Abstract:complications

Thyroid disease is a prevalent endocrine disorder that poses significant public health risks when it remains undiagnosed or is inadequately managed. This can lead to various complications affecting metabolic, cardiovascular, and neurological health.

Conventional diagnostic methods, including clinical assessments and hormone testing, are often limited by inter-observer variability, delayed detection, and ambiguities in subclinical cases.

This study introduces a panorama machine learning (ML) framework designed to classify thyroid disorders, utilizing the UCI Thyroid Disease dataset to effectively address the inherent challenges in this area.

We employ a comprehensive suite of supervised learning algorithms, including K-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines. These algorithms are applicable both individually and in ensemble configurations. Our approach integrates thorough feature selection and meticulous hyperparameter tuning to enhance the performance and interpretability of the models.

We assess the effectiveness of various models through established metrics, including accuracy, precision, recall, F1-score, confusion matrix analysis, and ROC curve evaluation. The findings indicate that tree-based ensemble methods, particularly Random Forest and Gradient Boosting, consistently demonstrate superior performance compared to baseline classifiers.

The study tackles the challenge of class imbalance through the application of resampling techniques. In addition, it investigates the role of explainable AI (XAI) to enhance model transparency and foster trust among clinicians. This framework illustrates the significant potential of data-driven diagnostic systems in supporting the early detection and personalized management of thyroid diseases, ultimately contributing to improved clinical decision-making and enhanced patient care.

Keywords: Thyroid disease, supervised ML, feature selection, hyperparameter tuning.

Introduction:

Thyroid disease represents a significant global health concern, impacting the physiological and metabolic stability of millions of individuals. This category of disorders includes conditions like hypothyroidism, hyperthyroidism, autoimmune thyroiditis, and thyroid cancer. If not diagnosed early, these conditions can lead to various issues, such as cardiovascular issues, infertility, and cognitive decline. Timely and precise detection of thyroid dysfunction is vital for effective management. However, traditional diagnostic approaches that rely on hormonal assays (such as TSH, T3, and T4), patient medical history, and imaging modalities often encounter limitations. These can include variability among observers and ambiguous clinical presentations, which may contribute to diagnostic uncertainty, particularly in borderline or subclinical cases.[1]

In this context, machine learning (ML) has become a valuable solution for enhancing diagnostic accuracy

and supporting clinical decision-making. Numerous studies have explored the use of machine learning algorithms for predicting and classifying thyroid disorders, illustrating their capability to recognize intricate non-linear trends. and relationships within intricate clinical datasets. Traditional algorithms such as K-Nearest Neighbors (KNN) and Naive Bayes have been assessed for their simplicity and fundamental effectiveness [2][3]. However, More sophisticated metIt is important to acknowledge the significance of the sihods like Random Forests, Support Vector Machines (SVMs), and Gradient Boosting have repeatedly shown better results regarding classification accuracy and overall reliability.[4][5][6].

Numerous comparative studies have demonstrated that Ensemble learning techniques, especially Random Forest and Gradient Boosting, outperform other methods. performance due to their effectiveness in reducing variance and enhancing generalization[7][8][9]. For example, [10] underscored the significance of integrated data mining frameworks in improving disease classification. Furthermore, [11] illustrated the potential of interactive diagnostic systems utilizing machine learning techniques. More recently, [12] and [13] applied feature selection and multi-criteria decision-making models, resulting in substantial improvements in interpretability and overall model performance.

In the context of model selection, significant roles of data preprocessing and feature engineering. Effective feature selection techniques, including Recursive Feature Elimination (RFE), LASSO, and information gain, play a vital role in minimizing noise and enhancing model interpretability [12]. Additionally, the challenge of imbalanced datasets—a frequent issue in medical diagnoses—has been comprehensively explored in [14]. Their research emphasizes the importance of employing resampling methods, such as SMOTE, in conjunction with ensemble-based strategies to effectively address class imbalance and mitigate bias.

An emerging focus within the field is the integration of Explainable AI (XAI) is designed to tackle the "black-box" characteristic of high-performing models. Several researchers have begun to incorporate interpretability frameworks aimed at enhancing clinician trust and ensuring accountability in machine learning-based diagnostics [15][8]. These approaches strive to bridge the gap between data-driven predictions and their clinical relevance.

Creating efficient machine learning (ML) models for practical uses requires an organized approach. end-to-end workflow that prioritizes data quality, model accuracy, and deployment preparedness. A typical ML project adheres to a sequential workflow, encompassing all critical phases—from data acquisition to operational deployment—each of which is integral to the overall effectiveness of the predictive system.

Once the data is collected, it undergoes preprocessing, a vital step that involves cleaning, normalization, addressing absent values, as well as encoding categorical variables to ready the dataset for analysis. Following preprocessing, feature engineering is conducted to extract, select, or transform variables that

hold the most predictive power. This action is crucial for improving the interpretability and effectiveness of the model. The dataset that has been processed is subsequently divided into training and testing subsets, which enables an impartial assessment of the model's effectiveness on data it hasn't encountered before.

