Received: 16 June 2025 | Revised: 14 August 2025 | Accepted: 24 August 2025 | Published online: 17 October 2025

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2025, Vol. 00(00) 1-17
DOI: 10.47852/bonviewJCCE52026503

BON VIEW PUBLISHING

Optimized Ensemble Machine-Learning-Driven Transparent Thyroid Cancer Prediction Using Explainable Artificial Intelligence

Syed Younus Ali[1], Bilal Shoaib Khan[1], Abdul Hanan Khan[1], Muhammad Adnan Khan[2,3,4], Asghar Ali Shah[5,*], Sagheer Abbas[6] and Khan Muhammad Adnan[7,*]

[1] *Department of Computer Sciences, Green International University, Pakistan*

[2] *School of Computing, Horizon University College, United Arab Emirates*

[3] *Applied Science Research Center, Applied Science Private University, Jordan*

[4] *Chitkara University Institute of Engineering and Technology, Chitkara University, India*

[5] *Department of Computer Science, Kateb University, Afghanistan*

[6] *Department of Computer Science, Prince Mohammad Bin Fahd University, Saudi Arabia*

[7] *Department of Software, Gachon University, Republic of Korea*

Abstract: Thyroid cancer disease diagnosis is a critical medical challenge, requiring accurate and reliable predictions to support clinical decision-making. Patients may suffer from an incomplete diagnosis while using traditional machine learning models. In this study, we present an optimized ensemble machine-learning framework for predicting thyroid cancer disease. The methodology integrates multiple classifiers, including support vector machine, random forest, Naïve Bayes, K-nearest neighbors, and decision tree. The final classification decision is determined with the help of soft voting by a total predictive support vector machine, which selects the classifier with the highest confidence score among the ensemble models. The ensemble strategy enhances predictive accuracy and robustness by combining the strengths of individual classifiers. The model was trained and evaluated, achieving an impressive accuracy of 0.9633 and an area under the receiver operating characteristic curve of 0.9914. The proposed method of this study is very accurate, but there is still a black box problem. To overcome this issue and to ensure interpretability, Explainable Artificial Intelligence techniques, including Shapley Additive Explanations and Local Interpretable Model-agnostic Explanations, are implemented, providing insights into feature contributions towards the performance of the proposed method and model decisions. The dataset contains a total of 30 features with 3,772 different cases consisting of two classes.

Keywords: thyroid cancer prediction, Explainable Artificial Intelligence (XAI), ensemble learning, soft-voting, machine learning, computational intelligence

1. Introduction

Thyroid cancer consists of various disorders affecting the functioning of the thyroid gland, a small gland in the neck, similar to a butterfly [1]. It is responsible for regulating many functions such as metabolism, energy production, growth, and development throughout the body. The two most common forms of thyroid disorder are hypothyroidism, wherein the thyroid is underactive and fails to produce sufficient thyroid hormones, and hyperthyroidism, an overproduction of thyroid hormones [2, 3]. These situations may result in a range of symptoms, including fatigue, mood changes, weight fluctuations, and irregular periods. Many factors can lead to thyroid disorders, including autoimmune conditions, iodine deficiency, genetics, and medications. This is typically done through proper doctor examination and blood tests to measure thyroid hormone levels [4]. Conditional on the exact illness, potential treatment options to correct disorders can include hormone replacement therapy, medications to regulate hormone levels, or even surgery in some cases. Thyroid disease covers a range of conditions, including hypothyroidism and cancer of the thyroid, and has been connected with mortality and death, although with varying strength [5]. Thyroid cancer represents the largest proportion of deaths associated with thyroid disorders [6]. Thyroid problems are difficult to diagnose and take a long time to diagnose, usually by extensive clinical evaluation and laboratory tests [7]. The use of cognitive computing for predicting flow status of a flexible rectifier is a new approach that leverages sophisticated machine learning (ML) techniques like Multilayer Perceptron and CatBoost to estimate the flow and reverse flow rates of fluids in flexible rectifiers [8].

***Corresponding author:** Asghar Ali Shah, Department of Computer Science, Kateb University, Afghanistan. Email: asghar.ali.shah@kateb.edu.af and Khan Muhammad Adnan, Department of Software, Gachon University, Republic of Korea. Email: adnan@gachon.ac.kr

2. Related Work

Previous techniques in the literature dedicated to the diagnosis of thyroid diseases have a strong dependence on classification algorithms when trained on an imbalanced dataset, while presenting moderate levels of performance scores. These methodologies typically do not address data balance problems, resulting in biased findings and reduced diagnostic performance. Moreover, current diagnostic methods that rely on these algorithms often lack Explainable Artificial Intelligence (XAI) models, further dulling the interpretability and trustworthiness of the predictions. Therefore, the advanced ML approach meets the aforementioned gaps by including robust data balancing techniques and utilizing XAI models, improving the accuracy and transparency of thyroid disorder diagnoses. The present study can potentially contribute towards facilitating the diagnostic process through improved accuracy, enabling early detection, and treating thyroid cancer disorders using AI technology.

Song et al. [9] performed an oversampling on serum Raman spectroscopy data collected from 47 thyroid carcinoma patients, using reweighing methods of the pillow, random forest (RF), and decision tree (DT) classifiers. In the Synthetic Minority Over-sampling Technique (SMOTE) method, accuracy was DT (75.4%), RF (81.5%), and Adaptive Boosting (84.61%). Liu et al. [10] discussed data up to October 2023, an RF framework for predicting post-thyroid ectopy quality of life derived from 286 thyroid cancer patients that had an accuracy of 89.7% validation rate. Kim et al. [11] investigated data from 187 patients treated with PD-1 or PD-L1 inhibitors, which were used in the article to predict adverse events with RT and logistic regression (LR) classifiers. RF provided an area under the receiver operating characteristic (ROC) curve (AUC) of 77.0%. Luong et al. [12] performed an RF, and LR was used to train the model used in the analysis of 355 surgical patients for indeterminate thyroid nodules data. The example output obtained from a test set was as follows: Accuracy 71.9%, AUC 85.9%, Sensitivity (true positive [TP] rate) 75.5%, and Specificity (true negative [TN] rate) 82.4%. Yang et al. [13] discussed data utilized to evaluate the prediction of recurrences of differentiated thyroid carcinoma using DT, RF, and LR classifiers, achieving accuracy levels of 84.7% to 89.7%. Idarraga et al. [14] performed a study based on 604 benign thyroid nodules and established an operating model with RT classifiers to predict malignancy, achieving a lower AUC of 64.0%. Xu et al. [15] investigated Serum thyroglobulin numbers amongst 1451 patients meeting the criteria for papillary thyroid carcinoma using an RF classifier, resulting in an AUC ranging from 71.0% to 81.0% in the development of a possible diagnostic and prognostic tool. Hang [16] performed testing of an RF classifier on 428 cases and implemented histogram equalization and alternative ultrasound image enhancement techniques. The accuracy was found to be 95.0%. Yang et al. [17] discussed the method for lncRNA and miRNA data of 506 thyroid carcinoma patients evaluated in support vector machine (SVM), RF, and DT classifiers. The maximum accuracy was recorded by RF. Qin et al. [18] investigated an AUC of 96.9% achieved from an ensemble method performed on magnetic resonance imaging (MRI) radionics in lymph node metastases of 109 thyroid carcinoma patients.

Sai et al. [19] performed electroglottogram signals to determine thyroid abnormality using Naïve Bayes (NB), DT, artificial neural network (ANN), LR, and RF classifiers. In this study, LR obtained an overall accuracy of 95.1%. Aiming for Early diagnosis of thyroid cancer using a Saudi Arabian dataset best accuracy of 90.9% was achieved with RT using NB, RF, SVM, and ANN classifiers. Olatunji et al. [20] performed and applied DNA methylation data from 92 samples to differentiate between follicular thyroid carcinoma and adenoma. Zhang et al. [21] investigated, and the accuracy achieved was 99.4% by applying the RF classifier. According to He et al. [22], RF classifiers achieved an accuracy of 90.6% on the analysis of biometric MRI data from 60 cases of papillary thyroid carcinoma. Garcia de Lomana et al. [23] performed gradient boosting (GB), RF, SVM, ANN, and LR to model thyroid hormone homeostasis. Reported accuracy ranged from 76.0% to 82.0%, depending on the model. Aksu et al. [24] discussed that the RF approach was used to assess radiomic data from 18F-FDG thyroid gland uptake to achieve an accuracy of 78.6% and an AUC of 84.9% for 60 cases of the thyroid gland. Chen et al. [25] investigated, for instance, the ultrasonographic feature of 1558 cases of thyroid nodules was analyzed using forest models, reaching an accuracy of 96.1%. The key findings of Su et al. [26] from using RF classifiers to analyze gut microbiome data related to thyroid dysfunction, but did not report performance metrics. Zhao et al. [27] performed a combination of 177 patients with partially cystic thyroid cancer and retrospective data, and high accuracy and AUC values were achieved by applying GB, ANN, RF, SVM, and LR classifiers; the best performing was RF with a specificity of 93.4%. Kwon et al. [28] performed LR, RF, and SVM classifier analysis using radiomics data from 96 cases for predicting BRAF mutation in papillary thyroid carcinoma in this study, achieving an accuracy of 64.3%.

Anu and Benifa [29] performed to handle raw data related to hyperthyroidism, using an ANN with K-means. Vivar et al. [30] discussed using an ANN with loss at its input network and coupled variations at test time to continuously determine feature value. An automated assessment algorithm is presented. This method is also used to classify hypothyroidism and hyperthyroidism in the UCI thyroid dataset. van Sonsbeek and Worring [31] performed an ANN and a novel method of information representation and argumentation to propose a decision support system for the evaluation of thyroid dysfunction in general. Ai et al. [32] performed capsule networks, a more contemporary kind of neural network, which was shown to achieve a high accuracy of 81.06% when used on ultrasonic thyroid imaging to identify possible symptoms of cancer of the thyroid.

Xi et al. [33] took a novel clinical dataset holding 724 patients with 1,232 nodules and used multiple classifiers, and the best classifier is RT, by 11% increases on accuracy, and 12% on F1 score. The two general measurements and average accuracy, and AUROC of the six models are 0.78.

Książek [34] took the dataset from Shengjing Hospital of China Medical University, holding 1232 records with 19 features, and performed 10 different classifiers, including XGBoost, LightGBM, and RF. The LightGBM classifier confirmed the maximum performance classification accuracy of 81.82% and an F1 score of 86.62%.

Akter and Mustafa [35] took the dataset from the UCI Machine Learning Repository, holding 6,916 records with 22 features, and performed different classifiers, achieving a maximum performance classification accuracy of 87%.

Kumar et al. [36] took the dataset, which included 1,250 male and female subjects whose ages ranged from one to one year, with 17 attributes, and performed multiple classifiers.

