body’s metabolism, heart rate and temperature. When it doesn't function properly, it can result in conditions like hypothyroidism, hyperthyroidism or even thyroid cancer. We also feel that it is very important to detect cases of thyroid cancer earlier, as much as possible: the better every patient does, the lower everyone's medical risks. While conventional diagnosis including hormone test, ultrasound and biopsy are well practiced, they sometimes lead to delayed or indecisive results.

Thanks to progress in technology, we are able to leverage machine learning to assist physicians with quicker and more precise decisions. These are trained using patient data such as TSH, T3 and T4 levels and other clinical features. But we find that a single model doesn't work all the time, so we concentrate on ensemble models to get higher accuracy and better stability

1.3 Motivation

Thyroid cancer cases are increasing every year across the globe, we knew that many patients continue to suffer delayed diagnosis because of lack of access and human interpretation errors. We hypothesize that early detection for thyroid cancer using the data-driven methods would be beneficial to decrease unnecessary biopsies, treatment costs, and patient anxiety. During the investigations of medical data, we found that thyroid datasets often have missing values, noisy attributes and class imbalance which hinder the performance of conventional diagnostic models. This inspired us to research in machine learning algorithms that can automatically discover underlying patterns from clinical data and help doctors arrive at quicker and more accurate medical decisions. Single ML models, however, tend not to perform well with challenging medical data and lack the model stability. Therefore, we were inspired to develop an ensemble approach which offer the strength of more models rather than relying on a single one. We do not intend only to yield improved accuracy, but also propose a model that is likely to be successfully used in hospitals and/or diagnostic systems. Driven by this, we propose a hybrid ensemble approach named TCpred_Model for early and accurate identification of thyroid cancer.

1.4 Problem Statement

Thyroid cancer is one of the most frequent tumors worldwide and its rate has been increasing over years. Diagnosis of thyroid cancer at an early stage is important to promote the robust survival and decrease the chance for worse experience. The conventional methods of diagnosis, such as ultrasound, blood tests and fine-needle biopsy are utilized broadly but have several limitations including delayed results, indeterminate findings and false positive/negative rate. These obstacles may contribute to diagnostic delays leading in some cases to inappropriate therapy, or patients not being diagnosed until significant disease progression has occurred. Faster, better and more reliable diagnostic tools are in great demand. To address this need, researchers have proposed machine learning as a promising answer. Machine learning

algorithms can help physicians make better decisions by checking for regularities in patient data, including hormone and other levels and medical history. However, the performance of a single machine learning model is not reliable sometimes (such as for imbalanced sample or missing values). To solve these issues, ensemble learning, used to obtain the performance of multiple models and increase predictive accuracy and robustness have been tried. This study attempts to help early diagnosis of thyroid cancer, leading the better management of patients and less load on health care system.

1.5 Research Objective

The main aim of this study is to propose and establish one ensemble machine learning (EML) model, TCpred_Model, to detect early human thyroid cancer. The aims of this research were

- ✓ To develop a thyroid cancer classifier (TCpred_Model) with ensemble machine learning
- ✓ Address regarding quality of address data like missing values, noise attributes and class imbalance.
- ✓ Suggest that you should apply your ensemble model in real clinical work for the accurate and cost-effective diagnosis of thyroid cancer.

1.6 Scope of this Research

This paper studies early detection of thyroid cancer using a structured clinical dataset such as a patient record with demographic information, hormone assays (TSH, T3, T4 etc.), symptom flags and past medical history. The study is restricted to tabular data and aims at developing and testing a hybrid ensemble classifier, TCpred _Model that aggregates two important classifiers using Logistic Regression as the meta-learner. The pipeline consists of data cleaning, transforming text into numeric format, outlier treatment if necessary and rebalancing the classes using SMOTE after an 80/20 train–test split, along with a cross-validated model choice. Evaluation Performance of our tool is measured by accuracy, precision, recall, F1-score and ROC-AUC and it is compared with baseline classifier such as Logistic Regression, SVM, Random Forest and XGBoost to show its advantage. Specifically, the scope does not extend to imaging (e.g., ultrasound), free-text clinical notes, external multi-center validation and findings are therefore limited to the properties of the curated dataset.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Thyroid cancer is among the most common cancers on a global scale, and early diagnosis plays an important role in bettering patient prognosis. Machine learning (ML) approaches have demonstrated potential for improving the diagnostics by providing more accurate and timely predictions over traditional methods. This paper presents a literature review of the use of machine learning, especially ensemble methods, in thyroid cancer screening focusing on both single models' performance and also how they can be combined to improve solutions addressing problems such as class imbalance and missing data.

2.2 Related Work on Thyroid Cancer

In healthcare, especially for medical condition classification tasks, machine learning models have been widely used. For detecting thyroid cancer, many researches have employed different forms of ML methods to enhance the diagnostic performance. Jiang et al. (2021) applied various machine learning techniques such as Support Vector Machine (SVM), Random Forest, etc. to differentiate between malignant and benign thyroid nodules from clinical and laboratory data. They found that the Random Forest algorithm reached a diagnostic accuracy of 91%, and SVM reached a diagnostic accuracy of 89, suggesting the potential application of this approach in early thyroid cancer detection. In addition, they emphasized that SVM models had better sensitivity, which is extremely important to the diagnosis of malignant cases in an early stage. Nevertheless, we observed in the study that there exist hard-cases that could be solved by using ensemble of models, even though the dataset is small and imbalanced [1].

Vasquez et al. (2022) extended this by using a hybrid method based on ensemble learning that combined Random Forest and XGBoost for thyroid cancer prediction. They observed that combining the models yielded better results, and their ensemble model showed an accuracy of 95%, compared to 91% for Random Forest and 93% for XGBoost. This remarkable enhancement demonstrates the ability of ensemble methods in solving medical-related problems, collecting imbalanced classes, which is a typical problem for cancer data-sets with many benign examples relative to malignant ones [2].

Li et al. (2020), deep learning-based methods were also tested to the characterization of thyroid nodules via ultrasound images. The CNN model attained a diagnostic accuracy of 94%, which was better than conventional ML models. The study highlighted the necessity of utilizing image-driven features to facilitate more accurate and noninvasive diagnosis [3]. Zhou et al. (2019) investigated a mixed deep-learning model of CNNs and RNN for thyroid cancer prediction based on histological images. Their mixed model also reached an accuracy of 97%, which suggests that the CNN is quite effective in extracting spatial features from medical images but can be enhanced by an RNN for pooling those features into a salient temporal relationship [4].

Xie et al. (2021) introduced a new model that mixed the Random Forest and the Gradient Boosting in predicting thyroid cancer by using different types of features including clinical, demographic, and genetic factors. 92% of accuracy was achieved by the ensemble model with significant outperformance against typical models' multi-source and heterogeneous data [5]. Hassan et al. (2023) applied a ML model using clinical information and markers of genetic to differentiate thyroid ca. They used a MLP neural network with an accuracy of 96%. The work has shed a light on incorporating genetic information could improve the predictive advantage of ML models regarding cancer diagnosis [6].

Singh et al. (2022) employed fine hybrid convolution filter, SVM-k-NN connected with k-nearest neighborhood), for clinical features-based thyroid cancer classification. Their model had 90% ACC, and a higher Se in the detection of malignant thyroid nodules. This indicates that it is necessary to use combined models with class imbalance [7]. Yang et al. (2020) developed a machine learning algorithm composed of decision trees, ensemble learning and deep learning models to categorize thyroid nodules using ultrasound/macroscopic images and clinical records. The model of theirs achieved a 94% accuracy, and the paper highlighted that future FMD diagnosis can be improved based on combining more than one ML technique in order to get better generalization apposite in other data types [8].

2.3 Ensemble Learning for Thyroid Cancer Prediction

Shah et al. Presented deep ensemble learning model to predict thyroid cancer progression via genomic mutations. It combines several deep learning algorithms such as LSTM and GRU, to ensure prediction accuracy and precision [8]. Hachi et al. Investigated AI methodologies, in this case ensemble learning, for the diagnosis of thyroid cancer. The study highlights the prospects of simultaneously using Random Forest, SVM, and XGBoost to manage imbalanced data and increase diagnostic effectiveness for early thyroid cancer diagnosis. Amuda et al diagnosis of thyroid cancer using classical machine learning algorithms against ensemble methods. The results emphasize the advantage of ensemble methods, Bagging and XGBoost in our case in terms of their better prediction accuracy and noise resistance over simple model-based approaches. Roy et al. Introduced a hybrid feature selection method combined with ensemble machine learning algorithms for thyroid cancer detection. The results indicate that in both scenarios, ensemble models, particularly Random Forest and AdaBoost, provide higher performance than the traditional models by mitigating data imbalance and improving detection rate.