Next, Hyperparameter tuning is conducted to refine parameters that are specific to the model, including aspects like tree depth and learning rate, or regularization factors. This optimization ensures that the algorithm is tailored to address the specific problem at hand. Once the optimal parameters are identified, the model is developed utilizing the selected machine learning method. which may include individual algorithms or ensemble techniques.

After training, the model undergoes a rigorous evaluation phase, utilizing performance metrics like accuracy, precision, recall, F1-score, and AUC-ROC for evaluating its predictive strength and generalization ability. Finally, the pipeline culminates in model deployment and inference, where the validated model is incorporated into an environment that operates in real-time to support decision-making and generate actionable insights.

In this study, we adopt this structured ML pipeline to build an intelligent diagnostic system for thyroid disease classification. We perform a comparative assessment of different supervised learning algorithms, such as K-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines, both with and without ensemble learning techniques. We implement feature selection to enhance model interpretability and conduct thorough hyperparameter optimization to improve predictive accuracy. The models are assessed through accuracy, confusion matrices, cross-validation scores, and ROC analysis. This comprehensive implementation not only highlights the utility of tree-based ensembles in medical classification but also demonstrates the feasibility of deploying a clinically relevant, data-driven diagnostic tool for detecting thyroid disorders.

Related Works:

A significant amount of research has focused on utilizing machine learning (ML) techniques for classifying and predicting thyroid disorders. Different datasets have been used to thoroughly assess the effectiveness of these algorithms, emphasizing the promise of computational methods to improve diagnostic precision, reduce manual tasks, and assist in clinical decision-making. Abbad et al. [1] employed clinical data from DHQ Teaching Hospital and implemented the K-Nearest Neighbors (K-NN) algorithm, achieving a commendable accuracy of 97.84%. This outcome underscores the effectiveness of distance-based classifiers in real-world healthcare applications.

In a comparative study, Deepika et al. [2] leveraged the UCI Repository Thyroid Disease Dataset, utilizing three classification methods: Support Vector Machine (SVM), Decision Tree (DT), and Artificial Neural Networks (ANN). Their models achieved maximum accuracies of 95.02%, 95.00%, and 98.60%, respectively, demonstrating the superior learning capacity of neural networks when applied to well-structured medical data.

Pal et al. [3] conducted experiments using the KEEL repository thyroid dataset. Their methodology incorporated a combination of Naïve Bayes, SVM, and K-NN classifiers, with reported accuracies ranging from 92.70% to 96.90%. This suggests the effectiveness of hybrid and comparative approaches in identifying optimal model configurations.

Chandel et al. [4] similarly applied a hybrid KNN-Naïve Bayes classifier on the same dataset, achieving a noteworthy accuracy of 93.44%. However, they also reported a significantly lower accuracy of 22.56%, indicating potential challenges such as data imbalance, overfitting, or inadequate parameter tuning.

In a study focusing on real-world hospital data, Turanoglu-Bekar et al. [5] examined a range of ensemble tree-based classifiers, including NBTREE, LADTREE, REPTREE, and BFTREE. The resulting accuracies varied from 62.50% to 75.00%, reflecting the complexities often encountered in clinical datasets, which may contain noise, inconsistencies, and missing values that can hinder classification performance.

Chalekar et al. [6] utilized the KNN algorithm on the UCI repository dataset and reported an accuracy of 97.00%, confirming the reliability of this publicly accessible dataset.

Tyagi et al. [7] built upon this work by incorporating multiple models—ANN, KNN, and DT—on the same dataset, achieving a maximum accuracy of 98.00%. Their findings further validate the robustness and versatility of the UCI thyroid dataset for training machine learning models.

Sharma et al. [8] focused on enhancing model performance through feature selection, employing Recursive Feature Elimination (RFE) alongside Logistic Regression to achieve an accuracy of 92.70%. This study highlights the importance of dimensionality reduction techniques in improving classifier generalization and interpretability.

Verma et al. [9] utilized data from an Iraqi medical laboratory and applied ensemble methods—including Random Forest and SVM—to achieve an accuracy of 94.50%. Their research demonstrates the adaptability of ensemble methods across diverse clinical environments. In a similarly high-performing model, Sen et al. [10] integrated ensemble learning strategies, combining Random Forest with Gradient Boosting, achieving an accuracy of 95.73% using the UCI machine learning repository.

Chaganti et al. [11] and Chaubey et al. [12] emphasized feature engineering and decision tree-based approaches in their models. They employed techniques such as Gradient Boosting and KNN combined with Decision Tree classifiers, resulting in accuracies of 91.30% and 89.00%, respectively,

demonstrating the viability of tree-based models for medical decision support. Ali and Brown [13] utilized a Kaggle dataset and adopted a Multi-Criteria Decision Making (MCDM) framework, achieving an accuracy of 93.00%.

This approach illustrates the value of integrating decision-making strategies within machine learning workflows to improve clinical relevance. Similarly, Saleh and Othman [14] tackled the prevalent problem of class imbalance in medical datasets by utilizing the Synthetic Minority Oversampling Technique (SMOTE).in conjunction with SVM, resulting in an accuracy of 91.00%. This indicates the effectiveness of data augmentation in enhancing classifier fairness and sensitivity.