Alawiyah et al. [37] developed a thyroid cancer recurrence prediction model using the XGBoost method, achieving an average accuracy of 97.74% and an F1 score of 95.94%, effectively predicting both recurrence and non-recurrence, aiding clinical decision-making for patient management. Ozturk et al. [38] used different machine learing (ML) models, including GB, RF, XGBoost, and AdaBoost, to perform in forecast the recurrence of thyroid cancer. The best prediction

accuracy at 95.3% was achieved by the XGBoost model, which suggests it could enhance forecasting in clinical settings. Firat Atay et al. [39] created a hybrid model integrating predictive classification techniques and association rule mining to forecast the recurrence of differentiated thyroid cancer with a remarkable accuracy of 96.7%. The model also facilitated the identification of critical management predictors like incomplete response and lymphadenopathy for optimized decision support systems. The study by Vu et al. [40] built a machine-learning model for thyroid cancer detection using clinical data, achieving an accuracy of approximately 82%.

In traditional previous work, as shown in Table 1, many ML models are deployed while having a "black-box" problem, where there is no simple understanding of the inner mechanisms by human experts of the model. The motivation behind this study is that the lack of transparency is a major downside in healthcare contexts, where doctors want to understand why a model made a certain prediction to have confidence in its outputs.

In our study, cluster-based SMOTE is used for balancing the dataset. Other preprocessing techniques, including missing value handling and normalization, are also used in this research. Then, we applied six different classifiers, including RF, DT, NB, Light, K-nearest neighbors (KNN), and SVMs. By using the ensemble learning technique, SVM is selected that achieves the highest accuracy among other classifiers. To ensure result transparency, Local Interpretable Model-Agnostic Explanation (LIME) and Shapley Additive Explanation (SHAP) tools are used for interpretability and XAI.

The main contribution is as follows:

1) An Enhanced Ensemble ML Approach is proposed for Transparent Thyroid Cancer Prediction with XAI that enhances the accuracy and effectiveness of thyroid cancer predictions.
2) Our study presents a state-of-the-art cluster-based SMOTE data balancing technique that addresses the limitations of existing approaches. These techniques decrease overfitting and eliminate noise from the data, enhancing model flexibility.
3) A series of preprocessing methods has been applied, including missing value imputation, outlier detection, normalization, and addressing imbalance and cluster-based SMOTE issues.
4) Introduce a new soft voting-based SVM classifier with parameter configuration.
5) The performance of the proposed classifier is predicated on various performance matrices with K-fold cross-validation.
6) Experimental results depict the leads of the proposed soft voting-based SVM prediction classifier with an accuracy of 97%, outperforming the performance of ruling ML models.
7) Furthermore, SHAP and LIME techniques have been applied to explain the efficiency of the proposed model concerning a particular feature

3. Methodology

Figure 1 shows the proposed model, which consists of two phases, namely the training phase and the validation or evaluation phase. Data balancing and preprocessing techniques have been applied to the dataset in the training phase.

3.1. Dataset

This research utilized thyroid-based datasets from well-known data repositories, Kaggle. The total size of samples in the dataset is 3,772, and the total attributes is 28. This is probably because thyroid conditions can influence or correlate with ocular health, hence constituting important latent features for prediction.

3.2. Clinical validity

The dataset can be found on Kaggle and contains information from actual patients that the Garavan Institute and J. Ross Quinlan from the New South Wales Institute in Sydney, Australia, gathered. It features 3,772 cases along with 30 attributes, which include demographics and laboratory tests for thyroid hormones such as T3, TT4, thyroid-simulating hormone (TSH), FTI, and medication status. Several peer-reviewed publications have cited this dataset, demonstrating that the thyroid and endocrine AI model's prediction features T3, TT4, TSH, and age align with important thyroid and endocrine healthcare benchmarks, thereby reinforcing its clinical relevance. Literature-based feature importance within AI models repeats the clinical setting and confirms the critical nature of these diagnostic parameters.

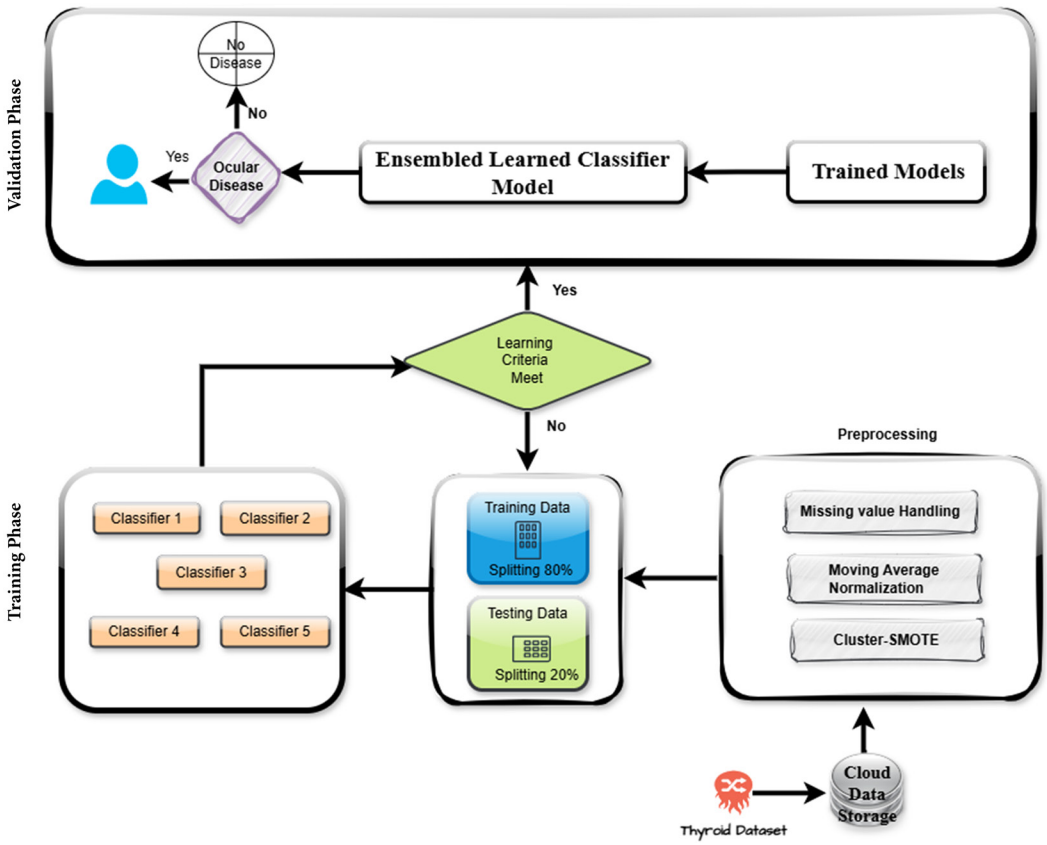
3.3. Data balancing

In this study, the cluster-based SMOTE is used to address class imbalance.

Table 1
Limitations of previous work

Citation	Methods	Accuracy	Dataset size	Generalization
Sai et al. [19]	Multiple Algorithm	90.9%	218	Yes
Olatunji et al. [20]	Random Forest	99.4%	92	-
Zhang et al. [21]	Random Forest	90.6%	60	-
He et al. [22]	Multiple Algorithm	82.0%/79.0%	1,074	Yes
Garcia de Lomana et al. [23]	Random Forest	78.6%	60	-
Aksu et al. [24]	Random Forest	96.1%	1,558	-
Chen et al. [25]	Random Forest	-	92	-
Su et al. [26]	Multiple Algorithm	86.0%	177	Yes
Zhao et al. [27]	Multiple Algorithm	64.3%	96	Yes
Xi et al. [33]	Multiple Algorithm	78%	724	Yes
Książek [34]	Multiple Algorithm	81.82%	1,232	Yes
Proposed	Multiple Algorithm	Our study	3,773	K-fold

Figure 1
The proposed model



3.4. Data preprocessing

In this study, missing values were handled using other features' values. The moving-average-based normalization technique is also used for handling negative, imaginary, or abrupt values.

3.5. Data splitting

Data retrieved from the cloud is divided into two sets: Training Data: 80% of the data is taken to train the machine-learning model. Testing Data: 20% of the data is set aside for model testing.

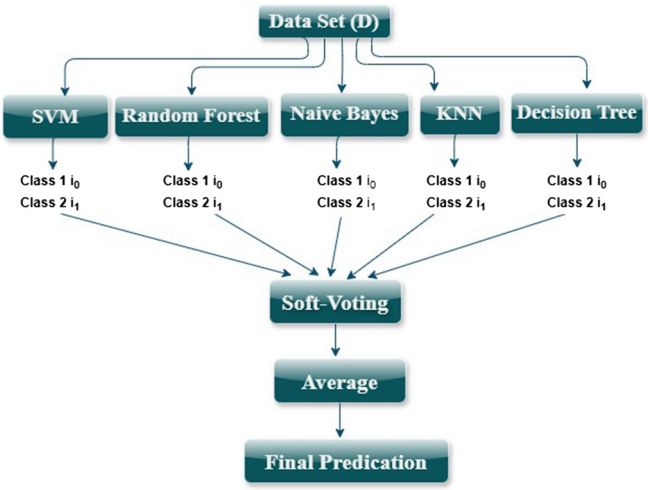
3.6. Normalization

For this dataset, normalization was done using smoothing using a moving average technique, which helps in reducing the impact of outliers, thus enhancing the model's performance considerably. To balance out the problem of sparse data, this work incorporates cluster-based SMOTE. In addition, the classifiers' outputs are combined using the soft-voting technique, which adds similarly weighted contributions from several classifiers for greater accuracy, as demonstrated in Figure 2.

Two classes are present in our dataset, Classes 1 and 2, and it is represented by $i[0,1]$. The ensemble learning phase would be predicted as follows:

$$\textit{Predication classifier } 1\ (SVM) \rightarrow [P(i_0/D), P(i_1/D)]$$
$$\textit{Predication classifier } 2\ (R.F) \rightarrow [P(i_0/D), P(i_1/D)]$$
$$\textit{Predication classifier } 3\ (N.B) \rightarrow [P(i_0/D), P(i_1/D)]$$
$$\textit{Predication classifier } 4(KNN) \rightarrow [P(i_0/D), P(i_1/D)]$$
$$\textit{Predication classifier } 5\ (D.T) \rightarrow [P(i_0/D), P(i_1/D)]$$

Figure 2
Soft-voting implemented on multiple ML classifiers



Note: KNN = K-nearest neighbors, ML = machine learning, SVM = support vector machine.

$$\textit{Predication classifier } 6\ (LightGBM) \rightarrow [P(i_0/D), P(i_1/D)]$$
$$\textit{Soft-voting} = \arg\ \textit{Max}\ [P(\textit{Predicitive classifier } 1), (\textit{Predicitive classifier } 2),$$
$$P(\textit{Predicitive classifier } 3),$$
$$P(\textit{Predicitive classifier } 4), P(\textit{Predictive classifier } 5),$$
$$P(\textit{Predictive classifier } 6)]]$$

Table 2
Implementation of soft voting on ML classifiers

	Soft voting			
	Classifier	**Accuracy**	**Class 0**	**Class 1**
0	DT	0.750	0.825	0.635
1	SVM	0.970	0.986	0.945
2	KNN	0.903	0.915	0.881
3	Naïve Bayes	0.710	0.742	0.674
4	Random forest	0.953	0.973	0.891
5	LightGBM	0.926	0.945	0.875

Note: DT = decision tree, KNN = K-nearest neighbors, ML = machine learning, SVM = support vector machine.

