Ensemble-Based Machine Learning Approach for Early Detection of Thyroid Cancer

***Abstract*—Thyroid disease ranks among the most prevalent endocrine disorders and poses significant public health concerns when not diagnosed early or adequately managed. If left untreated, the condition may lead to serious metabolic, cardiovascular, and neurological complications. Traditional diagnostic methods, relying on clinical assessments and hormonal tests, often encounter challenges such as inter-observer variability, diagnostic delays, and difficulty in detecting subclinical cases. To address these limitations, this study introduces a comprehensive machine learning (ML) framework for the classification of thyroid disorders, leveraging the UCI Thyroid Disease dataset. The modeling approach is refined by rigorous feature selection and hyperparameter optimization to enhance predictive accuracy and interpretability. Model performance is assessed using key metrics, including accuracy, precision, recall, F1-score, and confusion matrices. A diverse set of supervised learning algorithms, K-Nearest Neighbors, Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines, were employed both individually as standalone models and within ensemble configurations. A soft voting ensemble that integrates Gradient Boosting, Random Forest, and Decision Tree achieved a superior accuracy of 99.60%, outperforming individual models by ensuring better balance, particularly for minority class detection. The dataset's original class imbalance was preserved to maintain clinical relevance.**

***Index Terms*—Thyroid disease, feature selection, hyperparameter tuning, grid search, hard-voting, cross-validation.**

I. Introduction

Thyroid disease stands as one of the major global health concerns, affecting millions of individuals and disrupting metabolic and physiological processes. Disorders such as hypothyroidism, hyperthyroidism, autoimmune thyroiditis, and thyroid cancer can lead to serious health complications, including cardiovascular disease, infertility, and cognitive decline, if not identified and treated in a timely manner. As such, early and accurate diagnosis is a necessity for effective intervention and long-term disease management. Conventional practice methods typically involve hormonal assays, measuring levels of TSH, T3, and T4, alongside patient history and imaging techniques. While these methods form the cornerstone of thyroid evaluation, they are not without significant limitations. Diagnostic accuracy can be compromised by inter-observer variability, vague or overlapping clinical symptoms, and reduced sensitivity in detecting subclinical or borderline abnormalities. These constraints underscore the need for more robust, data-driven diagnostic tools capable of improving precision and facilitating earlier detection [1].

In this context, machine learning (ML) has emerged as a transformative tool for improving diagnostic precision and supporting evidence-based clinical decision-making. A growing body of research has highlighted the efficacy of ML algorithms in predicting and classifying thyroid disorders, showcasing their ability to detect intricate, non-linear patterns within complex and high-dimensional clinical datasets. While traditional models such as K-Nearest Neighbors (KNN) and Naïve Bayes have been explored for their simplicity and computational efficiency, they are often employed as benchmark models, providing a baseline for comparison against more advanced and robust approaches [2].

Advanced machine learning techniques, such as Random Forests, Support Vector Machines (SVMs), and Gradient Boosting, have consistently demonstrated superior classification accuracy, robustness, and reliability in diagnosing thyroid disorders. These algorithms showcase a significant evolution in diagnosis methodologies, offering enhanced capability to detect complex patterns and subtle clinical indicators, thereby improving the precision and confidence of thyroid disease assessments [6].

A growing body of comparative research highlights the remarkable performance of ensemble learning methods, which consistently outperform traditional model approaches. Their strength lies in combining multiple learners to minimize variance and boost generalization, allowing them to adapt robustly and accurately across a wide spectrum of complex, real-world datasets [7]. Integrated data mining frameworks play a transformative role in modern healthcare by converting complex clinical datasets into meaningful insights that significantly enhance disease classification accuracy. By unifying processes such as data preprocessing, feature selection, and matching learning, these frameworks streamline the diagnostic pipeline, improving precision, minimizing human error, and enabling earlier detection of medical conditions. Their capacity to handle heterogeneous data and uncover subtle, hidden patterns makes them essential for building robust, scalable, and intelligent diagnosis systems [10].