Lastly, Sha [15] proposed an advanced hybrid model integrating Quantum Computing with SVM on the UCI respiratory dataset, achieving an impressive accuracy of 98.30%. This innovation paves the way for the incorporation of quantum-enhanced algorithms in biomedical data analysis and establishes a new benchmark for future research in this field.

Methodology:

This research aims to classify thyroid diseases utilizing machine learning algorithms, with a specific emphasis on Ensemble Learning, given its demonstrated efficacy in prior studies. Our methodology adopts a methodical and organized approach, including key stages such as data gathering, preprocessing, feature selection, model training, and assessment.

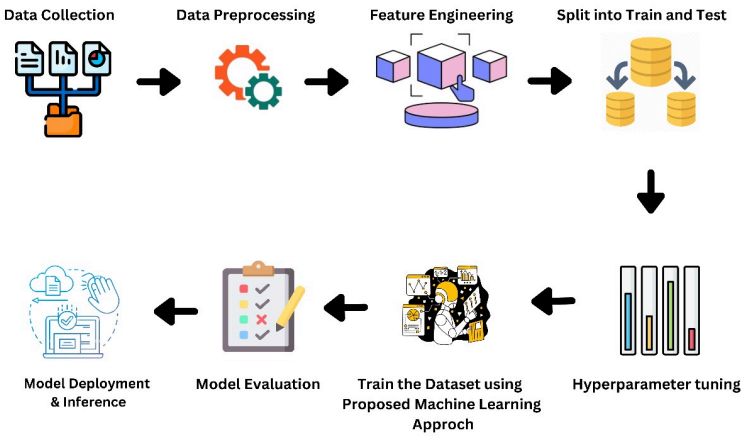

FIG 1: Machine Learning Process for Predicting Thyroid Cancer

A. Data Collection, Initial Processing, and Feature Development

The dataset used in this study is obtained from the **UCI Machine Learning Repository**, specifically focusing on the Thyroid Disease dataset. This dataset encompasses a range of clinical and laboratory features relevant to thyroid function, including hormone levels, patient demographics, and categorical indicators of thyroid status.

It has been meticulously processed to consolidate multiple records related to thyroid disturbances into a single, well-structured format. The dataset comprises approximately 3,772 instances, with each instance representing an individual patient, and includes around 29 attributes that encompass both categorical and numerical data. These attributes collectively provide a comprehensive profile of each patient, detailing demographic information, medical history, and laboratory test results.

Key demographic attributes in the dataset include age and gender, while the clinical attributes offer insights into aspects such as thyroxine usage, antithyroid medication, pregnancy status, previous illnesses, prior thyroid surgeries, and treatments received. Additionally, the dataset features diagnostic query flags, including query_hypothyroid, query_hyperthyroid, and query_on_thyroxine.