3.7. Soft voting

Ensemble learning chooses the best-performing classifier dynamically using soft voting in Table 2 instead of relying on a single model. It helps improve prediction accuracy by selecting the most confident classifier. Thereafter, the ML model analyzes the normalized dataset with soft voting and outputs predictions when the learning conditions are satisfied, and then the XAI is employed to break down the learning process and provide multiple different explanations in a case where the criteria were not met, ultimately enhancing the model once more in preparation for better results. By making such a system transparent, it allows healthcare providers to verify those predictions. In this stage, the data is stored, pre-stage data is reduced, and the normalization of data is achieved using cloud infrastructure to keep the model accessible and scalable. Furthermore, I propose a possible relationship between thyroid disorder and ocular disorder, suggesting that the thyroid dataset can provide more predictive features in disease detection when included in the model. Our model for ocular disease prediction uses a synthetic ML system with XAI that utilizes a preprocessing and normalization stage of data.

3.8. XAI

The model, according to Figure 1, generates predictions and explanations using XAI. It allows the ML model outputs to be interpretable, which is crucial for transparent decision-making. These explanations are intended to justify the predictions but can also help to learn the underlying reasons for predictions made by the model (in this case, risk factors of ocular diseases).

3.9. Validation phase

This phase begins with a person being tested for the disease. The decision node is used to check whether the person has any disease. If not, the process stops; if yes, the data is retrieved from the cloud for pre-reduction. This cloud-based storage system has all the appropriate datasets. The cloud helps to store some of the data to make predictions.

3.10. Cloud data storage

Data from the cloud is fetched to aid in the prediction process.

4. Simulation and Results

4.1. Histograms for numerical features

Figure 3 represents the histogram distribution of six numerical features. Age: The distribution of ages in the set is shown with this histogram. The distribution seems to be right-skewed, meaning there are more younger people in the data compared to older people. TSH: This histogram illustrates the distribution of TSH levels, which is a hormone secreted by the pituitary gland that activates the thyroid gland. It is right-skewed, meaning that most of the people in the dataset have low TSH levels. T3: This histogram shows the distribution of T3 levels, another thyroid hormone. It is also right-skewed, meaning the majority have low T3. TT4: This histogram shows the total thyroxine (TT4) distribution. It is positively skewed, indicating that the majority of people have lower levels. T4U: This

Figure 3
Histograms for numerical features

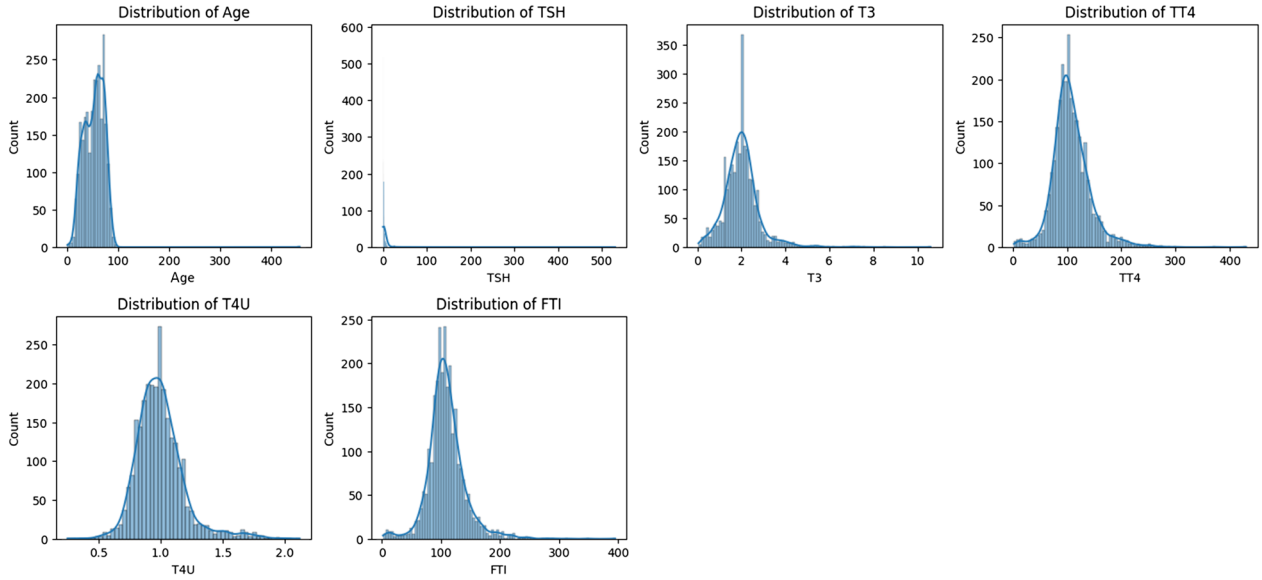
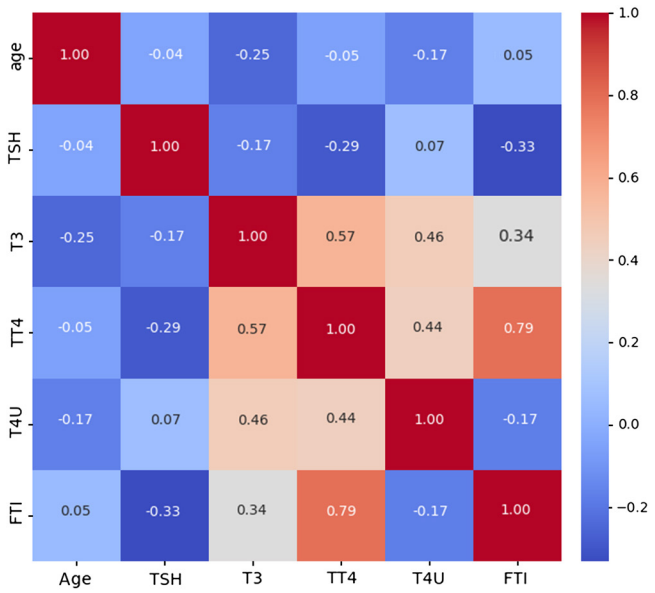


Figure 4
Correlation matrix heatmap



plot is a histogram of the T4U, a measurement of unbound or free thyroid hormone levels. It is slightly right-skewed. FTI: FTI is a free thyroid hormone, and this histogram displays where the FTI distribution falls on the chart. It is right-skewed, meaning that the majority of folks tend to have lower FTIs.

4.2. Correlation matrix heatmap

Figure 4 shows the correlation matrix, plotted in a heatmap, of the relationship among some factors in a dataset. The darkness of the color signifies the correlation, meaning that the closer to reading the color, the higher the positive correlation, with blue indicating a strong negative correlation and white indicating none. Strong Positive Correlations: TT4 and FTI: The strongest correlation (0.79) indicates that Total T4 and Free T4 have a strong association. T3 and TT4: There is a strong correlation (0.57) between T3 and Total T4 as well. Moderate Positive Correlations: T3 and T4U: A moderate correlation (0.46) indicates a positive relationship between T3 and Unbound T4 (T4U). TT4 and T4U: A moderate correlation (0.44) implies a positive relationship between Total T4 and Unbound T4. Weak Correlations: age and TSH, age and T3, age and TT4, TSH and T4U, T3 and TSH, and T4U and FTI: All these pairs have weak correlation and hence are less related. TSH and FTI: The relationship is weak and negative, which means there is a slight inverse relationship. Interpretation: From this heatmap, it can be seen that some of the variables are highly correlated with each other, especially with the Thyroid hormone-related measures. It indicates the possibility of redundancy among these variables, or at least that one measure can adequately summarize the information provided by multiple measures. Further Analysis: Feature Selection: The significant correlations indicate a need for feature selection to mitigate redundancy and adverse effects on model performance. Domain Expertise: Knowledge of how to interpret the relations in the medical world could be beneficial. It is advised to do some consultation with domain experts to find out such biological mechanisms. Predictive Modeling: The identified correlations can facilitate the development of effective predictive models for thyroid disorders

Pair plot for selected numerical features: The pair plot shown in Figure 5 visualizes the relationships between different numerical features in a dataset, with a focus on the potential impact of sex on these relationships. Distribution of Features: The diagonal of the plot shows the distribution of each feature (age, T3, TT4, T4U, and FTI). These distributions are shown as histograms with kernel density estimates. The distribution of FTI appears to be skewed to the right.

Scatter Plots: The off-diagonal plots are scatter plots showing the relationship between pairs of features. Each point represents an individual data point, and the color of each point represents the individual's sex (blue for female, orange for male).

Relationships Between Features: A slightly positive correlation is observed between TT4 and T3, with a tendency for higher values of T3 to correspond to higher values of TT4. A weak positive correlation is observed between FTI and T3. A moderately strong positive correlation is observed between TT4 and FTI. T4U and T3: No clear correlation is observed between T4U, and a slightly positive correlation is observed between T4U and FTI. A weak positive correlation is observed between T4U and TT4. This pair plot provides an initial overview of the relationships between the selected numerical features. The color-coding by sex reveals potential differences in these relationships based on sex. Further analysis, such as correlation coefficients and statistical tests, could provide more concrete evidence of these relationships and their significance. The information from this plot can be used to guide further analysis, model building, and hypothesis testing.

The distribution of sex shown in Figure 6 shows that the number of females in the dataset is more than the number of males in the dataset.

5. Experimental Setup

The experiments were conducted using the following hardware configuration available on Google Colab: Processor: Intel Xeon CPU (Single core, 2.20 GHz), Graphics Processing Unit: NVIDIA Tesla T4/ Tesla P100/Tesla K80 (depending on availability and session allocation by Colab). 16 GB of GDDR5 VRAM (for K80) or 16/32 GB HBM2 VRAM (for T4/P100). System Memory (RAM): System RAM 1.3/12.7 GB available. Disk Storage: Disk 32.4/107.7 GB storage is provided by the Colab environment. Operating System: Ubuntu 18.04 LTS (managed via Google Colab backend). Frameworks and Libraries: NumPy, Pandas, and other ML tools as needed.