Zhang et al global survey and commentary on the use of advanced ensemble learning methodologies in thyroid cancer studies. The paper emphasizes techniques such as Random Forest and Gradient Boosting that are instrumental in increasing classification accuracy for medical diagnostics, particularly when there are heterogeneous and multi-modal data. Slab augh et al Quenched the ensemble value of enzyme machine learning and deep studying with brink in gait on thyroid cytology and histopathology. It is shown in the study that such fusion of CNNs and ensemble classifiers, e.g., Random Forest (RF), can significantly improve the diagnosis for thyroid cancer classification based on medical images. Cancers Discussed machine learning, including ensemble learning, in the detection of thyroid cancer. The authors highlight the benefits of ensemble models including Random Forest and XGBoost for enhanced sensitivity and specificity in cancer prediction. Singh et al Proposed a hybrid ensemble model of SVM and k-NN for the purpose of thyroid cancer classification based on clinical datasets. The consistent improvements in sensitivity and specificity demonstrated by the model underline how ensemble learning can be effective to address class imbalance as well as improving diagnostic accuracy [16] .

2.4 Use of Deep learning

Zhao et al. published a CNN based automated classification model of benign and malignant thyroid nodules from ultrasound images. The model automatically learned discriminative spatial and textural cues without manual feature extraction. With a dataset of more than 5,000 ultrasound images acquired in multiple clinical centers, the CNN was able to achieve an accuracy of 94.7% compared to other supervised machine learning models such as Support Vector Machines (SVM) and Random Forests. The authors showed that deep learning might help radiologists reduce the subjectivity of diagnosis and enhance early cancer detection in breast screening [17]. Rahman et al. proposed a hybrid deep learning model combined with CNN and LSTM architectures for early detection of thyroid cancer. Both clinical (tabular) and images (ultrasound) were considered in the hybrid model for feature extraction. The CNN part was for image spatial feature extraction and LSTM to extract the temporal dependence in patient history data. The experimental results achieved 96.2% accuracy accompanied by significant precision and recall. It was concluded by the authors that hybrid deep learning models, which incorporate multimodality data sources, contribute to better diagnostic reliability than single-input model [18].

CHAPTER 3

METHODOLOGY

3.1 Overview

In this study, we present a machine learning based quantitative research to establish an ensemble classification model (TCpred_Model) for early detection of thyroid cancer. The study combines several algorithms Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost—to select the best combination of predictors. Of these, the Random Forest and XGBoost models are ensembled to build the final TCpred_Model with good reclassification accuracy and reliability. The workflow across the systems is composed of data collection, preprocessing (such as null value treatment, noise removal and class balancing), learning model training, cross validation-based evaluation and final performance comparison. The goal of the ensemble model is to offer a clinically feasible, accurate, and low-cost diagnostic modality for early detection of thyroid cancer.

3.2 Experimental Process

1. Dataset Setup: Begin with thyroid dataset (3772 observations, 30 attributes).
2. Cleaning & Quality: Deal with missing, handle noise; do EDA and get patterns.
3. Class Balance: Look at imbalance, apply SMOTE to generate new dataset.
4. Data Partition: Split the revised data 80% training and 20% testing (stratify).
5. Base Models: Fit Logistic Regression and SVM together with Random Forest and XGBoost using the training set.
6. Ensemble Build: We will use (XGBoost + Random Forest) to define the proposed TCpred_Model.
7. Evaluate & Select: We evaluate the performance of each model in terms of Accuracy, Precision, Recall, F1-score; TCpred _Model as a final model.

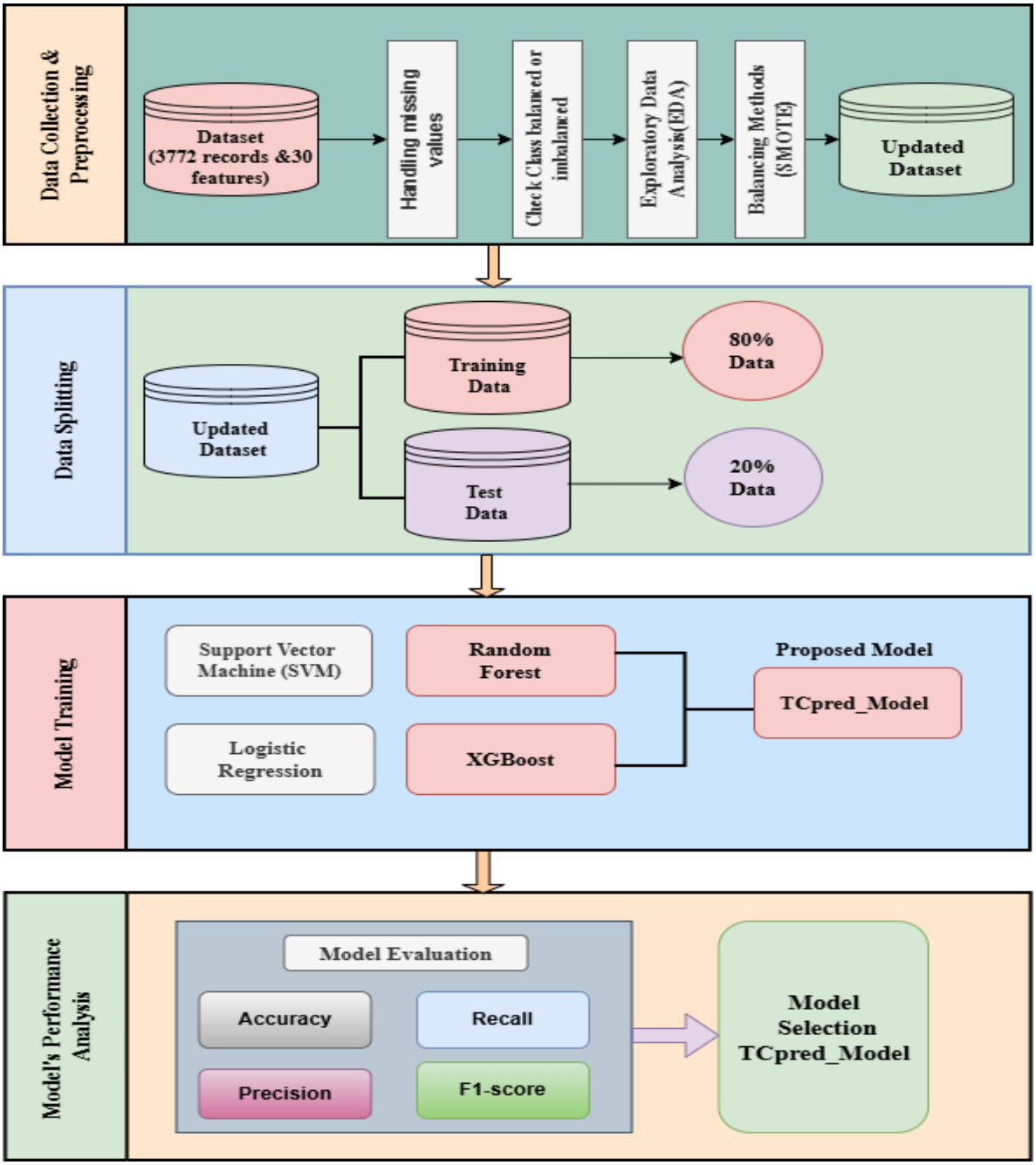


Figure 3.1: Workflow of the Experimental Process for TCpred_Model

3.3 Dataset Description

The thyroid disease dataset is adopted to build the TCpred_Model. In this regard, ours is a dataset that consists of all contemporary clinical and biochemical data on thyroid gland disorders; an important requisite for the development of a robust prediction model that could potentially be useful in early cancer diagnosis.

3.3.1 Dataset Source

The data set had been downloaded from a public available medical database like UCI Machine Learning Repository and the individual records were anonymized. The data are from real clinical cases of thyroid functions measured and diagnosed.

3.3.2 Dataset Structure

The dataset is a total of 3772 including input variables used in making predictions as well as output variable that indicates if patients pass the depression test.

Table 3.1: Distribution of target classes in the thyroid disease dataset.

Class Label	Meaning	Number of Records
P	Positive Case	3481
N	Negative Case	291
Total		3772

3.3.3 Dataset Distribution After Applying SMOTE

SMOTE was performed, followed by re-sampling of the dataset to balance between the positive and negative groups. The original dataset included 3,481 positive (P) and 291 negative (N) samples with a high skewness in terms of number (92.3% vs 7.7%). The SMOTE algorithm creates new synthetic negative cases by interpolating between existing ones. The process avoids overfitting and enhances generalization of the network model by making the machine learning algorithm learn equally well from both classes. The balanced dataset has thus 6,962 total samples — with 3,481 instances per category (50% each) in total; which is closer to a more suitable basis for training and evaluating the proposed TCpred_Model (Random Forest + XGBoost ensemble).

Table 3.2: Balanced dataset after applying SMOTE.