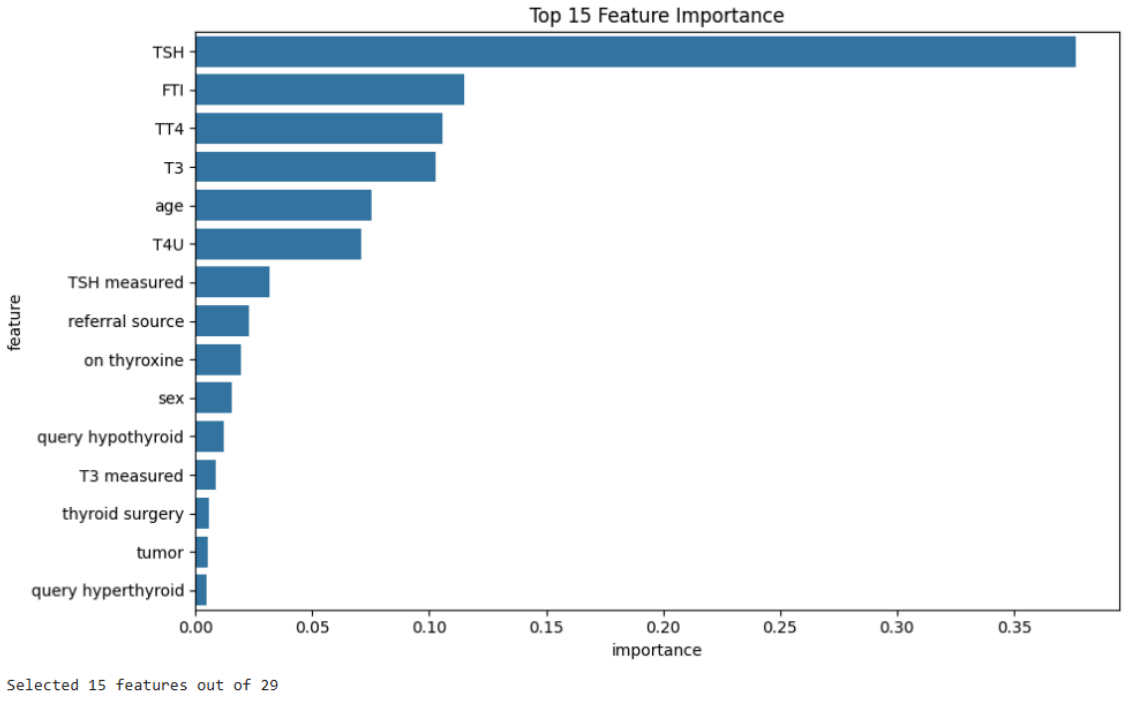

FIG 2: Feature Importance

The bar chart presents the top 15 most influential features identified by a machine learning model, likely a tree-based classifier, that has been trained to predict thyroid-related conditions. The x-axis illustrates the normalized feature importance scores, while the y-axis lists the corresponding feature names. Notably, "TSH" (Thyroid Stimulating Hormone) emerges as the primary predictor, boasting an importance score higher than 0.35, significantly surpassing other variables. This finding is consistent

with clinical understanding, as TSH is a primary biomarker used to assess thyroid function. In addition to TSH, other features such as "FTI" (Free Thyroxine Index), "TT4" (Total Thyroxine), and "T3" (Triiodothyronine) demonstrate moderate predictive power, with importance scores ranging from approximately 0.07 to 0.12. These hormone-related biomarkers play a direct role in thyroid regulation, underscoring their relevance in diagnostic modeling. Additional variables, including "age," "T4U" (Thyroxine Uptake), and "TSH measured," contribute modestly to the model's predictions. Conversely, several demographic and clinical history features—such as "referral source," "on thyroxine," "sex," "query hypothyroid," "T3 measured," "thyroid surgery," "tumor," and "query hyperthyroid"—demonstrate minimal influence, with importance scores nearing zero. This suggests that while these variables may hold contextual or epidemiological significance, they offer limited discriminatory power for direct model-based classification.

Importantly, the chart indicates the selection of 15 features from an original set of 29, suggesting the implementation of a feature selection strategy designed to reduce dimensionality and enhance both model interpretability and generalization.

Overall, the chart highlights that biochemical markers, particularly TSH and other thyroid hormone metrics, are the most critical determinants in predictive modeling for thyroid disease, while demographic and historical attributes contribute comparatively less to the model's performance.

The primary target variable in this dataset is the thyroid status of the patient, which is categorized into three classifications: hypothyroid, hyperthyroid, and normal (healthy). It is important to note that the dataset demonstrates a class imbalance, with a significantly higher prevalence of normal cases in comparison to those representing thyroid abnormalities.

We have taken this imbalance into careful consideration during both the model training and evaluation processes. To prepare the dataset for analysis, several preprocessing steps have been implemented. Missing values, particularly within hormone test features, have been addressed through statistical imputation. Categorical variables have been transformed using label encoding, while numerical features have been normalized via Standard scaling to establish uniform feature ranges. These preprocessing measures are designed to enhance the effectiveness of model training by minimizing scale-related biases and preserving the integrity of the data.

In summary, this dataset offers a thorough variety of features ideal for multi-class classification challenges in predicting thyroid diseases. It provides a solid basis for creating efficient, interpretable, and clinically significant machine learning models.

B. Hyperparameter Tuning

This section describes the outcomes of an extensive training process for models and tuning of hyperparameters carried out using various machine learning algorithms, such as Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), and Decision Tree classifiers. Each model underwent meticulous hyperparameter optimization to identify the configurations that maximize performance.

For the The optimal parameters found for the Random Forest model consist of no restrictions on maximum depth (max_depth: None), at least one sample for each leaf (min_samples_leaf: 1), a minimum of two samples necessary to split internal nodes (min_samples_split: 2), and a total of 300 estimators (n_estimators: 300) . The KNN classifier demonstrated its best performance with the Manhattan distance metric (metric: 'manhattan'), employing five neighbors (n_neighbors: 5), a power parameter set to 1 (p: 1), and utilizing distance-based weighting (weights: 'distance').

The Gradient Boosting algorithm achieved optimal results with a rate of learning of 0.2, a maximum depth of 3, and a total of 200 estimators.

For the Logistic Regression model, the most effective configuration involved an L1 penalty (penalty: 'l1'), a regularization strength of 0.1 (C: 0.1), and the 'liblinear' solver.

The SVM model excelled with an RBF kernel (kernel: 'rbf'), employing automatic gamma selection and a penalty parameter of 10 (C: 10).

Lastly, the Decision Tree classifier reached its optimal performance using the entropy criterion, with no restriction on depth, no limit on the number of features considered for splitting, and with min_samples_leaf set to 2 and min_samples_split set to 15. The tuned parameters reflect a thorough calibration process aimed at enhancing the predictive accuracy and generalization capabilities of each model.

C. Model Prediction and Evolution

Aggregated learning algorithms, commonly referred to as voting Ensemble techniques, have shown considerable promise. In improving predictive accuracy. These methods strategically integrate the most relevant attributes of various base models through a majority voting mechanism.