6. Analysis of classifier report after implementing cluster-based SMOTE

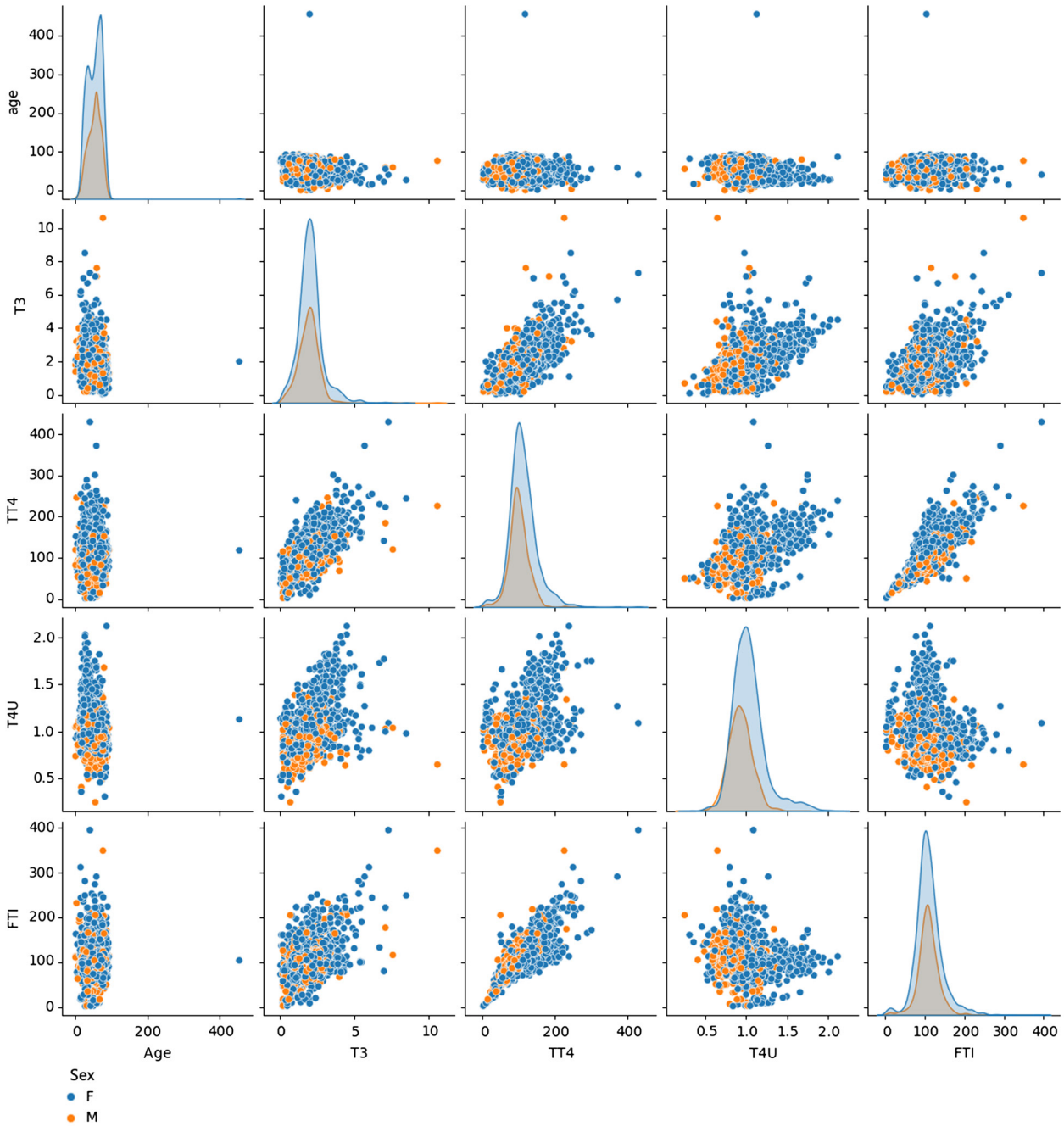
Table 3 presents the hyperparameters used in six different classifiers.

Cluster-based SMOTE aims to address the class imbalance issue, so after the implementation of cluster-SMOTE, our analysis report of the classifier is shown in Figure 7.

Table 4 describes the individual training time and computational complexity of all six classifiers, both training and testing. In an ensemble learning approach, the computational complexity depends on the individual complexities of the classifiers. Because here in this study, the ensemble learning method chooses SVM to predict the final model.

Accuracy: According to Table 5, below 75.0% of citations were accurately correct for the DT. SVM: 97.0% best accuracy. KNN gave 90.33% accuracy, which was less than SVM. Naïve Bayes' accuracy was 71.0%, the worst performance. RT accuracy is 95.33%, near SVM Of the classifiers mentioned, SVM has the best

Figure 5
Pair plot for selected numerical features



accuracy, whereas Naïve Bayes performs the worst. LightGBM gives an accuracy of 0.96.

AUC: As per Table 5, DT 74.0% of predictions are correct. SVM 0.990% best accuracy. KNN 0.979%, which is lower than SVM. Naïve Bayes 0.878% the lowest accuracy. Almost the same as SVM, with a high accuracy of 0.990%. The best accuracy of SVM is on par with other classifiers, and Naïve Bayes significantly underperformed. RT AUC of 0.976%. LightGBM gives an AUC of 0.740.

Recall: in Table 5, DT has 0.750 recall (75% positive cases), SVM. Its recall is 0.970 (high), 97% case-positive. KNN recall is 0.903, so it is good in sensitivity. Naïve Bayes's lowest recall score at 0.710, that is, only 71% of the positive cases were identified. RT has a 0.953 recall, meaning that there is a catch of 95.33% of predictions. Naïve Bayes is more insensitive because SVM and RT outperform Naïve Bayes for positive case detection. LightGBM gives a recall of 0.92 for case-positive.

Precision: Table 5 DT precision is 0.881 (88.1% of pos). Confident Positive from SVM has 0.972 precision, and KNN precision is 0.938. It holds for given data when the classifier predicts positive around positive confidence. Naïve Bayes precision is 0.900 (ouch for

Figure 6
Distribution of sex

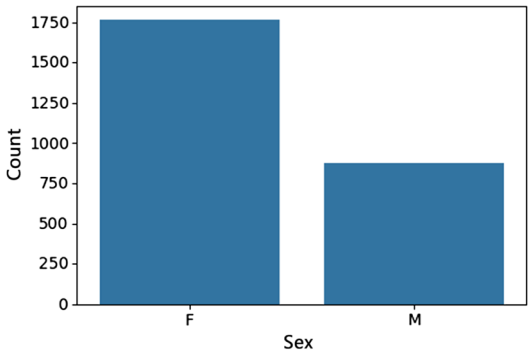
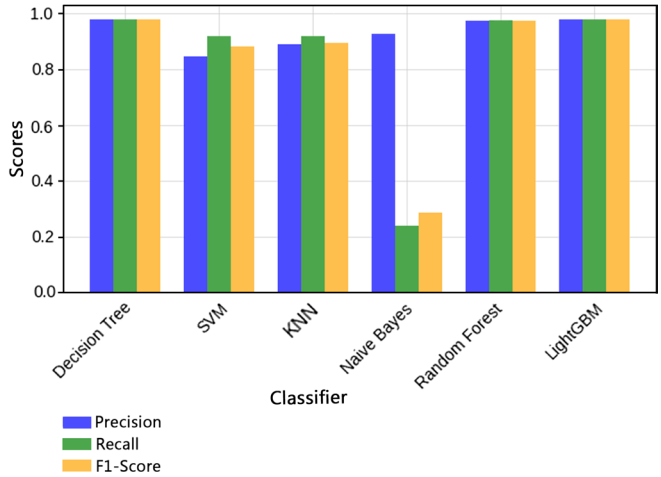


Table 3
Hyperparameters for classifiers

	Classifier	Hyperparameters
1	DT	Max depth =14
2	SVM	Kernel: "linear," degree: Default 3
3	KNN	Weights: "uniform"
4	Naïve Bayes	Alpha: Default 1.0
5	Random forest	"Max depth = 40, Features used for splitting = 50%"
6	LightGBM	Boosting_type: "gbdt"

Note: DT = decision tree, KNN = K-nearest neighbors, SVM = support vector machine.

Figure 7
Precision, recall, and F1 score comparison



Note: KNN = K-nearest neighbors, SVM = support vector machine.

lower accuracy, but still pretty large). After SVM, RT gives the second-highest accuracy score: 0.962. Here again, SVM leads in accuracy, and RF. DT has also performed well, and whereas Naïve Bayes is not accurate, its precision is good enough. LightGBM gives a precision of 0.949.

F1 score: According to Table 5 and Figure 7, the DT F1 score of 0.793 is a balanced, but moderate, performance. SVM is the best F1 score: 0.971, with good precision and recall. KNN score is 0.914; Naïve Bayes score is 0.764. The RT score is 0.956. The two best-performing classifiers out are SVM and RF; Naïve Bayes performs poorly, owing to its low recall. LightGBM gives an F1 score of 0.933.

Table 4
Training time and complexity

	Classifier	Training Time	Complexity	
			Training	Testing
1	DT	0.0161 seconds	O(nmlog(n))	O(d)
2	SVM	0.6944 seconds	O(n2)	O(d)
3	KNN	0.0039 seconds	O(nd)	O(d)
4	Naïve Bayes	0.0204 seconds	O(nd)	O(d)
5	Random forest	0.3607 seconds	O(nlog(n)m)	O(d)
6	LightGBM	0.1083 seconds	O(n)	O(d)

Note: DT = decision tree, KNN = K-nearest neighbors, SVM = support vector machine.

7. Analysis of Classifiers Using Comparison

SVM performs best across all metrics, making it the most reliable classifier for this task, as shown in Table 6 and Figure 8.

The outcomes of the SVM classifier's performance are compared with other models in Table 7, including DT, KNN, Naïve Bayes, RF, and LightGBM, according to the table. The SVM performance comparison involves two tests in Table 7: the t-test is based on average metrics (in this case, accuracy), and the DeLong test is based on the AUC. The test results include both the test statistic (t or Z value) and the corresponding p-value. A small p-value (commonly <0.05) indicates the tested models differ meaningfully. The results confirm SVM outperforms DT and Naïve Bayes both in accuracy and AUC, as their p-values are very low. Against KNN, RF, and LightGBM, the p-values exceed 0.05, suggesting a lack of performance difference equivalently performs with these models. All in all, the table illustrates that SVM's model performance enhancements are overwhelmingly statistically valid, whereas SVM's performance enhancements over other strong models are marginal.

Table 8 and Figure 8 show that the AUC probabilities of the DT classifier achieve a reasonable level of accuracy, with an AUC of 0.81, indicating that it is moderately good at distinguishing between positive and negative classes. The SVM classifier demonstrates excellent performance with an AUC of 0.99, indicating that it is highly effective at classifying positive and negative instances. The RT classifier achieves a high level of accuracy with an AUC of 0.96, suggesting its strong ability to distinguish between positive and negative classes. The KNN classifier exhibits good performance with an AUC of 0.94, demonstrating its effectiveness in classifying positive and negative instances. The Naïve Bayes classifier performs well with an AUC of 0.93, indicating its capability to discriminate between positive and negative classes. The LightGBM classifier achieves a reasonable level of accuracy with an AUC of 0.97.

The overall comparison shows that the best-performing classifier is SVM, which shows the highest AUC scores, suggesting it performs better than the other classifiers. DT has a moderate AUC, indicating a less effective performance compared to another classifier. DT, KNN, and Naïve Bayes classifiers achieve good performance with AUC scores close to 0.9, indicating their capability for classification.