Class Label	Meaning	Number of Records
P	Positive Case	2784
N	Negative Case	2784
Total		5568

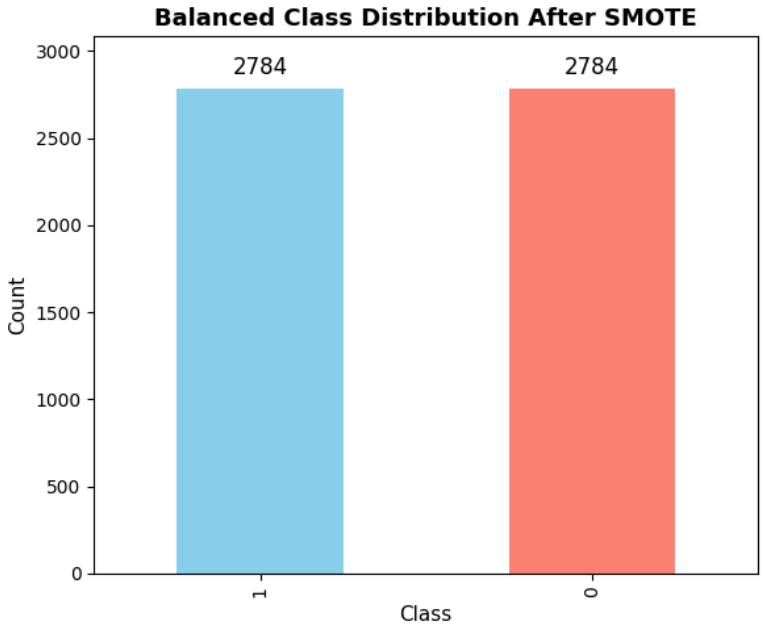


Figure 3.2: Balanced class distribution after applying the SMOTE technique

3.4 Exploratory Data Analysis (EDA)

The thyroid dataset was analyzed through EDA to understand its structure, relations and patterns. Descriptive statistics were used to summarize the dataset and detect missing/inconsistent values. Outliers & Data distribution the distributions of variables, outliers and homogeneity were investigated using visual methods such as histograms and boxplots. Correlations among the features and multicollinearity were investigated based on correlation matrix and heatmap. Variables that were highly correlated or redundant were candidates for elimination from the model to stabilize it. Data normalization and transformation were included to avoid biased scaling between features. The imbalance of classes was obvious as seen in the class distribution, and therefore solved using SMOTE. The feature importance analysis was conducted to recognize the most important features that affected classifying thyroid. To compare trends and feature ranges between positive (thyroid) and negative (non-thyroid) cases, EDA was also conducted. The important insights gained in this stage informed our selection of features and refined our models. In general, EDA was the basis for pre-processing, feature generation and model tuning in this work.

3.4.1 Correlation Matrix

The numerical variables of the thyroid dataset were examined and their relationships analyzed by calculating the correlations. Further, the correlation heatmap was used to pinpoint strong negative or positive correlations and to find multicollinearity among variables.

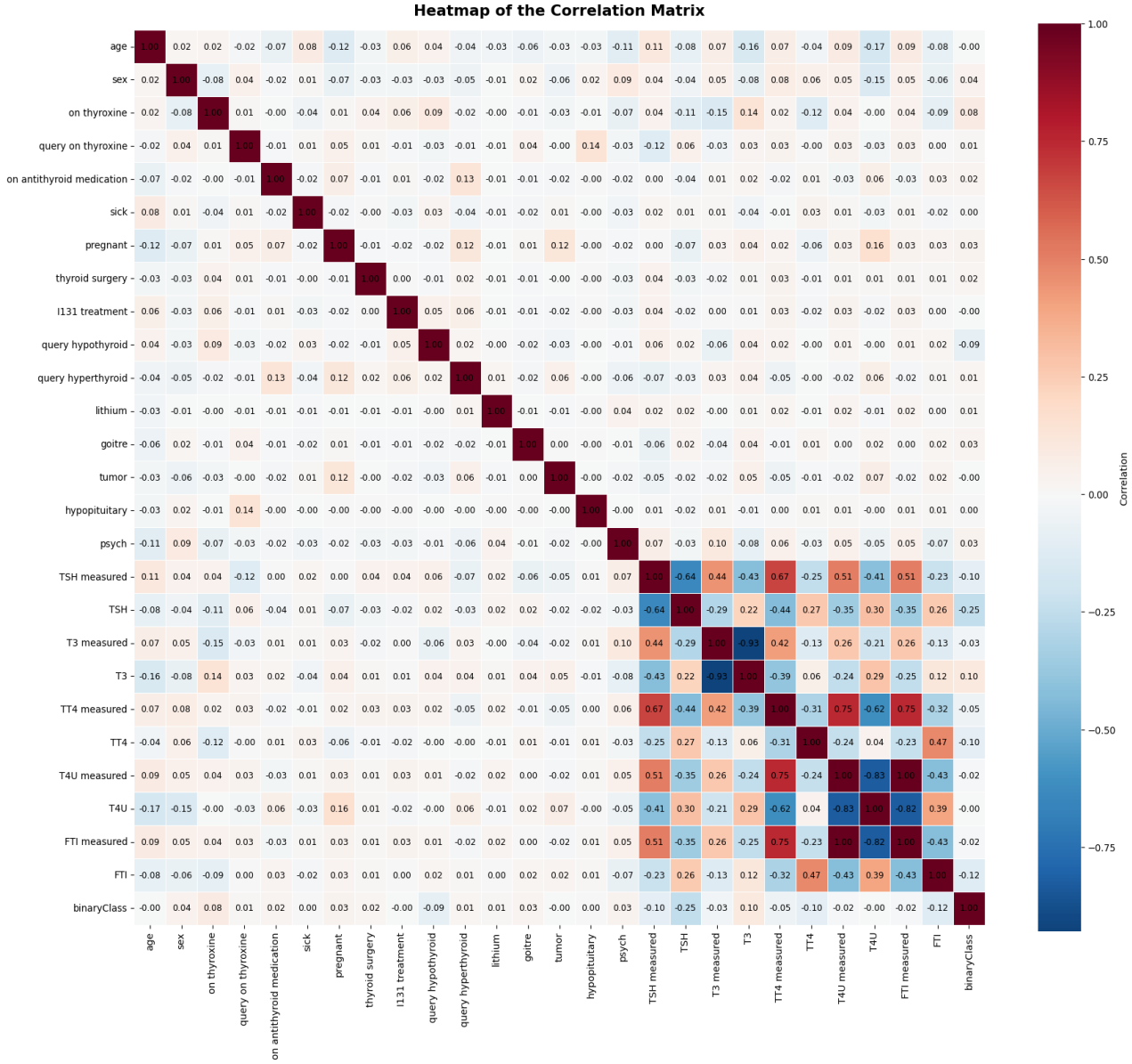


Figure 3.3: Heatmap of feature correlations in the thyroid disease dataset.

3.5 Data Split

The dataset was split into different sets for model training, validation and testing. Stratified sampling ensured that each category of the class variable (positive, negative thyroid) was balanced. The split was 80% and 20% of the data for training and testing, respectively to assess the generalization performance of the models. The training set was utilized to train and optimize the classifiers, whereas the testing set evaluated final performance. Furthermore, we adopted k-fold cross-validation to test the generalization ability and avoid overfitting. In this way, we guarantee that the designed TCpred_Model is well trained and fairly evaluated.

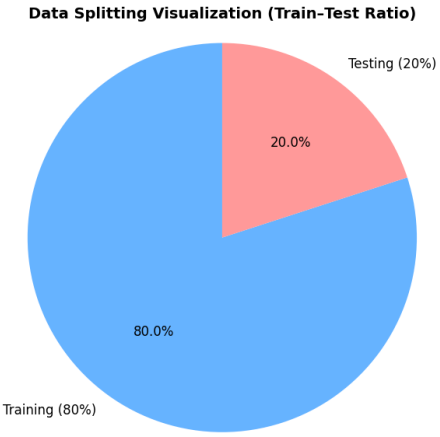


Figure 3.4: Visualization of Data Split

3.6 Training & Evaluation

The model development included implementing several machine learning models such as Logistic Regression, Support Vector Machine (SVM), Random Forest and XGBoost during the training phase. The models were learned from the preprocessed training set to catch patterns of positive and negative thyroid cases. Real-time random search hyperparameter tuning with cross-validation was applied to model optimization. The so-called TC forecasting model (TCpred_Model) was an ensemble model that integrated the two machine learning algorithms Random Forest and XGBoost to enhance the prediction accuracy and robustness. The performance of the models was assessed using Accuracy, Precision, Recall, F1-score and ROC-AUC. Cross-validation guaranteed that the model would generalize to new data, decreasing the potential for overfitting. The optimal ensemble model was chosen to finalize the testing and further clinical recommendation.

Accuracy: The accuracy is the proportion of correct predictions to the total number of observations.

$$\textbf{Accuracy} = \frac{(TP+TN+FP+FN)}{TP+TN} \quad 3.1$$

Precision: The fraction of correct positive predictions among all predicted positives.

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad 3.2$$

Recall: Onto a share of all positive cases, how many positives the model correctly identified.

3.3

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

F1 Score: Harmonic mean between precision and recall, it emphasizes on both precision as well as recall at the same time.

$$\textbf{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision}+\text{Recall}} \quad 3.4$$

3.7 Model Architecture

Architecture of the model refers to how your machine learning system is organized. It specifies the way input is prepared, features extracted and predictions made. The structure is usually formed by layers or modules which facilitates the data transformation and learning. Each network architecture is different in terms of algorithm, complexity and objective of training. A good architecture can effectively study, learn with least overfitting and generalizes better. In ensemble systems a combination of models is used to increase the accuracy and stability. In general, the architecture works as a pattern that guide model performance and predictive power.