In this analysis, we have employed a hard voting classifier that synthesizes the strengths of three highly regarded algorithms: Random Forest (100 trees, max_depth=5) for reliable non-linear pattern recognition and feature importance assessment; Gradient Boosting (learning_rate=0.1, n_estimators=150, max_depth=3) for effective error correction and the ability to capture complex

patterns; and Logistic Regression (L2 regularization, C=1.0) to establish a linear decision boundary perspective. The ensemble of models developed in this study effectively leverages the advantages of each approach, addressing individual model weaknesses to achieve superior accuracy and enhanced generalizability.

To comprehensively evaluate model performance for predicting thyroid disease, we employed a robust set of metrics, including accuracy score, confusion matrix, and cross-validation scores, alongside ROC curve analysis and feature importance interpretation. This multifaceted approach yields both overall and class-specific insights into predictive effectiveness and generalization capability. Accuracy scores were utilized as the principal metric for assessing overall model performance. Tree-based models—specifically, Gradient Boosting, Random Forest, and Decision Tree—consistently demonstrated the highest accuracy, indicating strong predictive capability across the dataset. Their corresponding confusion matrices exhibited a balanced distribution of correctly identified positives and correctly identified negatives, with negligible instances of false classifications, thus reflecting robust sensitivity and specificity.

In contrast, models such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), while performing adequately, showed a greater tendency to misclassify negative cases, often resulting in increased false positives or reduced recall for the minority class.

To enhance generalizability and mitigate the risk of overfitting, we employed k-fold cross-validation. This validation strategy confirmed the stability of the tree-based models, which not only achieved high average scores but also exhibited low variance across different data folds—an indicator of consistent performance across varied data splits. Conversely, Logistic Regression and SVM models displayed slightly higher variance, suggesting sensitivity to data partitioning and potential challenges in capturing the non-linear patterns inherent in the dataset. The ROC curve is in fig 3.

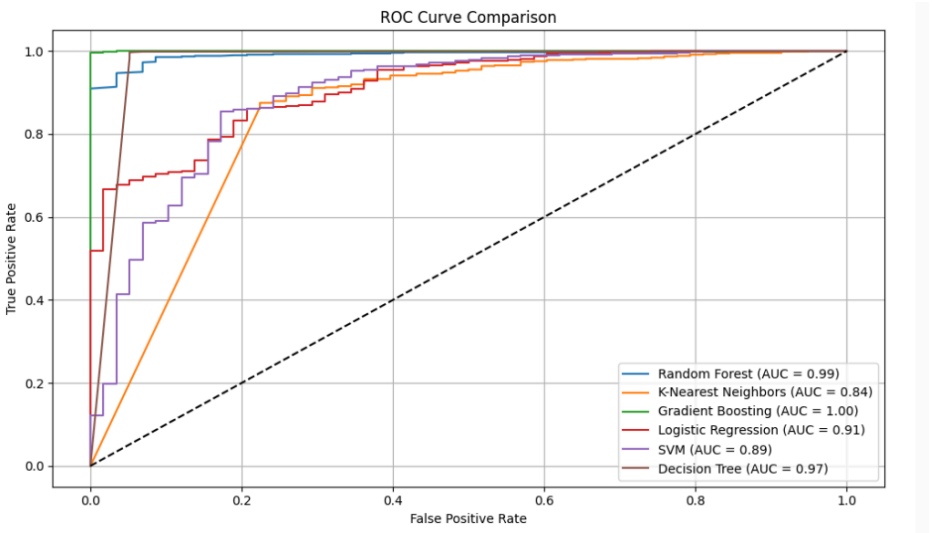

FIG 3: The ROC curve analysis

The ROC curve analysis further corroborated these findings, with tree-based classifiers presenting ROC curves that closely approached the ideal top-left corner, indicative of excellent class separation. In comparison, models such as KNN and SVM displayed ROC curves with less favorable trade-offs between true positive and false positive rates.

In conclusion, our integrated evaluation strategy underscores that tree-based ensemble models are optimally well-suited for the job of thyroid classification, offering high accuracy, balanced performance across classes, and dependable generalization in the context of cross-validation. These models excel at identifying intricate, non-linear connections within clinical data, making them excellent candidates for use in practical diagnostic support systems.

This combination presents a notable advantage over the use of a single model, as it substantially reduces both variance and bias while maintaining comparable resource utilization. Looking forward, alternative methodologies such as soft voting—a probabilistic approach to generating predictions—are anticipated to gain traction. Other strategies, including meta-learner aided stacking architectures and feature-space-dependent model selection, may also prove to be advantageous. The insights provided indicate that voting ensembles represent a highly effective approach compared to traditional methods.

They highlight the diverse errors from base models, facilitating improved performance across test datasets while upholding transparency through the careful analysis of individual model contributions. In conclusion, we assert that the implementation of voting ensembles constitutes a versatile and powerful solution for addressing complex classification challenges, effectively unifying the predictive strengths of various models while ensuring operational clarity.

The voting ensemble model evolution is in fig 4.