Table 5
Precision, recall, and F1 score comparison

	Classifier	Accuracy	AUC	Recall	Precision	F1
0	DT	0.750	0.740	0.750	0.881	0.792
1	SVM	0.970	0.990	0.970	0.971	0.970
2	KNN	0.903	0.979	0.903	0.937	0.913
3	Naïve Bayes	0.710	0.878	0.710	0.899	0.763
4	Random forest	0.953	0.976	0.953	0.962	0.955
5	LightGBM	0.926	0.961	0.926	0.949	0.933
	Soft voting ensemble	0.869	0.921	0.869	0.933	0.888

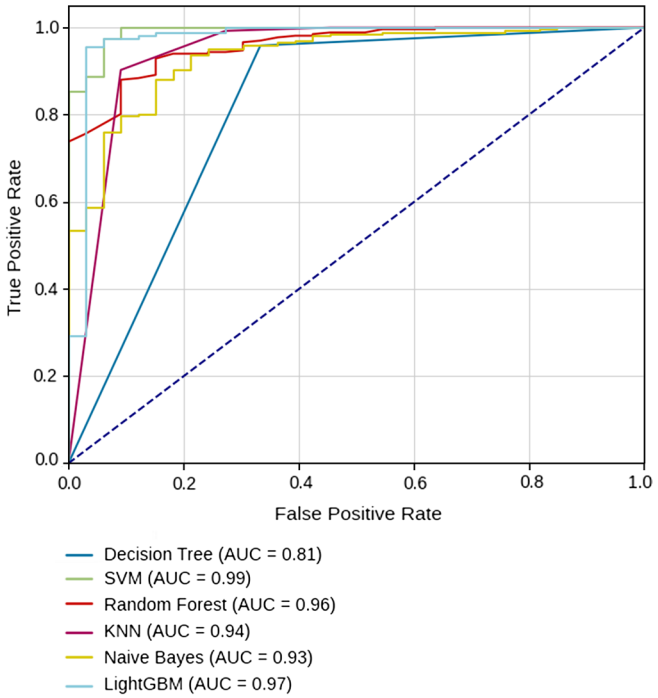
Note: AUC = area under the receiver operating characteristic curve, DT = decision tree, KNN = K-nearest neighbors, SVM = support vector machine.

Table 6
Accuracy comparison

	Classifier	Accuracy
0	DT	0.750
1	SVM	0.970
2	KNN	0.903
3	Naïve Bayes	0.710
4	Random Forest	0.953
5	LightGBM	0.926

Note: DT = decision tree, KNN = K-nearest neighbors, SVM = support vector machine.

Figure 8
Classifiers with AUC probabilities



Note: AUC = area under the receiver operating characteristic curve, KNN = K-nearest neighbors, SVM = support vector machine.

Table 7
Statistical comparison of SVM performance against other classifiers using the T-test and the DeLong test

	T-test (*t*)	*T*-test (*p*)	DeLong *Z*	DeLong *p*
SVM vs. DT	3.87	0.0001	2.92	0.0035
SVM vs. KNN	1.76	0.079	1.21	0.225
SVM vs. Naïve Bayes	4.71	0.00001	3.76	<0.001
SVM vs. RF	0.98	0.325	0.97	0.332
SVM vs. LightGBM	1.38	0.167	0.85	0.395

Note: DT = decision tree, KNN = K-nearest neighbors, RF = random forest, SVM = support vector machine.

Table 8
AUC probabilities of classifiers

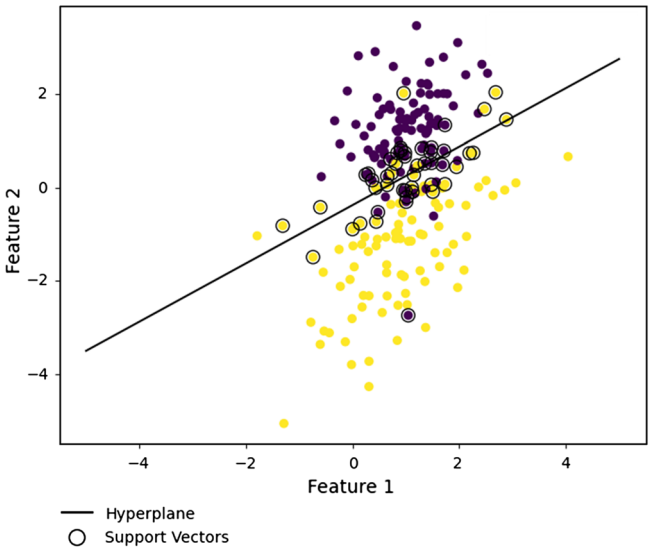
Classifiers	AUC
DT	0.812
SVM	0.991
Random forest	0.958
KNN	0.939
Naïve Bayes	0.933
LightGBM	0.97

Note: AUC = area under the receiver operating characteristic curve, DT = decision tree, KNN = K-nearest neighbors, SVM = support vector machine.

8. Analysis of Classifiers Through AUC

Train the SVM classifier on these two classes of data points in Figure 8. The hyperplane separates the data points from their respective classes. Primarily, data points known as support vectors determine the hyperplane. Figure 9 represents two classes of data points, each with a distinct color (purple and yellow). The black line in the figure is the hyperplane. This line represents the decision boundary of the SVM classifier. The classifier will classify any data point above the hyperplane as belonging to one class (yellow) and any data point below it as belonging to the other class (purple). The data points in circles

Figure 9
SVM visualization



Note: SVM = support vector machine.

Figure 10
Confusion matrix for training data

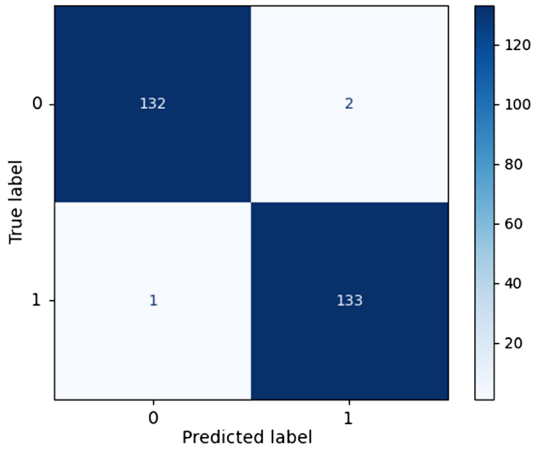
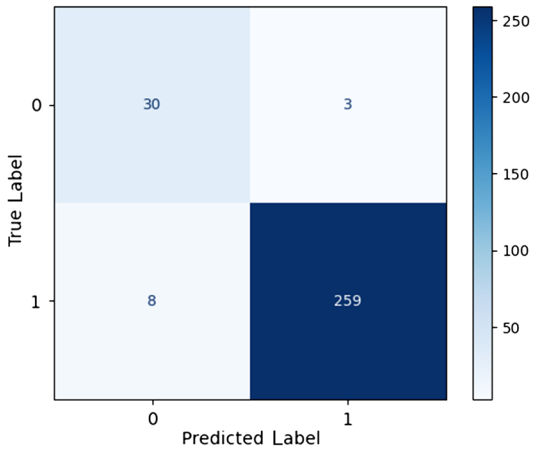


Figure 11
Confusion matrix for testing data



are the support vectors. These are the data points that are closest to the hyperplane and therefore have the greatest influence on the position of the hyperplane. Support vectors define this hyperplane. The purpose of the SVM visual is to provide a clear explanation of how the SVM classifier works to classify data points into a particular class by fitting data points with a hyperplane. By separating data points with the largest margins, the support vectors are those techniques closest to the hyperplane, making them essential for both defining the decision boundary and classifying the classifier.

Confusion Matrix: In Figure 10, 132 (top left): This cell indicates TPs, where the classifier correctly predicted class 0 for instances that were class 0. 2 (top right): This cell indicates false positives (FPs), where the classifier incorrectly predicted Class 1 for instances that were Class 0. 1 bottom left): This cell indicates false negatives (FNs), where the classifier incorrectly predicted Class 0 for instances that were Class 1. 133 (bottom right): This cell indicates TNs, where the classifier correctly predicted Class 1 for instances that were Class 1.

$$\text{Accuracy} = \frac{TP+TN}{\text{TP+FP+TN+FN}} \tag{1}$$

$$\text{Precision} = \frac{TP}{\text{TP+FP}} \tag{2}$$

$$\text{Recall} = \frac{TP}{\text{TP+FP}} \tag{3}$$

$$\text{Specificity} = \frac{TN}{\text{TN+FP}} \tag{4}$$

In Figure 11, TP: The classifier correctly predicted Class 1 (259). TN: The classifier correctly predicted Class 0 (30). FP: The classifier incorrectly predicted Class 1 when the actual class was 0 (3). This is also known as a Type I error. FNs: The classifier incorrectly predicted Class 0 when the actual class was 1 (8). This is also known as a Type II error.

9. Validation

The K-fold cross-validation results for the six ML classifiers are DT, SVM, KNN, Naïve Bayes, RF, and LightGBM in Table 9. Each model was evaluated over five folds, recording accuracy for each fold, and performance metrics were averaged across all folds. These include average accuracy, average AUC, average recall, average precision, and average F1 score. Based on the results, SVM comes out on top with an average accuracy of 0.97 and the highest AUC with 0.99, excelling in recall, precision, and F1 score as well. RF follows closely at 0.95 accuracy and 0.98 AUC, with strong performance from LightGBM at 0.92 accuracy. KNN performs well in AUC and precision, but falls short in recall and F1. DT and Naïve Bayes are weaker performers for this dataset, with low accuracy and recall, despite Naïve Bayes having high precision. The table makes it clear that SVM is the most reliable classifier for this classification task, standing out as the most dependable across multiple scenarios.

To further validate the performance of the proposed mode in Table 10, two additional publicly available datasets are used to confirm the proposed model's robustness and generalization capability.

In Table 11, the comparison of the previously published work is shown, which clearly shows that the proposed model achieved a maximum accuracy of 96.66% as compared to state-of-the-art published articles.

The mean SHAP value in Figure 12 shows both classes. The contribution of that feature to the model's prediction for Classes 0

Table 9
K-fold cross-validation performance comparison of machine learning classifiers

Classifier	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average accuracy	Average AUC	Average recall	Average precision	Average F1
Decision tree	0.74	0.76	0.75	0.75	0.75	0.75	0.74	0.75	0.88	0.79
SVM	0.97	0.97	0.96	0.98	0.97	0.97	0.99	0.97	0.97	0.97
KNN	0.90	0.91	0.91	0.90	0.92	0.90	0.98	0.90	0.94	0.91
Naïve Bayes	0.71	0.71	0.72	0.70	0.71	0.71	0.88	0.71	0.90	0.76
Random forest	0.95	0.96	0.96	0.95	0.96	0.95	0.98	0.95	0.96	0.96
LightGBM	0.92	0.93	0.93	0.92	0.93	0.92	0.96	0.92	0.95	0.93

Note: AUC = area under the receiver operating characteristic curve, KNN = K-nearest neighbors, SVM = support vector machine.