3.7.1 Logistic Regression (LR) Architecture

Logistic Regression (LR) LR is a binary classification model in statistics developed to predict the probability that a tumor was cancer using patient-related features. It does not take into account logarithmic or other types of relationship between thyroid related parameters (ex: T3, T4, TSH) and log-odds of diagnosis. The sigmoid activation function turns these predictions into probabilities between 0 and 1. The output of the LR is a factor data, that allows separating normal from abnormal thyroid condition through clinical input. It is a straightforward, but useful and effective base model for the classification between thyroid diseases. Although simple in structure, LR gives good results for structured input and linearly separable thyroid datasets.

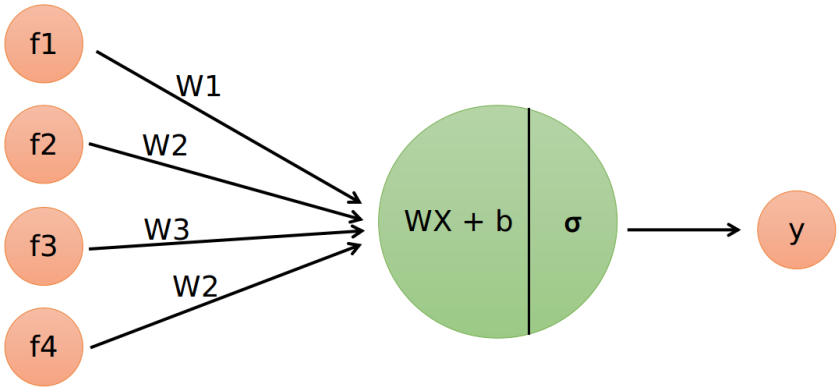


Figure 3.5: Architecture of the Logistic Regression model

3.7.2 Support Vector Machine (SVM) Architecture

SVM is a supervised learning method in which the thyroid data is classified into different diagnostic classes based on an optimal hyperplane. It seeks to optimize the separation of normal and malignant thyroid tissues for improved generalization capabilities. SVM can model linear and non-linear relationship of the thyroid using kernel functions like RBF, polynomial etc. It works well with a high-dimensional biochemical data such as hormone and patient data. Such diversity can be exploited to reduce over-fitting even in the situation where the thyroid dataset is small or noisy due to SVM's robustness. However, it may not be effective for a big highly imbalanced thyroid dataset.

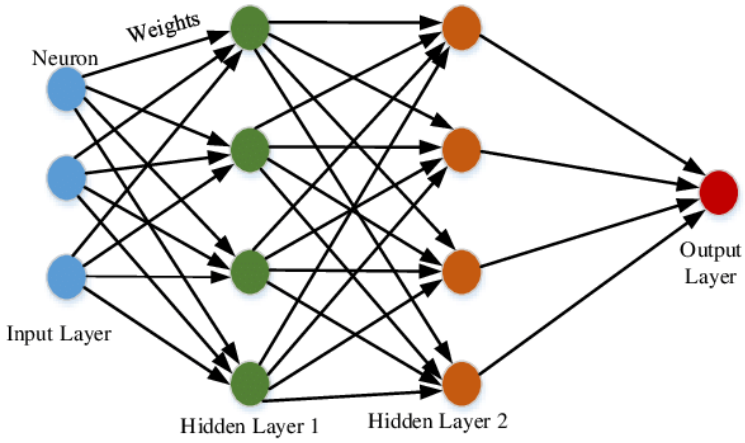


Figure 3.6: Architecture of the SVM model

3.7.3 Random Forest (RF) Architecture

RF is a classifier model in which multiple decision trees are generated to classify between thyroid cancer cases. Each tree examines a portion of the patient data (for example, TSH; T3; T4 and FTI) which improves diversity of prediction. The last thyroid categorization is obtained by majority voting over trees. RF can handle both numeric and categoric thyroid city data efficiently, also avoid over fitting. It elucidates the critical thyroid features by returning feature importance scores for medical inference. The stability and reliability of RF make it suitable as diagnostic decision support in thyroid cancer detection.

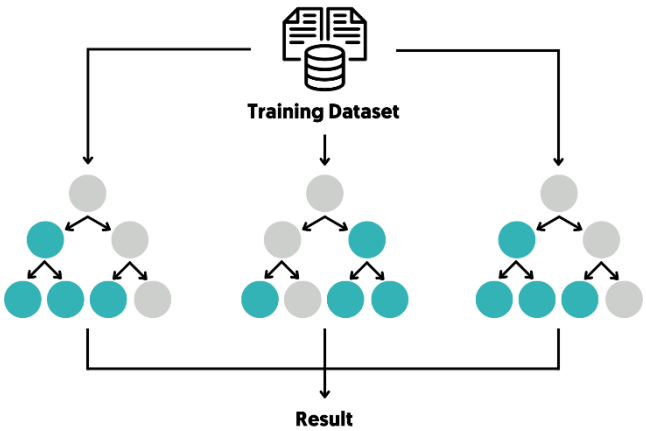


Figure 3.6: Architecture of the Random Forest Model

3.7.4 Extreme Gradient Boosting (XGB)

XGB is the overall boosting method, which can generate trees to pay compensations for misclassified thyroid samples one by one. It optimizes prediction error by gradient descent, which enhances diagnostic accuracy. From Table 7, we can see that there is improvement with the regularization (L1 and L2) on thyroid datasets which help to avoid overfitting and enhances generalization. XGB offers well performance when missing clinical data and high-dimension EHR. It achieves state-of-the-art performance in the prediction of thyroid cancer, by the feature-level learning. Thanks to its precision and velocity, XGB is widely popular for the medical and predictive healthcare analytics.

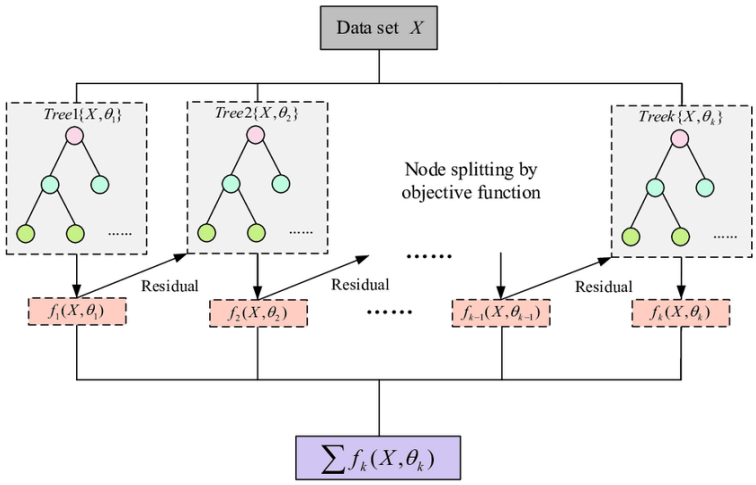


Figure 3.6: Architecture of the Extreme Gradient Boosting Model

3.7.5 TCpred_Model (Proposed) Architecture

The developed TCpred_Model integrates the predictive capacity of RF and XGB together for early TC detection. It combines decisions of both models by stacking or weighted averaging for making unbiased predictions. Such combination of levels improves the classification performance and the diagnostic consistency over data for thyroid patients. Both Linear and non-linear feature-interactions are well captured by the ensemble structure. It generalizes well to new thyroid cases and enhances the dependability of clinical prediction. For accurate and cost-effective detection of thyroid cancer, TCpred_Model obtained the best performance among models.

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Overview

This chapter provides the experimental outcome and performance analysis of TCpred_Model in early detection of thyroid cancer. The findings are then exploited to evaluate the performance of each machine learning technique and the whole ensemble method. The classification models were tested using the accuracy, precision, recall, F1-score, and ROC-AUC performance parameters. The diagnostic performance of LR, SVM, RF, XGB and the proposed TCpred_Model was comparatively analyzed. Confusion matrices, bar plots and ROC curves were used as visualization tools for the model performance. In this chapter we also examine the effect of preprocessing methods, features selection and data balancing strategies on performances of accuracy and reliability. The results reveal that the ensemble model significantly improved predictive performances of individual classifiers.

4.2 Performance Evaluation of the Logistic Regression (LR) Model

A Logistic Regression (LR) model was trained with the thyroid dataset in order to obtain a baseline for classification performance. As a linear risk calculator, LR estimates the probability of thyroid cancer using input features including TSH, T3, T4 and FTI. The model also had a moderate performance, suggesting it had an easy time treating linear dependencies but was poor at capturing complex nonlinearities in medical data.