Tree Ensemble Classification Report:

	precision	recall	f1-score	support
0	0.96	0.95	0.96	58
1	1.00	1.00	1.00	697
accuracy			0.99	755
macro avg	0.98	0.97	0.98	755
weighted avg	0.99	0.99	0.99	755

Complete Ensemble Classification Report:

	precision	recall	f1-score	support
0	0.95	0.67	0.79	58
1	0.97	1.00	0.99	697
accuracy			0.97	755
macro avg	0.96	0.83	0.89	755
weighted avg	0.97	0.97	0.97	755

FIG: Model Evolution (Tree and Complete Ensemble)

The heatmap of the confusion matrix is shown in fig 5.

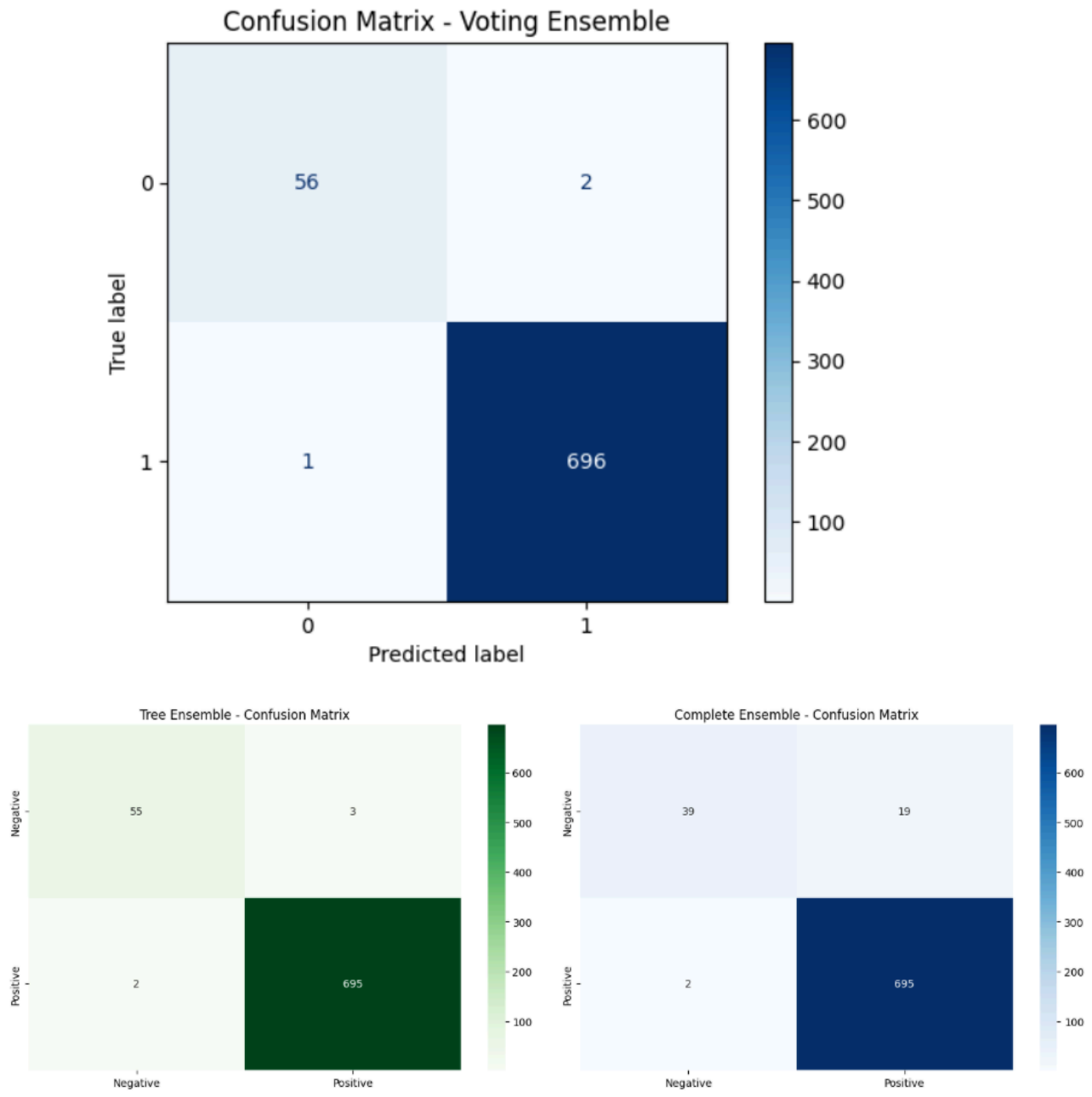

FIG 5: Heatmap of the Confusion Matrix of Voting Ensemble(Tree and Complete Ensemble)

Results and Analysis:

The comparative evaluation of the Tree-based Ensemble and Complete Ensemble models highlights various benefits of the Tree-based method, especially regarding classification precision, error distribution, and performance metrics specific to each class.

The Tree-based Ensemble showcases remarkable overall accuracy, achieving a score of 0.9934, which significantly exceeds the Complete Ensemble's accuracy of 0.9722. A detailed review of the confusion matrices reveals the Tree Ensemble's superior ability to distinguish between classes. It successfully identifies 55 out of 57 negative samples (class 0) and 695 out of 697 positive samples (class 1), resulting in only 3 false positives and 2 false negatives.