Table 10
Comparison table of two different datasets

	Classifier	Accuracy	AUC	Recall	Precision	F1
Dataset 1	DT	0.895	0.910	0.90	0.89	0.89
	SVM	0.901	0.923	0.89	0.90	0.89
	KNN	0.878	0.882	0.88	0.88	0.88
	Naïve Bayes	0.864	0.875	0.87	0.86	0.86
	Random forest	0.915	0.938	0.92	0.91	0.91
	LightGBM	0.920	0.942	0.92	0.92	0.92
	Soft voting	0.926	0.950	0.93	0.93	0.93
Dataset 2	DT	0.915	0.934	0.92	0.92	0.92
	SVM	0.921	0.947	0.92	0.92	0.92
	KNN	0.894	0.899	0.90	0.89	0.89
	Naïve Bayes	0.889	0.894	0.89	0.89	0.89
	Random forest	0.934	0.962	0.93	0.93	0.93
	LightGBM	0.940	0.968	0.94	0.94	0.94
	Soft voting	0.944	0.974	0.94	0.94	0.94

Note: AUC = area under the receiver operating characteristic curve, DT = decision tree, KNN = K-nearest neighbors, SVM = support vector machine.

and 1 is shown. The importance of the feature concerning the model prediction for Class 1. Features with higher SHAP values contribute more to the predictions made by the model. Positive SHAP values for a feature indicate that the feature is increasing the probability of the model predicting Class 1. Negative SHAP values related to any feature also contribute to decreasing the probability of target class 1 outputted by the model. TT4, T3, FTI: Features positively contribute most to Class 1 prediction. They are probably important predictors of the outcome. Age: It has a moderate impact on predicting Class 1. Referral_source_SVI represents the source of the referral used in the dataset.

Referral_source_other, TSH: These features also have positive but smaller impacts. The remaining features have very little impact on the model's predictions, meaning they are likely less important for the model to decide on the outcome. The diagram illustrates the key features making an impact on the model predictions. It indicates the extent to which various factors influence the model's output.

SHAP values are plotted in Figure 13 and also show feature value effect in Table 12, with the horizontal axis indicating how much each feature was contributing to the output of the model. The color of the dots indicates the feature value. Low values are shown by blue dots, high values by red dots, and all the in-between is a smooth transition. The "violin" SHAP, by the distribution of SHAP values for each feature. It shows how the feature impacts predictions across different instances in the dataset. The SHAP values for TT4 are spread out from −0.1 to 0.8. This indicates that TT4 has a significant impact on the model's output. The majority of the SHAP values are positive, suggesting that higher TT4 values generally increase the model's output. There are also some negative SHAP values, indicating that lower TT4 values can sometimes decrease the model's output. This means that because TT4 has a negative coefficient, the model has identified TT4 as an important predictor of height, and the higher the TT4 value, the higher the predicted value of length in general. This implies that TT4 is an important predictor for the outcome being predicted. Nonetheless, it should be emphasized that the simple relationship is not strictly linear and that TT4 can have deleterious effects on the output at lower values as well.

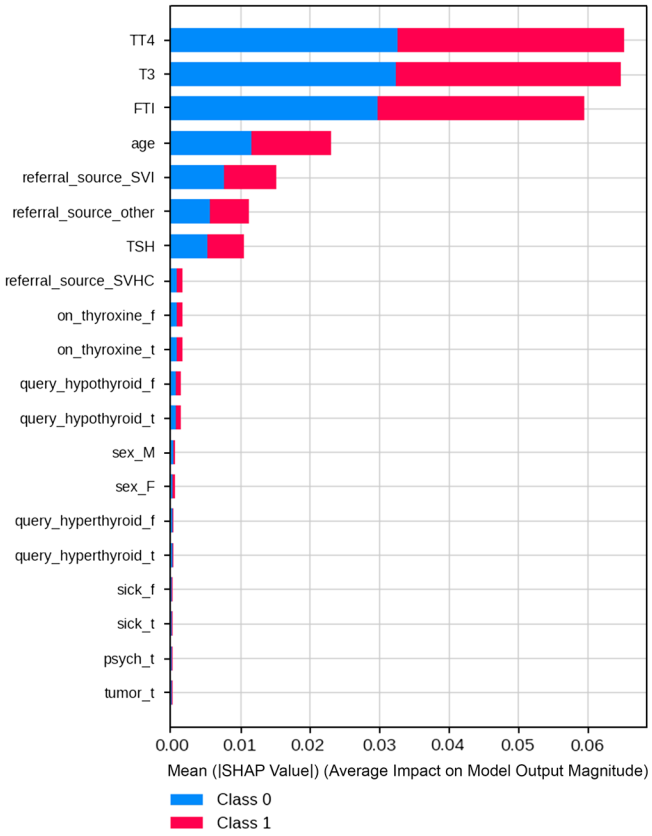
Figure 14 shows the effect of various features on the base value of a prediction. The base value is 0.05214. Higher value: age = 53:

Table 11
Comparison with previously published work

Citation	Data collection place	Methods	Imbalance technique	Accuracy (%)	Dataset size	Features
Peng et al. [8]	Hospital	DT	SMOTE	84.6	47	-
Song et al. [9]	Thyroid cancer patients	Random forest	Not available	89.7	286	17
Liu et al. [10]	Patients treated with PD-1 or PD-L	RT and logistic regression	Not available	77.0	187	7
Kim et al. [11]	Thyroid nodules from surgical C	RT and logistic regression	Not available	71.9	355	-
Luong et al. [12]	Clinical data	DT, random forest, logistic regression	Not available	84.7		-
Yang et al. [13]	Laboratory	RT and logistic regression	Not available	64.0	604	-
Idarraga et al. [14]	Laboratory	Random forest	Not available	71.0–81.0	1,451	-
Xu et al. [15]	Ultrasound images thyroid	Random forest	Histogram equalization, Laplacian operator, logarithm transform, and Gamma correction	95.0	428	Image
Luong et al. [12]	Clinical data	DT, random forest, logistic regression	Not available	84.7		-
Yang et al. [13]	Laboratory	RT and logistic regression	Not available	64.0	604	-
Idarraga et al. [14]	Laboratory	Random forest	Not available	71.0–81.0	1,451	-
Xu et al. [15]	Ultrasound images thyroid	Random forest	Histogram equalization, Laplacian operator, logarithm transform, and Gamma correction	95.0	428	Image
Olatunji et al. [20]	DNA methylation data for follicular thyroid	Random forest	Not available	99.4	92	6
Zhang et al. [21]	MRI data	Random forest	Not available	90.6	60	107
He et al. [22]	Thyroid hormone homeostasis	Multiple algorithms	Not available	82.0/79.0	1,074	-
Garcia de Lomana et al. [23]	Radiomics data from 18F-FDG thyroid gland uptake	Random forest	Not available	78.6	60	18
Aksu et al. [24]	Ultrasonographic	Random forest	Not available	96.1	1,558	Image
Chen et al. [25]	Gut microbiome data	Random forest	Not available	-	92	-
Su et al. [26]	Thyroid cancers	Multiple algorithms	Not available	86.0	177	Image
Zhao et al. [27]	Radiomics data for BRAF mutation prediction	Multiple algorithms	Not available	64.3	96	86
Xi et al. [33]	Novel clinical dataset	Multiple Algorithm	Not available	78	724	19
Książek [34]	Shengjing Hospital of China Medical University	Multiple algorithms	Not available	81.82	1,232	19
Proposed	Thyroid	Multiple algorithms	Yes	96.66	3,773	28

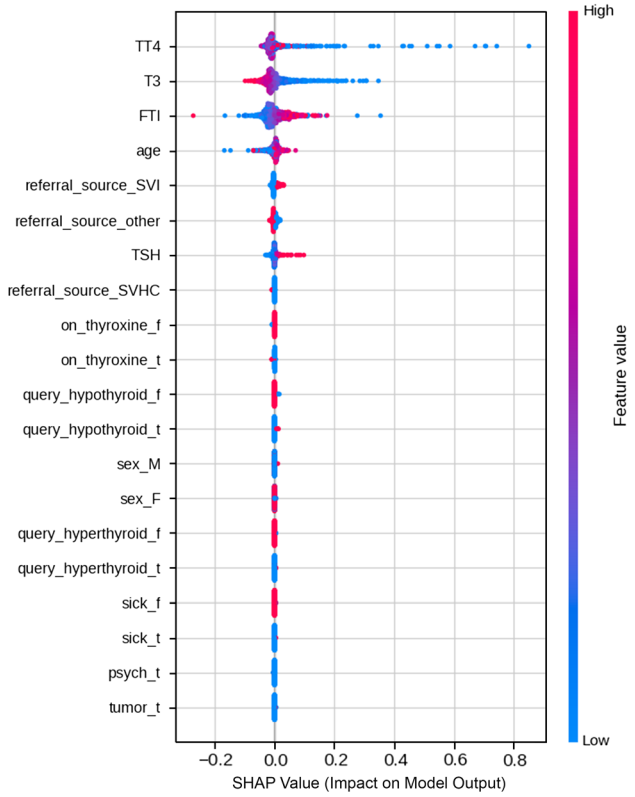
Note: DT = decision tree, KNN = K-nearest neighbors, MRI = magnetic resonance imaging, SMOTE = Synthetic Minority Over-sampling Technique.

Figure 12
Feature importance from SHAP value



Note: SHAP = Shapley Additive Explanation, TSH = thyroid-simulating hormone.

Figure 13
SHAP summary impact of each feature on the model's output



Note: SHAP = Shapley Additive Explanation, TSH = thyroid-simulating hormone.

This feature decreases the base value by 0.03214, which means a higher age leads to a lower prediction. Lower value: TT4 = 103: This feature increases the base value by 0.03014, which means a higher TT4 value leads to a higher prediction. T3 = 2: This feature increases the base value by 0.01014, which means a higher T3 value leads to a higher prediction. Referral_source_other = 1: This feature increases the base value by 0.01014, which means that if the referral source is other, the prediction is higher. referral_source_SVI = 0: This feature increases the base value by 0.01014, which means if the referral source is SVI, the prediction is higher. FTI = 111: This feature increases the base value by 0.00914, which means a higher FTI value leads to a higher prediction. TSH = 0.85: This feature increases the base value by 0.00914, which means a higher TSH value leads to a higher prediction.

LIME XAI in Figure 15 shows the prediction probabilities of a machine-learning model for a patient's condition. The model predicts the patient is more likely to be negative (with a probability of 0.98) than sick (with a probability of 0.02).

The features that model considered the features in making its prediction, along with the corresponding contribution of each feature to the final prediction probability.

Referral_source_SVH: This feature contributes 0.02 to the probability of the patient being negative. referral_source_SVI: This feature contributes 0.02 to the probability of the patient being negative. lithium_t ≤ 0.00: This feature contributes 0.01 to the probability of the patient being negative. 0.00 < referral_source_. This feature contributes 0.01 to the probability of the patient being negative.

1.60 < T3 ≤ 2.00: This feature contributes 0.01 to the probability of the patient being negative. 90.00 < TT4 ≤ 105.00: This feature contributes 0.01 to the probability of the patient being negative.

37.00 < age ≤ 55.00: This feature contributes 0.01 to the probability of the patient being negative. Pregnant_t ≤ 0.00: This feature contributes 0.02 to the probability of the patient being sick. 108.00 < FTI ≤ 125.00: This feature contributes 0.02 to the probability of the patient being sick. Tumor_t > 0.00: This feature contributes 0.01 to the probability of the patient being sick.