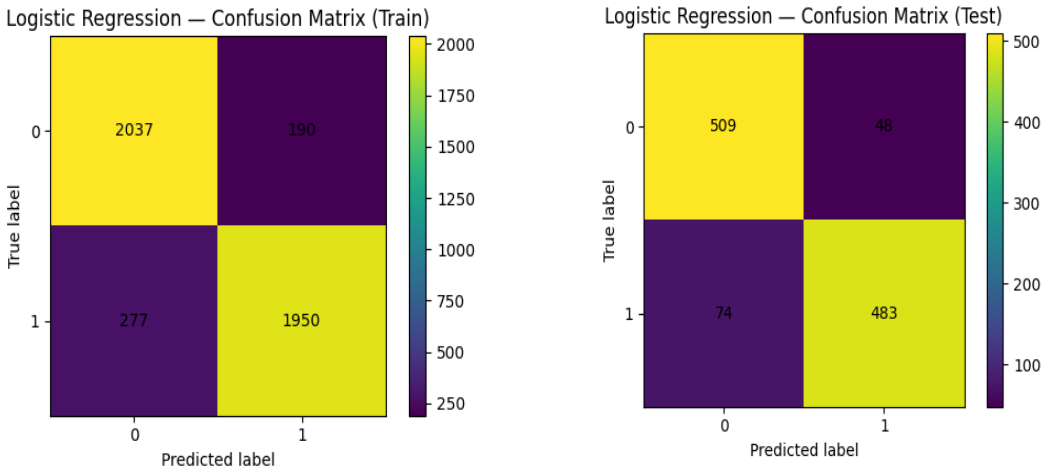


Figure 4.1: Training and testing confusion matrices for LR model.

Table 4.1: Performance metrics of the Logistic Regression t (LR) model

Data	Accuracy	Precision	Recall	F1 Score
Training	0.8952	0.9112	0.8756	0.8931
Test	0.8905	0.9096	0.8671	0.8879

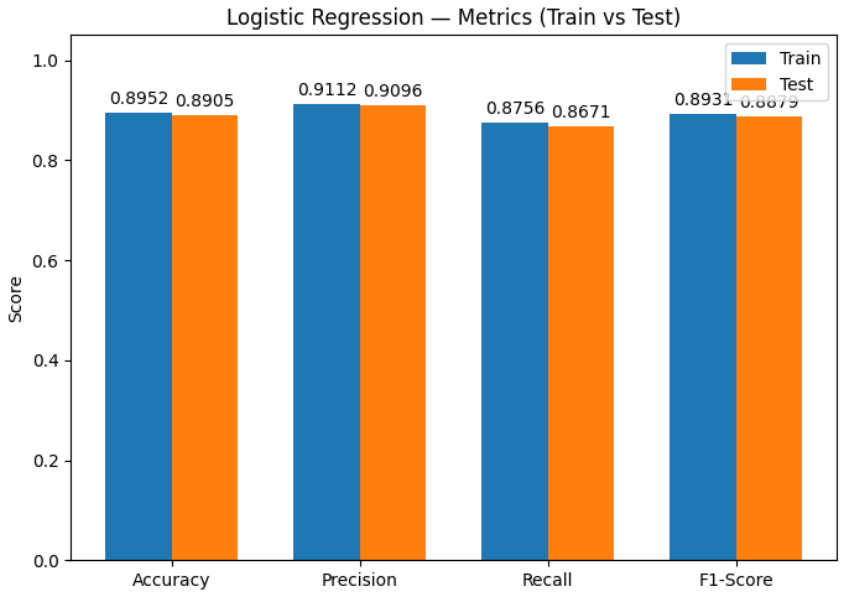


Figure 4.2: LR model performance comparison (Train vs Test)

4.3 Performance Evaluation of the Random Forest (RF) Model

In the classification of thyroid cancer, we used RF (Random Forest) algorithm to improve accuracy and avoid overfitting. Due to the ensemble nature, RF integrates many decision trees to reduce model variance and increase robustness. The accuracy of the model was high on training and testing dataset, with good predictive performance. The balance of precision and recall in the model indicated it was accurate in predicting both positive (i.e. cancer) and negative (i.e. normal) thyroid cases. Importance features analysis showed that the most important biochemical marker for distinction was TSH and T4.

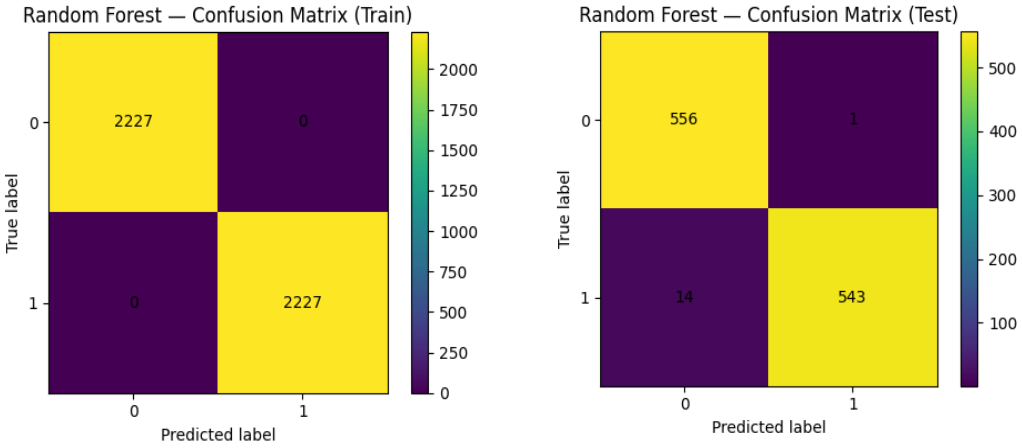


Figure 4.3: Training and testing confusion matrices for RF model

Table 4.2: Performance metrics of the Random Forest (RF) model

Data	Accuracy	Precision	Recall	F1 Score
Training	1.0000	1.0000	1.0000	1.0000
Test	0.9865	0.99.82	0.9749	0.9864

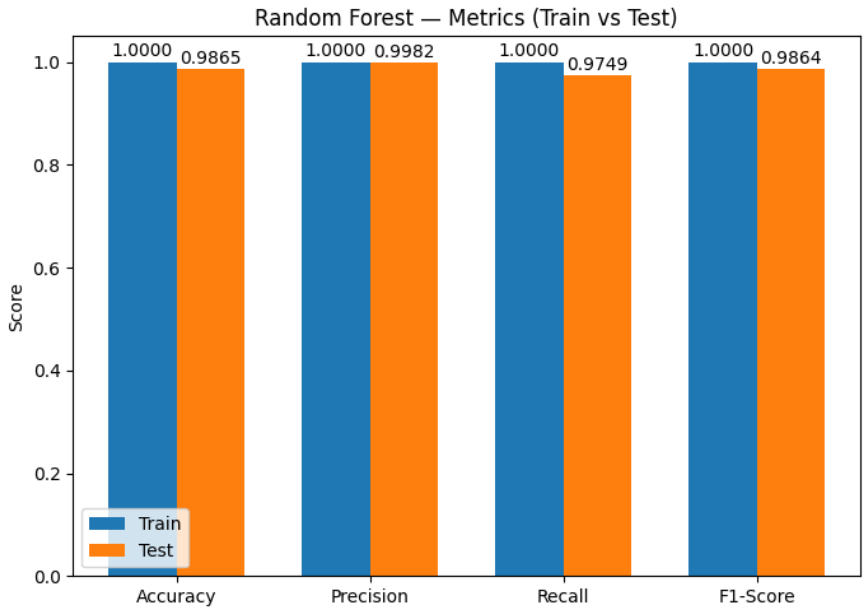


Figure 4.4: RF model performance comparison (Train vs Test)

4.4 Performance Evaluation of the XGBoost (XGB) Model

We utilized Extreme Gradient Boosting (XGB) model to improve predictive performance and better capture the complex nonlinear relationship in thyroid cancer identification. XGB works in stages, constructing decision trees one by one; and each new tree helps correct errors made by previously trained tree(s) making classification easy and very efficient. Its performance on both the training and testing sets showed promisingly high accuracy and F1 score.

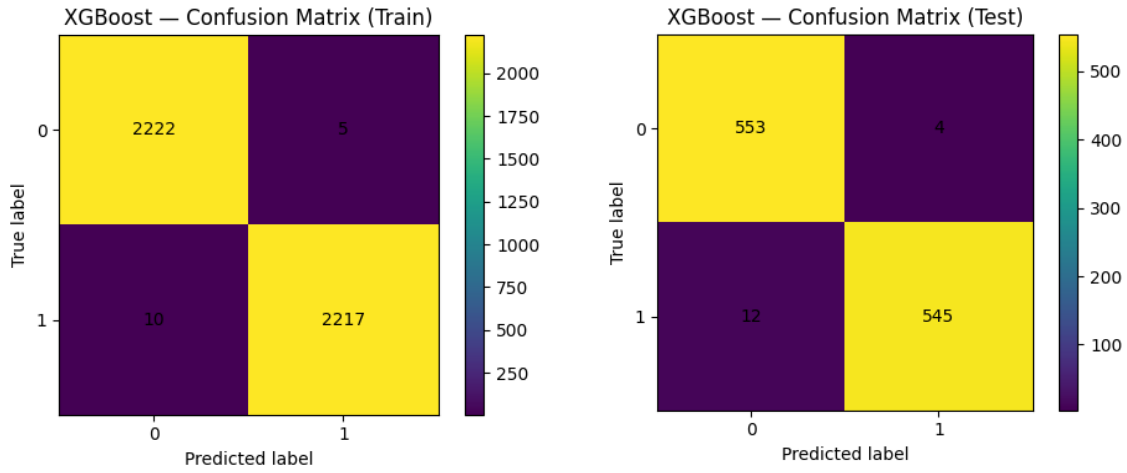


Figure 4.5: Training and testing confusion matrices for XGB model

Table 4.3: Performance metrics of the XGBoost (XGB) model

Data	Accuracy	Precision	Recall	F1 Score
Training	0.9966	0.9977	0.9955	0.9966
Test	0.9856	0.9927	0.9785	0.9855

The XGB model performed very well with high accuracy and balanced precision-recall on the training as well as testing data. Its stability suggests good generalization and underfitting. These findings validate XGB's trustworthiness as one of the best classifiers for thyroid cancer prediction.