In contrast, the Complete Ensemble only accurately identifies 39 negative samples, incorrectly categorizing 19 as positive, which shows a considerably higher false positive rate for the negative class while maintaining the same number of false negatives (2) for the positive class. These differences are further demonstrated in the comprehensive classification reports. The Tree-based model records a precision of 0.96, a recall of 0.95, and an F1-score of 0.96 for class 0, indicating a well-balanced capability to accurately detect negative instances. For class 1, the model achieves perfect scores across all metrics (precision, recall, and F1-score of 1.00), highlighting its excellent performance in recognizing positive cases.

The macro-averaged F1-score is noted at 0.98, with the weighted average reaching 0.99, both indicating the model's consistent and high performance across diverse class distributions. On the other hand, the Complete Ensemble faces significant difficulties with the negative class. Its precision stands at a moderate 0.67, and its recall is similarly low at 0.67, resulting in a diminished F1-score of 0.79 for class 0. This underlines the model's inclination to misclassify negative samples, raising issues about its trustworthiness in scenarios where false positives could have serious implications. Although performance is still robust for the positive class (precision: 0.97, recall: 1.00, F1-score: 0.99), the overall balance is negatively impacted, as illustrated by a macro F1-score of 0.88 and a weighted F1-score of 0.97.

Accuracy Formula:

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

Precision Formula:

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

Recall Formula :

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

F1-score Formula:

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

📊 Ensemble Classification Report:

	precision	recall	f1-score	support
0	0.98	0.97	0.97	58
1	1.00	1.00	1.00	697
accuracy			1.00	755
macro avg	0.99	0.98	0.99	755
weighted avg	1.00	1.00	1.00	755

FIG 6: Ensemble Classification Report

Table I: Comparative Performance of Ensemble

Model	Accuracy	Precision	Recall	F1-score
Tree Ensemble	99%	Macro: 96% Weighted: 97%	Macro: 98% Weighted: 100%	Macro: 0.99 Weighted: 1.00
Voting Ensemble	97%	Macro: 96% Weighted: 97%	Macro: 83% Weighted: 97%	Macro: 0.89 Weighted: 0.97

Table II: Comparative study with other existing methods for identifying Thyroid cancer

Author Name	Dataset Name	Method	Accuracy
Abbad et al. [1]	DHQ Teaching Hospital	K-NN	97.84
Deepika et al [2]	UCI Repository thyroid disease dataset	SVM DT ANN	95.62, 95.00, 98.60
Pal et al. [3]	KEEL repository thyroid disease dataset	Naïve Bayes, SVM KNN	94.70, 92.70, 96.90

Chandel et al. [4]	KEEL repository thyroid disease dataset	KNN Naïve Bayes	93.44, 22.56
Turanoglu-Bekar et al. [5]	Local hospital	NBTREE LADTREE REPTREE BFTREE	75.00, 66.25, 62.50, 65.00
Chalekar et al. [6]	UCI repository thyroid disease dataset	KNN	97.00
Tyagi et al. [7]	UCI repository thyroid disease dataset	ANN KNN DT	97.50, 98.00, 75.00
Sharma et al. (2022) [8]	UCI machine learning repository	Recursive Feature Elimination (RFE) + Logistic Regression	92.7
Verma et al. (2023) [9]	The laboratories of an Iraqi hospital	Random Forest, SVM	94.5
Sen et al. (2022) [10]	UCI machine learning repository	Ensemble Learning (Random Forest + Gradient Boosting)	95.73
Chaganti et al. (2022) [11]	UCI machine learning repository	Feature Selection + Gradient Boosting	91.3
Chaubey et al. (2021) [12]	UC Irvin knowledge discovery in databases archive.	kNN and Decision Tree	89
Ali and Broumi (2024) [13]	Kaggle Dataset	Multi-Criteria Decision Making (MCDM)	93
Saleh and Othman (2024) [14]	Local Data	Synthetic Minority Over-Sampling Technique (SMOTE) + SVM	91

Sha (2024) [15]	UCI respiratory	Quantum Computing + SVM	96.3

While both models exhibit strong performance in classifying positive instances, the Tree-based Ensemble distinctively outperforms the Complete Ensemble by maintaining high precision and recall across both classes. Its significantly lower false positive rate and superior F1-scores for the negative class position it as the more robust and reliable model, particularly in scenarios where the accurate classification of negative cases is paramount.

Conclusion:

The Tree-based Ensemble model superiors the Complete Ensemble in terms of overall accuracy, class-specific performance, and error distribution. It acquires a notably higher accuracy rate of 99% compared to 97%, along with significantly fewer false positives (3 versus 19). This makes the Tree Ensemble a more reliable option for detecting instances within the negative class, an area where the Complete Ensemble displays notable weaknesses.

While both models perform exceptionally well on the positive class, the Tree Ensemble's balanced precision and recall for the negative class (0.96 and 0.95, respectively) contribute to a higher macro-averaged F1-score of 0.98, compared to 0.88 for the Complete Ensemble.