In Figure 16, referral_source_SVHD: 0.00, referral_source_SVI: 0.00, pregnant_t: 0.00, FTI: 111.00, lithium_t: 0.00, referral_source_other: 1.00, tumor_t: 1.00, T3: 2.00, TT4: 103.00, age: 53.00. rows with an orange background color highlight features related to the patient's medical history and demographics. These include: pregnant_t, FTI, lithium_t, age. The rows with a blue background color highlight features related to the patient's diagnosis and treatment. These include: referral_source_SVHD, referral_source_SVI, referral_source_other, tumor_t, T3, TT4.

LIME Feature Importance: Figure 17 shows the feature importance of different features in a model, and is briefly explained in Table 13. Green bars: Positive impact on the model output. Higher values in these features are associated with a positive prediction. Red bars: Negative impact on the model output. Higher values in these features are associated with a negative prediction.

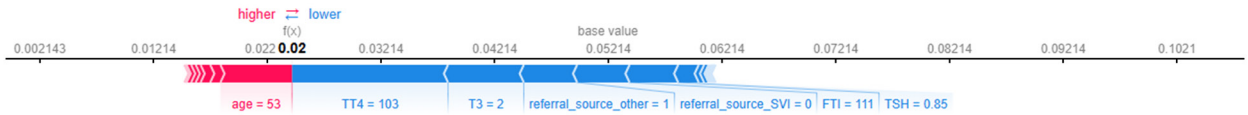
The features and their impact are: referral_source_SVHD ≤ 0.00: Negative impact. referral_source_SVI ≤ 0.00: Negative impact. pregnant_t ≤ 0.00: Positive impact. 108.00 < FTI ≤ 125.00: Positive impact. lithium_t ≤ 0.00: Negative impact. 0.00 < referral_source_other

Table 12
Summary of SHAP value distribution (impact on model output)

Feature name	SHAP value distribution (impact on model output)	Feature value effect
TT4	Highest positive and negative impact spread	Higher TT4 values push the model output higher
T3	High impact, primarily positive SHAP value spread	Elevated T3 increases model output
FTI	Substantial effect, both directions	Higher FTI results in higher model output
age	Moderate, some high and low impact points	High age both increases and decreases output
referral_source_SVI	Low, mostly around zero	Minor effect based on source
referral_source_other	Low, near zero	Minor effect
TSH	Low, close to zero impact	High/Low values produce small changes in output
referral_source_SVHC	Minimal impact	Minor effect
on_thyroxine_f	Minimal	Negligible effect
on_thyroxine_t	Minimal	Negligible effect
query_hypothyroid_f	Very low, around zero	Almost no effect
query_hypothyroid_t	Very low, around zero	Almost no effect
sex_M	Minimal	Negligible effect
sex_F	Minimal	Negligible effect
query_hyperthyroid_f	Minimal	Negligible effect
query_hyperthyroid_t	Minimal	Negligible effect
sick_f	Minimal	Negligible effect
sick_t	Minimal	Negligible effect
psych_t	Minimal	Negligible effect
tumor_t	Minimal	Negligible effect

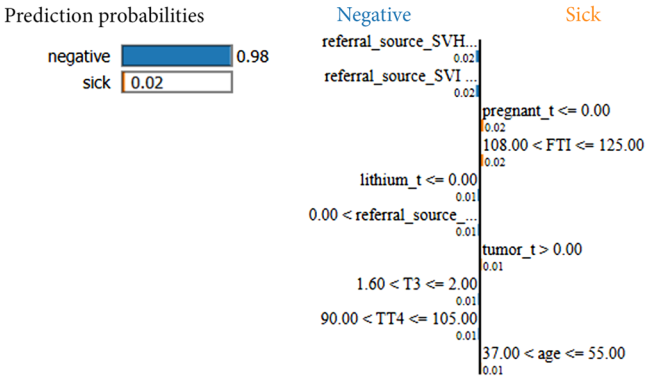
Note: SHAP = Shapley Additive Explanation, TSH = thyroid-simulating hormone.

Figure 14
SHAP contributes to the model's predictive probability for a single instance



Note: SHAP = Shapley Additive Explanation.

Figure 15
LIME feature importance



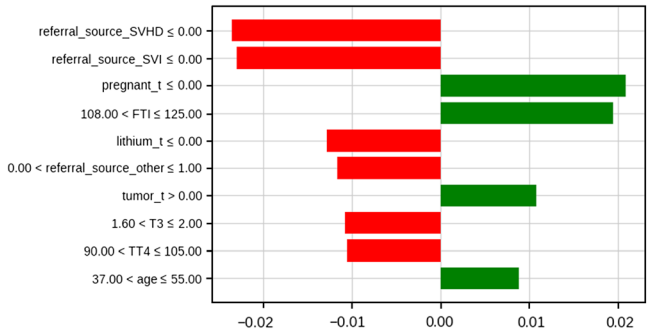
Note: Lime = Local Interpretable Model-Agnostic Explanation.

Figure 16
LIME feature impact summary

Feature	Value
referral_source_SVHD	0.00
referral_source_SVI	0.00
pregnant_t	0.00
FTI	111.00
lithium_t	0.00
referral_source_other	1.00
tumor_t	1.00
T3	2.00
TT4	103.00
age	53.00

Note: Lime = Local Interpretable Model-Agnostic Explanation.

Figure 17
LIME feature importance and prediction factor



Note: Lime = Local Interpretable Model-Agnostic Explanation.

Table 13
Summary of LIME contribution

Feature/rule	Condition or value	Contribution direction	Contribution to class ("sick")
referral_source_SVH...	Yes	Negative (towards "negative")	−0.02
referral_source_SVI...	Yes	Negative	−0.02
pregnant_t ≤ 0.00	True	Negative	−0.02
108.00 < FTI ≤ 125.00	True	Negative	−0.02
lithium_t ≤ 0.00	True	Negative	−0.02
1.60 < T3 ≤ 2.00	True	Negative	−0.01
90.00 < TT4 ≤ 105.00	True	Negative	−0.01
37.00 < age ≤ 55.00	True	Negative	−0.01
tumor_t > 0.00	False/not triggered	Negligible	(Not listed)

Note: Lime = Local Interpretable Model-Agnostic Explanation.

Table 14
Comparison of SHAP and LIME with selected features

Features	SHAP	LIME
TSH	0.85	0.18
T3	2.00	0.01014
TT4	103.00	0.03014
FTI	111.00	0.00914

Note: Lime = Local Interpretable Model-Agnostic Explanation, SHAP = Shapley Additive Explanation.

≤ 1.00: Negative impact. tumor_t > 0.00: Positive impact. 1.60 < T3 ≤ 2.00: Negative impact. 90.00 < TT4 ≤ 105.00: Negative impact. 37.00 < age ≤ 55.00: Positive impact. LIME is a technique that provides local explanations for predictions of complex models. It approximates the behavior of the model in the neighborhood of a specific data point.

Interpretation:

The graph suggests that features such as "pregnant_t," "108.00 < FTI ≤ 125.00," "tumor_t > 0.00," and "37.00 < age ≤ 55.00" have a positive impact on the model's prediction, whereas features such as "referral_source_SVHD," "referral_source_SVI," "lithium_t," "0.00 < referral_source_other ≤ 1.00," "1.60 < T3 ≤ 2.00," and "90.00 < TT4 ≤ 105.00" have a negative impact.

SHAP and LIME comparison:

According to Table 14, two ML explainability methods, SHAP and LIME, evaluate the importance of four thyroid-related features (TSH, T3, TT4, and FTI). SHAP assigns higher importance values—particularly for TT4 and FTI—whereas LIME gives much lower scores across all features. This difference indicates that SHAP captures global feature influence, while LIME emphasizes the impact of features on specific predictions, leading to noticeable variations in their rankings and values.

This research presents a promising AI-based method for the timely detection of thyroid cancer. Our study uses an open-source thyroid disease dataset comprising 3,772 case observations. To face challenges in the presence of class imbalance, the cluster SMOTE technique is used for data balancing, five different ML classifiers were used, and after that, an ensemble learning mechanism is applied to find the best classifier with the help of soft voting, SVM classifier is selected through this learning approach, and the diagnostic accuracy of 0.97 is achieved, which outperforms another current state-of-the-art classifier. In the service of interpretability, SHAP and LIME XAI mechanisms have been employed that render the model's decisions transparent, allowing clinicians to understand how the model came to its predictions. This model can serve as a decision-support tool for clinicians with a future goal of integration into electronic health records. Such a system could help health practitioners become more adept at identifying and diagnosing potential cases early or in high-volume or resource-constrained settings, as an early diagnostic tool. There are boundaries that future studies may address. Although the dataset is sizeable and diverse, it could still include further data to enhance the robustness and generalizability of the model to a wider population. These advancements are expected to improve the performance of the model itself. It allows the model to be applied to a range of different patients, eventually accounting for a higher degree of certainty of the clinical relevance of the model.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination. The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/bhargavchirumamilla/thyroid-cancer-risk-dataset. The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/sikandaraidev/thyroid-dataset.

Author Contribution Statement

Syed Younus Ali: Conceptualization, Software, Data curation, Writing – original draft, Writing – review & editing. **Bilal Shoaib Khan:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Abdul Hanan Khan:** Software, Validation. **Muhammad Adnan**

Khan: Formal analysis, Investigation. **Asghar Ali Shah:** Validation, Investigation. **Sagheer Abbas:** Formal analysis, Project administration. **Khan Muhammad Adnan:** Methodology, Resources.

References

[1] Pace, D. (2024). Thyroid: The butterfly of metabolism: How to prevent, take care of oneself, and stay healthy. Switzerland: Springer. https://doi.org/10.1007/978-3-031-55276-2

[2] de Luca, R., Davis, P. J., Lin, H.-Y., Gionfra, F., Percario, Z. A., Affabris, E., ..., & Incerpi, S. (2021). Thyroid hormones interaction with immune response, inflammation and non-thyroidal illness syndrome. *Frontiers in Cell and Developmental Biology*, *8*, 614030. https://doi.org/10.3389/fcell.2020.614030

[3] Sinkó, R., Mohácsik, P., Kővári, D., Penksza, V., Wittmann, G., Mácsai, L., ..., & Gereben, B. (2023). Different hypothalamic mechanisms control decreased circulating thyroid hormone levels in infection and fasting-induced non-thyroidal illness syndrome in male thyroid hormone action indicator mice. *Thyroid*, *33*(1), 109–118. https://doi.org/10.1089/thy.2022.0404

[4] Muscente, J. (2024). Medical laboratory testing (blood). In I. Khasnabish, T. Chikwinya, & J. Muscente (Eds.), *Complex cases in clinical ophthalmology practice* (pp. 377–388). Springer Nature. https://doi.org/10.1007/978-3-031-73467-0_28

[5] Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, *72*(1), 7–33.