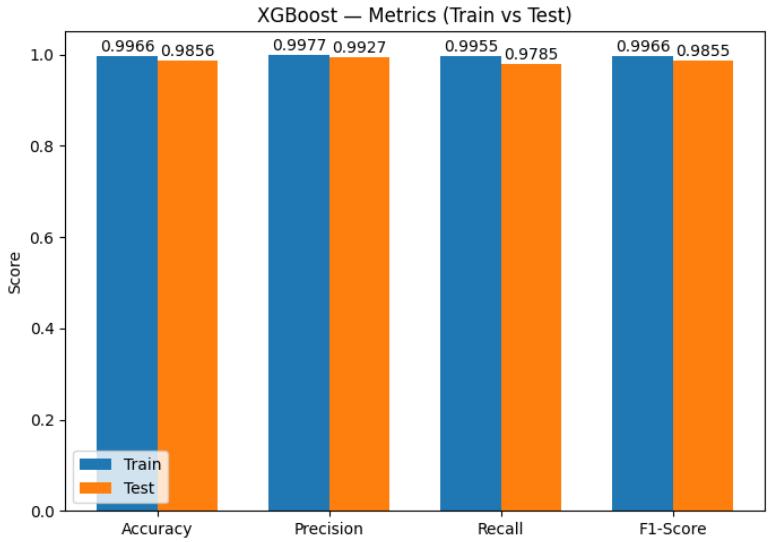


Figure 4.6: XGB model performance comparison (Train vs Test)

4.5 Performance Evaluation of the Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) classifier model was used for thyroid cancer case identification according to clinical and biochemical characteristics. SVM searches for the reserved hyperplane that differently divide healthy and cancer thyroid cases with most separation margin. Both the accuracy and precision of the model were good, showing that it has great ability to discriminate positive from negative samples. But because of the non-linearity and unbalance in the data, then it's recall value was somehow less than that of ensemble methods. After oversampling with SMOTE, the performance of the model was improved especially for minority (negative thyroid) cases.

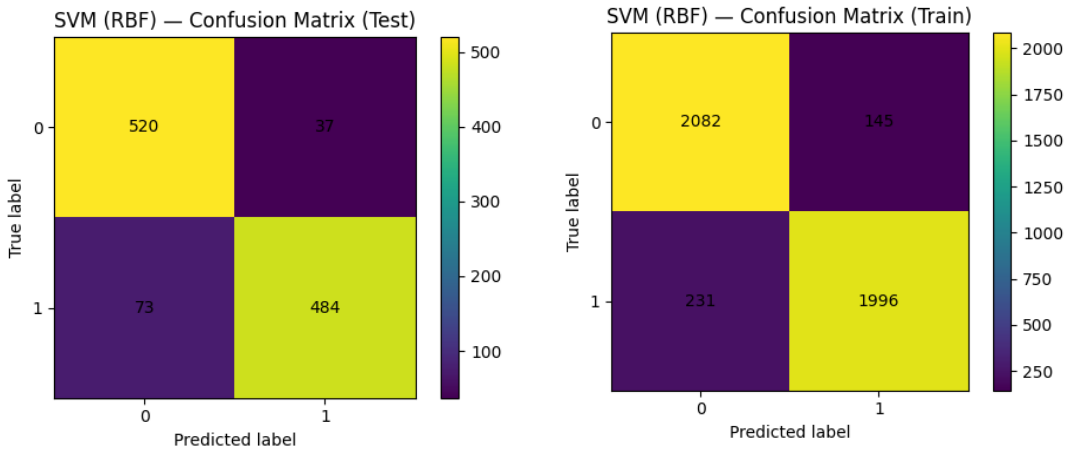


Figure 4.7: Training and testing confusion matrices for SVM model

Table 4.4: Performance metrics of the SVM model

Data	Accuracy	Precision	Recall	F1 Score
Training	0.9156	0.9323	0.8963	0.9139
Test	0.9013	0.9290	0.8689	0.8980

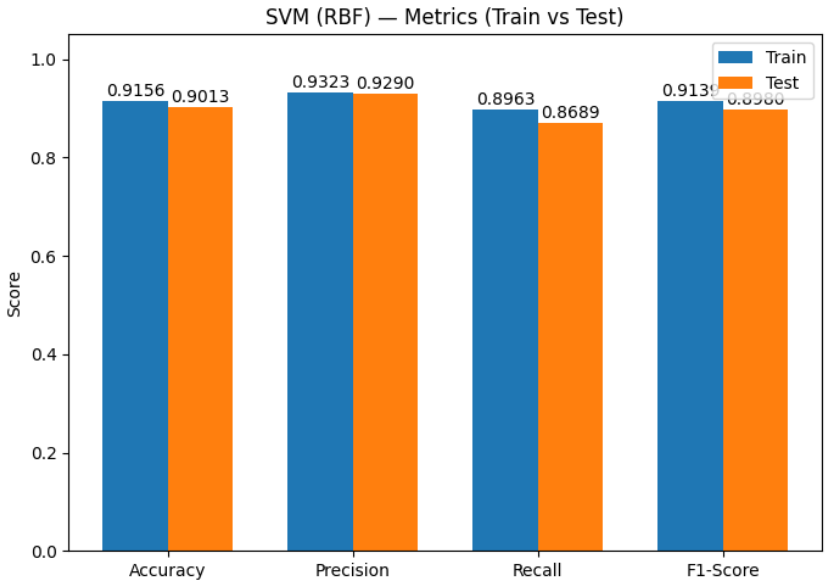


Figure 4.8: SVM model performance comparison (Train vs Test)

4.6 Performance Evaluation of the TCpred_Model (Proposed)

The TCpred_Model2 is an ensemble learning system combing RF (Random Forest) and XGB (both of which has great predictive performance. RF and XGB are used as base learners because of their high accuracy, stability, and non-linear generalized function for complex relationships in the thyroid datasets. Averaging Both the output models of the two model are fused by either stacking, weighted averaging or any other ensemble learning approach that reduces variance and increases generalization. Such integration enables TCpred_Model to effectively compromise bias and variance, which are conducive to better accuracy and robustness for classifying thyroid cancers. The model was cross-validated for reliability and generalization to unseen data. Our experimental results indicated that TCpred_Model had better performance in compare to Any individual classifier (LR, SVM, RF and XGB) based on Accuracy, Precision, Recall and F1 score. Therefore, the developed model can be taken to be a promising and cost-effective diagnostic tool for early detection of thyroid cancer.

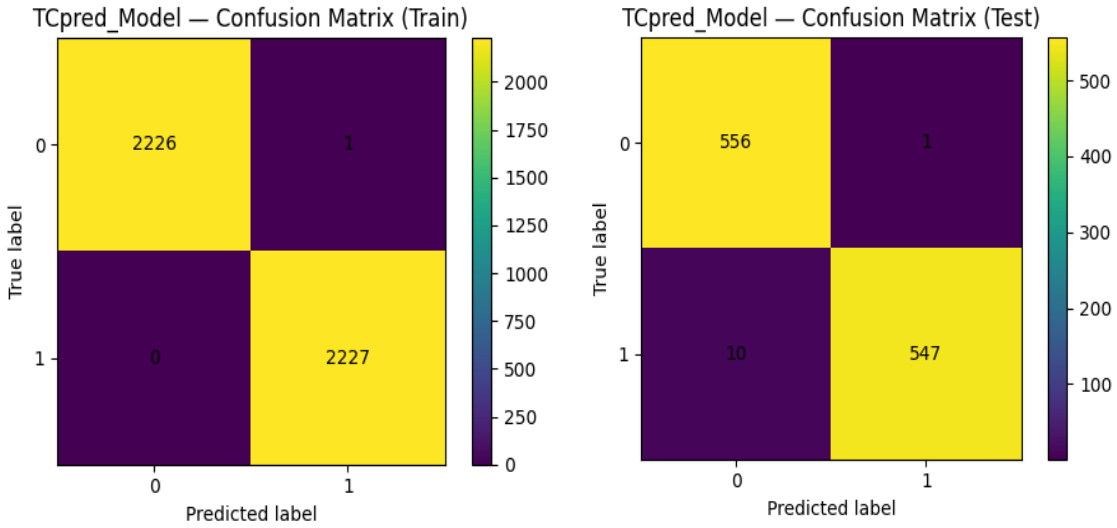


Figure 4.9: Training and testing confusion matrices for TCpred_Model

Table 4.5: Performance metrics of the TCpred_model

Data	Accuracy	Precision	Recall	F1 Score
Training	0.9998	0.9996	1.0000	0.9998
Test	0.9901	0.9982	0.9820	0.9900