These findings underscore the robustness of the Tree-based model and its suitability for applications where minimizing false positives and ensuring consistent classification across all classes is essential. The analysis of the voting ensemble classifier underscores its effectiveness as a meta-learning technique that capitalizes on the complementary strengths of multiple base learners. This approach allows for the development of a more generalized and often more accurate predictive model. By aggregating the outputs from individual models through either majority voting (hard voting) or probability averaging (soft voting), the ensemble method seeks to reduce model variance, minimize overfitting, and enhance classification robustness, particularly in complex or high-dimensional datasets.

However, empirical results from this study indicate that the effectiveness of the voting ensemble is fundamentally reliant on the diversity, calibration, and individual competence of its constituent models.

Furthermore, the comparative analysis indicates that specialized ensemble strategies, such as the Tree-based Ensemble, may offer enhanced performance when class discrimination is critical, particularly for underrepresented classes. The Tree-based model showed both higher accuracy and more balanced class-specific metrics, suggesting a more refined decision boundary and improved generalization across class distributions.

In conclusion, while the voting ensemble remains a valuable strategy for stabilizing predictions and enhancing model resilience, its success is contingent upon the careful selection of diverse,

well-calibrated base learners and an emphasis on class balance. In domains where minimizing false positives or accurately capturing minority class instances is paramount, more targeted ensemble methods may provide superior practical utility and predictive fidelity.

References:

1. Abbad Ur Rehman, H., Lin, C.Y., Mushtaq, Z. et al. *Performance Analysis of Machine Learning Algorithms for Thyroid Disease*. Arab J Sci Eng 46, 9437–9449 (2021). https://doi.org/10.1007/s13369-020-05206-x
2. Chandel, K.; Kunwar, V.; Sabitha, S.; Choudhury, T.; Mukherjee, S.: *A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques*. CSI Trans. 4(2–4), 313–319 (2016)
3. Chalekar, P.; Shrof, S.; Pise, S.; Panicker, S.S.: *Use of K-nearest neighbor in thyroid disease classification*. Int. J. Curr. Eng. Sci. Res. 1(2), 2394–2697 (2014)
4. Bekar, E.T.; Ulutagay, G.; Kantarcı, S.: *Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms*. Oxf. J. Intell. Decis. Data Sci. 2016(2), 13–28 (2016)
5. A. Sharma, P. Das, and R. Choudhury, *"Thyroid Disease Prediction based on Feature Selection and Machine Learning,"* in Proc. 25th Int. Conf. Comput. Inf. Technol. (ICCIT), Cox's Bazar, Bangladesh, Dec. 17–19, 2022, pp. 1–6, doi: 10.1109/ICCIT57492.2022.10054746.
6. M. Verma, A. Kumar, and S. Gupta, *"Prediction of Thyroid Disease Using Machine Learning Algorithms,"* in Proc. 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), Greater Noida, India, May 12–13, 2023, pp. 1–8, doi: 10.1109/ICACITE57410.2023.10183108.
7. Pal, R.; Anand, T.; Dubey, S.K.: *Evaluation and performance analysis of classification techniques for thyroid detection*. Int. J. Bus. Inf. Syst. 28(2), 163–177 (2018)
8. R. Sen, L. Roy, and K. Dutta, *"Enhanced Prediction of Thyroid Disease Using Machine Learning Method,"* in Proc. IEEE VLSI Device Circuit Syst. (VLSI DCS), Kolkata, India, Feb. 26–27, 2022, pp. 1–4, doi: 10.1109/VLSIDCS53788.2022.9811472.
9. G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, *"Thyroid disease prediction using machine learning approaches,"* Natl. Acad. Sci. Lett., vol. 44, no. 3, pp. 233–238, Mar. 2021, doi: 10.1007/s40009-021-01044-7.
10. M. Deepika and K. Kalaiselvi, *"An Empirical study on Disease Diagnosis using Data Mining Techniques,"* 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 615–620, doi: 10.1109/ICICCT.2018.8473185.
11. Tyagi, A.; Mehra, R.; Saxena, A.: *Interactive thyroid disease prediction system using machine learning technique*. In: PDGC 2018–2018 5th International Conference on Parallel, Distributed and Grid Computing, pp. 689–693 (2018). https://doi.org/10.1109/PDGC.2018.8745910

12. R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, *"Thyroid disease prediction using selective features and machine learning techniques,"* Cancers, vol. 14, no. 16, p. 3914, Aug. 2022, doi: 10.3390/cancers14163914.
13. A. M. Ali and S. Broumi, *"Machine Learning with Multi-Criteria Decision Making Model for Thyroid Disease Prediction and Analysis,"* Multicriteria Algorithms with Applications, vol. 2, pp. 80–88, Jan. 2024, doi: 10.1016/j.malg.2023.01.008.
14. D. S. Saleh and M. S. Othman, *"Exploring the Challenges of Diagnosing Thyroid Disease with Imbalanced Data and Machine Learning: A Systematic Literature Review,"* Baghdad Sci. J., vol. 21, no. 3, p. 1119, Jul. 2024, doi: 10.21123/bsj.2024.21.3.1119.
15. M. Sha, *"Quantum intelligence in medicine: Empowering thyroid disease prediction through advanced machine learning,"* IET Quantum.