[6] Schneider, S. A., Tschaidse, L., & Reisch, N. (2023). Thyroid disorders and movement disorders—A systematic review. *Movement Disorders Clinical Practice*, *10*(3),360–368. https://doi.org/10.1002/mdc3.13656

[7] Zhang, X., Lee, V. C., Rong, J., Liu, F., & Kong, H. (2022). Multi-channel convolutional neural network architectures for thyroid cancer detection. *PLOS ONE*, *17*(1), e0262128. https://doi.org/10.1371/journal.pone.0262128

[8] Peng, Y., Yang, X., Li, D., Ma, Z., Liu, Z., Bai, X., & Mao, Z. (2025). Predicting flow status of a flexible rectifier using cognitive computing. *Expert Systems with Applications*, *264*, 125878. https://doi.org/10.1016/j.eswa.2024.125878

[9] Song, H., Dong, C., Zhang, X., Wu, W., Chen, C., Ma, B., ..., & Lv, X. (2022). Rapid identification of papillary thyroid carcinoma and papillary microcarcinoma based on serum Raman spectroscopy combined with machine learning models. *Photodiagnosis and Photodynamic Therapy*, *37*, 102647. https://doi.org/10.1016/j.pdpdt.2021.102647

[10] Liu, Y. H., Jin, J., & Liu, Y. J. (2022). Machine learning–based random forest for predicting decreased quality of life in thyroid cancer patients after thyroidectomy. *Support Care Cancer*, *30*(3), 2507–2513. https://doi.org/10.1007/s00520-021-06657-0

[11] Kim, W., Cho, Y.-A., Kim, D.-C., Jo, A.-R., Min, K.-H., & Lee, K.-E. (2021). Factors associated with thyroid-related adverse events in patients receiving PD-1 or PD-L1 inhibitors using machine learning models. *Cancers*, *13*(21), 5465. https://doi.org/10.3390/cancers13215465

[12] Luong, G., Idarraga, A. J., Hsiao, V., & Schneider, D. F. (2022). Risk stratifying indeterminate thyroid nodules with machine learning. *Journal of Surgical Research*, *270*, 214–220. https://doi.org/10.1016/j.jss.2021.09.015

[13] Yang, P. P., Huang, J. P., Wang, Z. D., & Qian, L. X. (2021). A predictive model and survival analysis for local recurrence in differentiated thyroid carcinoma. *Minerva Endocrinology*, *47*(3), 286–294. https://doi.org/10.23736/s2724-6507.21.03393-9

[14] Idarraga, A. J., Luong, G., Hsiao, V., & Schneider, D. F. (2021). False negative rates in benign thyroid nodule diagnosis: Machine learning for detecting malignancy. *Journal of Surgical Research*, *268*, 562–569. https://doi.org/10.1016/j.jss.2021.06.076

[15] Xu, S., Huang, H., Zhang, X., Huang, Y., Guan, B., Qian, J., ..., & Liu, J. (2021). Predictive value of serum thyroglobulin for structural recurrence following lobectomy for papillary thyroid carcinoma. *Thyroid*, *31*(9), 1391–1399. https://doi.org/10.1089/thy.2021.0209

[16] Hang, Y. (2021). Thyroid nodule classification in ultrasound images by fusion of conventional features and Res-GAN deep features. *Journal of Healthcare Engineering*, *2021*(1), 9917538. https://doi.org/10.1155/2021/9917538

[17] Yang, F., Zhang, J., Li, B., Zhao, Z., Liu, Y., Zhao, Z., ..., & Wang, G. (2021). Identification of potential lncRNAs and miRNAs as diagnostic biomarkers for papillary thyroid carcinoma based on machine learning. *International Journal of Endocrinology*, *2021*(1), 3984463. https://doi.org/10.1155/2021/3984463

[18] Qin, H., Que, Q., Lin, P., Li, X., Wang, X.-R., He, Y., ..., & Yang, H. (2021). Magnetic resonance imaging (MRI) radiomics of papillary thyroid cancer (PTC): A comparison of predictive performance of multiple classifiers modeling to identify cervical lymph node metastases before surgery. *La Radiologia Medica*, *126*(10), 1312–1327. https://doi.org/10.1007/s11547-021-01393-1

[19] Sai, P. V., Rajalakshmi, T., & Snekhalatha, U. (2021). Non-invasive thyroid detection based on electroglottogram signal using machine learning classifiers. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, *235*(10), 1128–1145. https://doi.org/10.1177/09544119211028070

[20] Olatunji, S. O., Alotaibi, S., Almutairi, E., Alrabae, Z., Almajid, Y., Altabee, R., ..., & Alhiyafi, J. (2021). Early diagnosis of thyroid cancer diseases using computational intelligence techniques: A case study of a Saudi Arabian dataset. *Computers in Biology and Medicine*, *131*, 104267. https://doi.org/10.1016/j.compbiomed.2021.104267

[21] Zhang, H., Zhang, Z., Liu, X., Duan, H., Xiang, T., He, Q., ..., & Liang, Z. (2021). DNA methylation haplotype block markers efficiently discriminate follicular thyroid carcinoma from follicular adenoma. The *Journal of Clinical Endocrinology & Metabolism*, *106*(4), e1011–e1021. https://doi.org/10.1210/clinem/dgaa950

[22] He, J., Zhang, H., Wang, X., Sun, Z., Ge, Y., Wang, K., ..., & Hu, S. (2021). A pilot study of radionics signature based on parametric MRI for preoperative prediction of extrathyroidal extension in papillary thyroid carcinoma. *Journal of X-Ray Science and Technology*, *29*(1), 171–183. https://doi.org/10.3233/XST-200760

[23] Garcia de Lomana, M., Weber, A. G., Birk, B., Landsiedel, R., Achenbach, J., Schleifer, K.-J., ..., & Kirchmair, J. (2021). In silico models to predict the perturbation of molecular initiating events related to thyroid hormone homeostasis. *Chemical Research in Toxicology*, *34*(2), 396–411. https://doi.org/10.1021/acs.chemrestox.0c00304

[24] Aksu, A., Karahan Şen, N. P., Acar, E., & Çapa Kaya, G. (2020). Evaluating focal 18F-FDG uptake in thyroid gland with radiomics. *Nuclear Medicine and Molecular Imaging*, *54*(5), 241–248. https://doi.org/10.1007/s13139-020-00659-2

[25] Chen, D., Hu, J., Zhu, M., Tang, N., Yang, Y., & Feng, Y. (2020). Diagnosis of thyroid nodules for ultrasonographic characteristics

indicative of malignancy using random forest. *BioData Mining*, *13*(1), 14. https://doi.org/10.1186/s13040-020-00223-w

[26] Su, X., Zhao, Y., Li, Y., Ma, S., & Wang, Z. (2020). Gut dysbiosis is associated with primary hypothyroidism with interaction on gut-thyroid axis. *Clinical Science*, *134*(12), 1521–1535. https://doi.org/10.1042/CS20200475

[27] Zhao, H.-N., Liu, J.-Y., Lin, Q.-Z., He, Y.-S., Luo, H.-H., Peng, Y.-L., & Ma, B.-Y. (2020). Partially cystic thyroid cancer on conventional and elastographic ultrasound: A retrospective study and a machine learning—Assisted system. *Annals of Translational Medicine*, *8*(7), 495. https://doi.org/10.21037/atm.2020.03.211

[28] Kwon, M.-R., Shin, J. H., Park, H., Cho, H., Hahn, S. Y., & Park, K. W. (2020). Radiomics study of thyroid ultrasound for predicting BRAF mutation in papillary thyroid carcinoma: Preliminary results. *American Journal of Neuroradiology*, *41*(4), 700–705. https://doi.org/10.3174/ajnr.A6505

[29] Anu, K. P., & Benifa, J. B. (2022). A comprehensive analysis using neural network-based model for thyroid disease prediction. In *2022 International Conference on Augmented Intelligence and Sustainable Systems*, 72–78. https://doi.org/10.1109/ICAISS55157.2022.10011021

[30] Vivar, G., Mullakaeva, K., Zwergal, A., Navab, N., & Ahmadi, S.-A. (2020). Peri-diagnostic decision support through cost-efficient feature acquisition at test-time. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference*, 572–581. https://doi.org/10.1007/978-3-030-59713-9_55

[31] van Sonsbeek, T., & Worring, M. (2020). Towards automated diagnosis with attentive multi-modal learning using electronic health records and chest x-rays. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures: 10th International Workshop*, 106–114. https://doi.org/10.1007/978-3-030-60946-7_11

[32] Ai, X., Zhuang, J., Wang, Y., Wan, P., & Fu, Y. (2022). Res-Caps: An improved capsule network and its application in ultrasonic image classification of thyroid papillary carcinoma. *Complex & Intelligent Systems*, *8*(3), 1865–1873. https://doi.org/10.1007/s40747-021-00347-4

[33] Xi, N. M., Wang, L., & Yang, C. (2022). Improving the diagnosis of thyroid cancer by machine learning and clinical data. *Scientific Reports*, *12*(1), 11143. https://doi.org/10.1038/s41598-022-15342-z

[34] Książek, W. (2024). Explainable thyroid cancer diagnosis through two-level machine learning optimization with an improved naked mole-rat algorithm. *Cancers*, *16*(24), 4128. https://doi.org/10.3390/cancers16244128

[35] Akter, S., & Mustafa, H. A. (2024). Analysis and interpretability of machine learning models to classify thyroid disease. *PLOS ONE*, *19*(5), e0300670. https://doi.org/10.1371/journal.pone.0300670

[36] Salman, K. & Sonuç, E. (2021). Thyroid disease classification using machine learning algorithms. In *Journal of physics: conference series*, *1963*(1), 012140. https://doi.org/10.1088/1742-6596/1963/1/012140

[37] Alawiyah, T. A., Wibisono, T., & Mulyani, Y. S. (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*, *6*(3), 1035–1045. https://doi.org/10.47709/cnahpc.v6i3.4101

[38] Ozturk, C., Sagir, O., & Vural, U. (2024). Machine learning approaches to predict thyroid cancer recurrence: A comparative study. In *2024 9th International Conference on Computer Science and Engineering*, 1–6. https://doi.org/10.1109/UBMK63289.2024.10773518

[39] Atay, F. F., Yagin, F. H., Colak, C., Elkiran, E. T., Mansuri, N., Ahmad, F., & Ardigò, L. P. (2024). A hybrid machine learning model combining association rule mining and classification algorithms to predict differentiated thyroid cancer recurrence. *Frontiers in Medicine*, *11*, 1461372. https://doi.org/10.3389/fmed.2024.1461372

[40] Vu, T. A., Huyen, N. A., Huy, H. Q., & Huong, P. T. V. (2023). Enhancing thyroid cancer detection through machine learning approach. In *2023 12th International Conference on Control, Automation and Information Sciences*, 188–193. https://doi.org/10.1109/ICCAIS59597.2023.10382297

How to Cite: Ali, S. Y., Khan, B. S., Khan, A. H., Khan, M. A., Shah, A. A., Abbas, S., & Adnan, K. M. (2025). Optimized Ensemble Machine-Learning-Driven Transparent Thyroid Cancer Prediction Using Explainable Artificial Intelligence. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE52026503