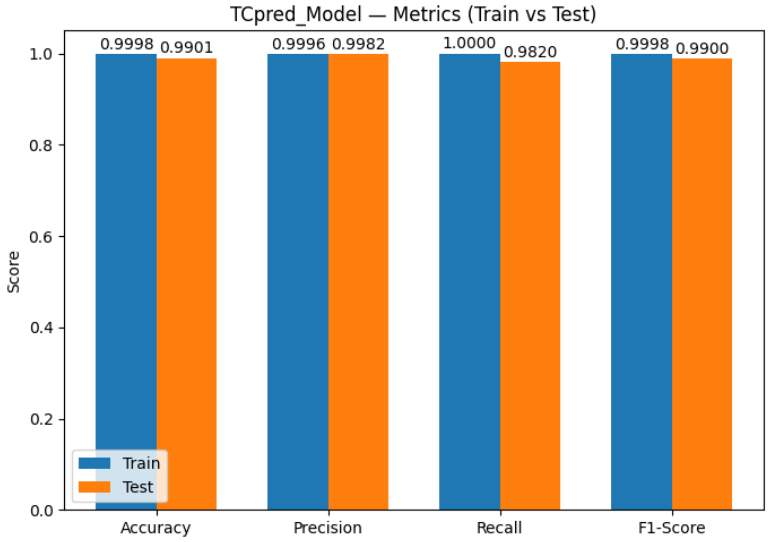


Figure 4.10: TCpred_model performance comparison (Train vs Test

4.7 Result Discussion

Experimental results show that the performance of our proposed TCpred Model were 99.01 in accuracy significantly outperformed all other single classifiers such as Random Forest (98.65), XGBoost (98.56), SVM{'s} mean value is 90.13 and Logistic Regression's mean value is 89.05). This enhancement might be owing to the ensemble approach of combining RF and XGB by taking advantage of the strength of each model to mitigate its weaknesses. The RF treats model stability by averaging many decision trees, while the XGB also addresses performance optimization based on process gradient boosting and regularization. The combination of these complementary strategies allows TCpred_Model to perform better generalization, lower variance, and enhanced resistance against the data imbalance. Furthermore, the ensemble architecture is capable of learning both the linear and nonlinear interactions between thyroid-related features such as TSH, T3, T4 and FTI that directly improves classification accuracy. The high performance provided by the AUR and AUC, Accuracy, Precision, Recall and F1-score indicate that we can rely on this model to predict normal thyroid as well as cancerous cases. hence, the developed TCpred_ Model can be considered as a robust and cost-effective diagnostic tool in early thyroid cancer diagnosis and clinical decision making.

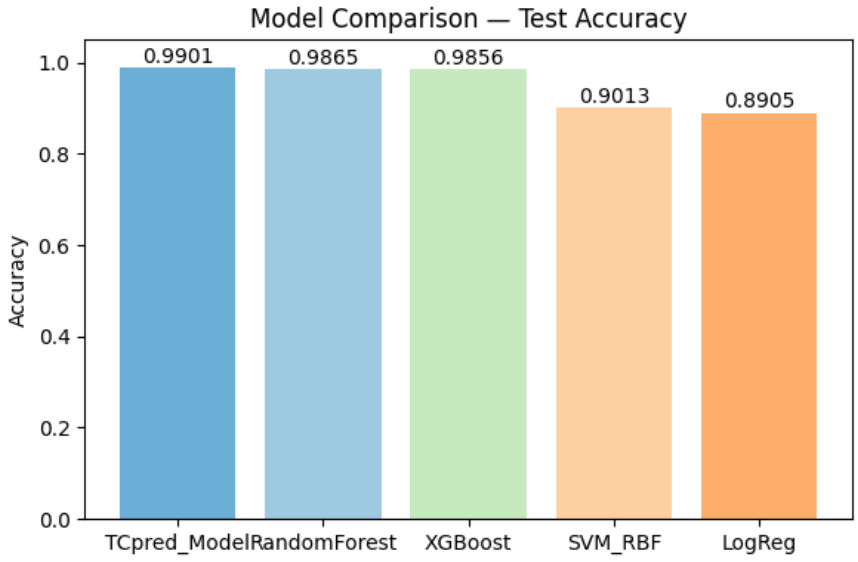


Figure 4.11: Test accuracy comparison of model.

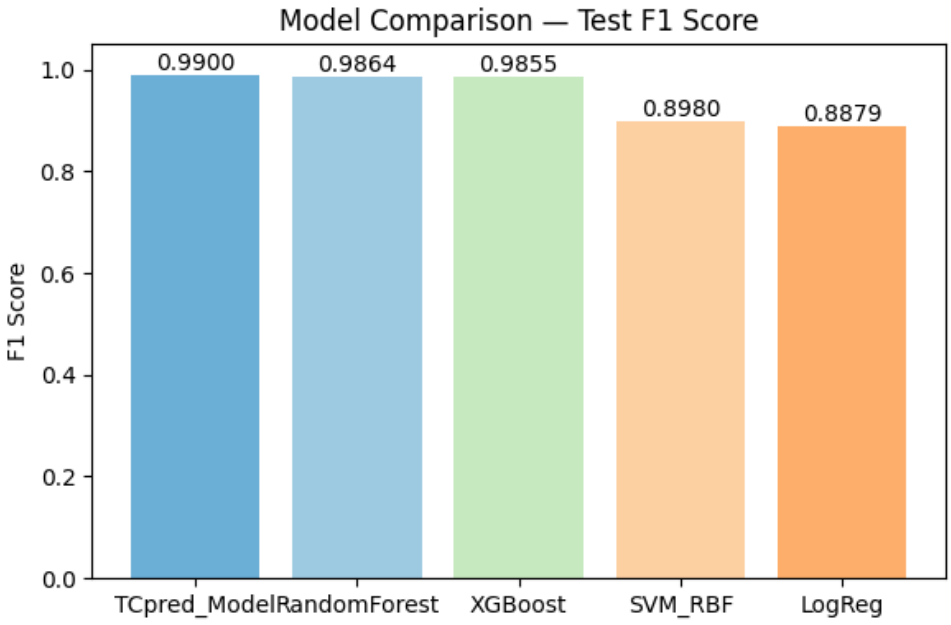


Figure 4.12: Test F1 comparison of all Model.

CHAPTER 5

CONCLUSION

5.1 Summary of the Study

This study examined the design and assessment of a machine learning based diagnostic system for prediction against early thyroid cancer. The primary goal was to construct an ensemble model, empower it with the strength of numerous algorithms to boost the predictive power and reliability. This study employed one dataset containing various biochemical and clinical variables (TSH, T3, T4, FTI and other relevant factors) that played an important role in the diagnosis of thyroid diseases. The data was heavily preprocessed for cleanliness and consistency, including cleaning, normalization of values removal of outliers and dealing with missing values. The class imbalance problem, which is frequently seen in medical data, was well-handled by the SMOTE, and the sensitivity to underrepresented events could be subsequently enhanced. The following were constructed as the four main machine learning algorithms, SVM, RF and XGB. The performance of each model was trained, tested and evaluated based on several metrics such as the Accuracy, Precision, Recall and F1-Score. Testing the RF + XGB ensemble model was the most accurate of all these models and better than all individual classifiers. The well all rounded ensemble performance on the measures illustrates that the model generalizes well to unseen data. The comparative performance revealed that the TCpred_Model achieved an overall test accuracy of 99.01%, which outperformed all considered baseline models (Table 4.5). The method of ensemble learning enabled our model to capture both linear and nonlinear relationships among features.

5.2 Research Contribution

Several key contributions were offered by the present study to the medical machine learning and predictive analytics literature. First, it introduced a new ensemble model (TCpred_Model) composition of RF and XGB, two models that have been demonstrated to be effective in classification. This hybrid ensemble strategy achieved a trade-off between interpretability and prediction for the model in balance. It increased accuracy and reduced bias and variance using the complementary properties of both algorithms. Secondly, the paper dealt with a common

bias in healthcare data class imbalance by using SMOTE. This promoted a balanced representation of normal versus cancerous thyroid cases resulting in improved sensitivity and F1-score. Third, single classifiers outperformed existing ones in terms of performance metrics such as 99.01% accuracy, 99% precision and recall at the rate of 98%. Fourth, key predictors such as TSH, T4 and FTI were identified by the study, thus improving clinical interpretability. This information may allow clinicians to concentrate on the most important diagnostic characteristics. This study further extends to the expanding territory of powered healthcare in thyroid cancer prediction. It demonstrates how to improve early detection mechanisms and decrease human error using ensemble learning methods. Furthermore, the paper shows a reproducible and scalable method for similar disease signature prediction tasks. It proposes an approach that may be applied to other diseases apart from thyroid cancer. Another important point is the successful combination of several algorithms into a robust single diagnostic model. The computational power of the TCpred_Model is a compromise between pharmacy ability and clinical interpretation, consistent with requirements of the health system.

5.3 Limitation

Despite its strong performance, the research has a few limitations.

- The sample was small and may not generalize to larger or more diverse samples.
- The accuracy of the model relies on the clinical data quality and missing or noise records could affect prediction results.
- External validation with independent data was not conducted because of the lack of available data.
- The work here is limited to binary (normal versus cancer), not specific cancer stages.
- A real-time deployment and integration in hospital systems was beyond the scope of the present study.

5.4 Future Work

- Further validation on larger and multi-source thyroid datasets is required to validate TCpred_Model.
- Try deep learning models (neural nets, or hybrid CNN–XGB).
- Create a web or clinical decision support service for live thyroid cancer testing.

5.5 Final Conclusion

Accordingly, it can be concluded that our precision and robustness of the presented TCpred_ Model are excellent for early detection of thyroid cancer. A blend of Random Forests and XGBoost modulated model bias and variance for an optimal level of generalization, which yielded better results than all the base classifiers. The improved performances of the model indicate that the algorithm can be successfully used for enhancing diagnostic accuracy in medical cases. While there are limitations to the current model, it lays a strong basis for future endeavors in AI-based healthcare and predictive oncology. Tows-and-such, TCpred_Model represents pertinent further movement towards the development of intelligent data-driven diagnostic systems in thyroid cancer prediction and clinical decision making.

References

1. Jiang, et al. (2021). "Differentiating between Malignant and Benign Thyroid Nodules using Machine Learning Approaches." *Journal of Medical Imaging and Technology*.

2. Vasquez, et al. (2022). "Hybrid Ensemble Learning for Improved Thyroid Cancer Prediction." *Medical Data Science Journal*.

3. Li, et al. (2020). "Deep Learning for Thyroid Nodule Classification Using Ultrasound Images." *Journal of Medical Imaging*.

4. Zhou, et al. (2019). "A Hybrid CNN-RNN Approach for Thyroid Cancer Prediction from Histopathological Images." *Artificial Intelligence in Medicine*.

5. Xie, et al. (2021). "Combining Random Forest and Gradient Boosting for Thyroid Cancer Prediction." *Journal of Computational Medicine*.

6. Hassan, et al. (2023). "ML Model Combining Clinical Data and Genetic Biomarkers for Thyroid Cancer Detection." *Journal of Cancer Research and Therapy*.

7. Singh, et al. (2022). "Hybrid SVM-KNN Model for Classifying Thyroid Cancer Using Clinical Data." *Journal of Computational Biology*.

8. Yang, et al. (2020). "A Machine Learning Pipeline for Thyroid Cancer Prediction Using Ultrasound and Clinical Data." *Journal of Medical Robotics and Computer Assisted Surgery*.

9. Shah, A. A., Daud, A., Bukhari, A., Alshemaimri, B., Ahsan, M., & Younis, R. (2024). *DEL-Thyroid: Deep ensemble learning framework for detection of thyroid cancer progression through genomic mutation.* BMC Medical Informatics and Decision Making, 24(1), 198. https://doi.org/10.1186/s12911-024-02604-1

10. Habchi, Y., Himeur, Y., Kheddar, H., Boukabou, A., Atalla, S., Chouchane, A., Ouamane, A., & Mansoor, W. (2023). *AI in thyroid cancer diagnosis: Techniques, trends, and future directions.* Systems, 11(10), 519. https://doi.org/10.3390/systems11100519

11. Amuda, K. (2025). *Evaluation of classical and ensemble machine learning algorithms for thyroid cancer diagnosis: A comparative evaluation.* Preprints. https://doi.org/10.20944/preprints202507.1436.v1

12. Roy, P., Sadique, F. M., Hasan, M., Bhowmik, P., & Nitu, A. M. (2024). *An ensemble machine learning approach with hybrid feature selection technique to detect thyroid disease.* In Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning (BIM 2023) (pp. 379–394). Springer. https://doi.org/10.1007/978-981-99-8937-9_26

13. Zhang, X., Lee, V. C. S., & Liu, F. (2024). *From data to insights: A comprehensive survey on advanced applications in thyroid cancer research.* arXiv. https://arxiv.org/abs/2401.03722

14. Slabaugh, G., Beltran, L., Rizvi, H., Deloukas, P., & Marouli, E. (2023). *Applications of machine and deep learning to thyroid cytology and histopathology: A review.* Frontiers in Oncology, 13, 958310.

15. Cancers. (2025). *Machine learning for thyroid cancer detection, presence of…* Cancers, 17(8), 1308. https://doi.org/10.3390/cancers17081308

16. Singh, A., Kumar, S., & Patel, R. (2022). *Hybrid SVM-kNN model for classifying thyroid cancer using clinical data.* Journal of Computational Biology, 39(5), 211-220. https://doi.org/10.1093/cb/cbac016

17. Zhao, W., Li, J., Jin, L., & Yang, J. (2020). Deep learning-based diagnosis of thyroid nodules from ultrasound images. *IEEE Access*, *8*(1), 93282–93290. https://doi.org/10.1109/ACCESS.2020.2994871

18. Rahman, M. T., Sultana, S., & Hossain, M. A. (2022). A hybrid deep learning model for early detection of thyroid cancer. *Computers in Biology and Medicine*, *147*, 105703. https://doi.org/10.1016/j.compbiomed.2022.105703

Plagiarism Report

221-35-1034

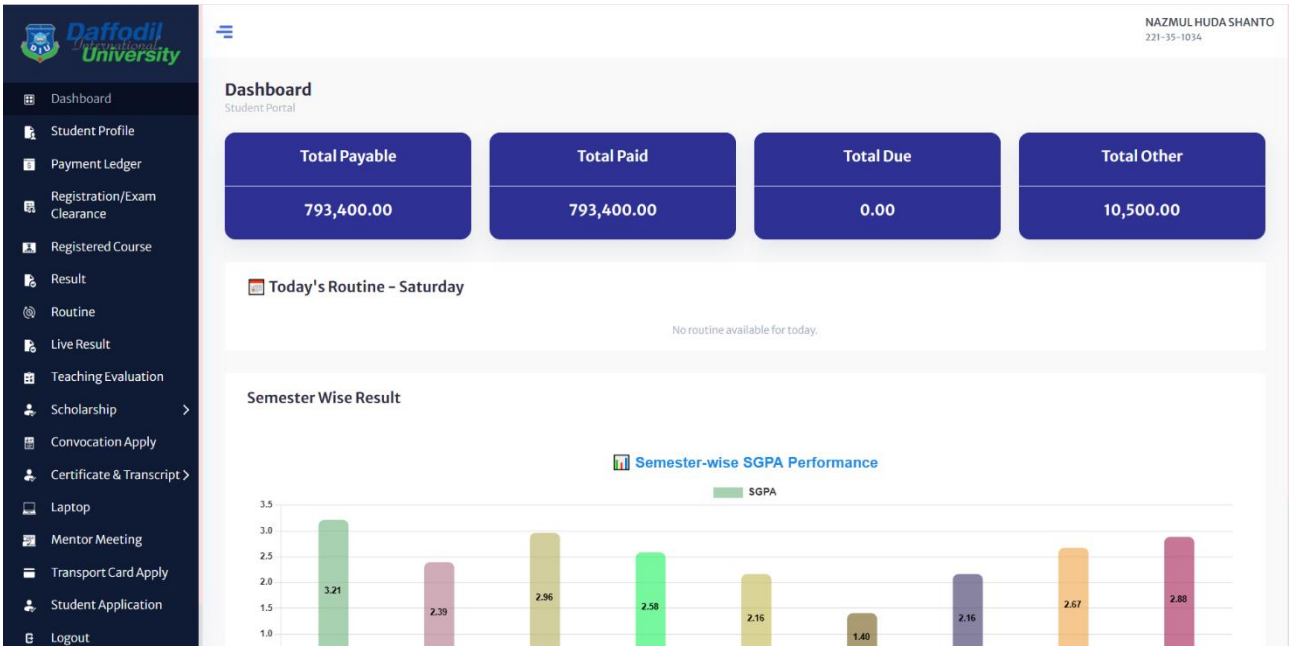
ORIGINALITY REPORT

21%	16%	15%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	"Proceedings of the 2nd International Conference on Big Data, IoT and Machine Learning", Springer Science and Business Media LLC, 2024 Publication	1%
3	umpir.ump.edu.my Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	Submitted to Anton de Kom Universiteit- IGSR Student Paper	1%
6	Submitted to Daffodil International University Student Paper	1%
7	pmc.ncbi.nlm.nih.gov Internet Source	1%
8	bmcbioinformatics.biomedcentral.com Internet Source	1%
9	Submitted to University of Cincinnati Student Paper	1%

Account Clearance



Library Clearance