Interactive diagnostic systems powered by machine learning (ML) are poised to transform clinical decision-making by delivering intelligent, adaptive, and highly personalized support to healthcare professionals. By combining sophisticated predictive algorithms with intuitive, user-friendly interfaces, these systems enable clinicians to input patient data, explore multiple diagnostic scenarios, and receive context-specific recommendations in real-time. Their ability to continuously learn from new data and outcomes allows them to evolve alongside medical knowledge, improving diagnostic accuracy and adapting to emerging patterns of care.

Crucially, the integration of explainable AI enhances trans-

parency and interpretability, building clinician trust and ensuring ethical, accountable usage. These systems not only reduce the cognitive load on practitioners but also promote consistency, efficiency, and evidence-based decision-making across clinical workflows. As a result, ML-driven interactive diagnostic tools are becoming instrumental in delivering personalized, high-quality patient care, marking a significant leap forward in the development of next-generation, intelligent healthcare solutions [11].

In recent developments, the fusion of advanced feature selection methods with multi-criteria decision-making frameworks has yielded significant strides in model clarity and effectiveness. This hybrid approach not only sharpens the focus on the most informative attributes but also harmonizes diverse performance metrics. As a result, models that are both more interpretable and strategically optimized. Such innovations mark a pivotal shift toward more transparent, explainable, and high-performing predictive systems across complex domains [12][13].

In the realm of model selection, data preprocessing and feature engineering serve as foundational steps for optimizing predictive performance. Employing robust feature selection techniques, such as Recursive Feature Elimination (RFE), LASSO regression, and information gain, helps reduce data redundancy, eliminate noise, and significantly enhance model interpretability. Equally important is addressing class imbalance, a common issue in medical diagnostic datasets that can severely distort model predictions. Extensive research has highlighted the critical impact of this imbalance, underscoring the need for thoughtful strategies to ensure fair, accurate, and clinically meaningful outcomes [14].

In this study, we developed a machine learning-based diagnostic system for the classification of thyroid disorders. Through rigorous evaluation techniques, including k-fold cross-validation, confusion matrix analysis, and performance metrics, the model's reliability and behavior were thoroughly validated.

The following key contributions highlight the strength and novelty of the proposed approach:

- Advanced feature selection techniques were applied to enhance model transparency and reduce complexity, while extensive hyperparameter optimization was conducted to fine-tune each model for maximum predictive accuracy.
- Constructed a soft-voting ensemble, which significantly enhanced predictive reliability and diagnostic performance.

Ultimately, this work underscores the feasibility and value of deploying a clinically relevant, ML-driven diagnostic solution tailored for accurate and scalable detection of thyroid diseases.

II. Related Works

A substantial body of research has explored the application of machine learning (ML) techniques for the classification and prediction of thyroid disorders, underscoring their transformative potential in modern healthcare. Leveraging diverse clinical datasets, these studies consistently demonstrate the capacity of computational models to improve diagnostic accuracy, streamline clinical workflows, and support evidence-based decision-making.

Abbad et al. [1], Chandel et al. [2], and Chalekar et al. [3] applied the K-Nearest Neighbours(K-NN) and KNN-Naive Bayes models, achieving an accuracy of 97.84%, 93.44%, and 97.00%, respectively. These studies reinforce the simplicity and real-world efficacy of distance-based classifiers. However, their performance is often hindered by sensitivity to class imbalance, local data variations, and high computational complexity during prediction. Chandel et al. [2] particularly noted that the performance dropped sharply to 22.56% under certain conditions, likely due to data imbalance, overfitting, or inadequate hyperparameter tuning.

Turanoglu-Bekar et al. [4], Sen et al. [8], Chaubey et al. [9], and Chaganti et al. [12] explored various tree-based ensemble classifiers (NBTREE, LADTREE, REPTREE, BFTREE, Gradient Boosting, and Decision Trees) on clinical datasets, reporting accuracies ranging from 62.50% to 95.73%. While boosting methods, especially Gradient Boosting, performed well in handling complex data, tree-based approaches struggled with noisy, incomplete datasets and were prone to overfitting if not carefully tuned. Additionally, some models lacked scalability and required substantial computational resources.

Sharma et al. [5]combined Recursive Feature Elimination (RFE) with Logistic Regression, achieving 92.70% accuracy, demonstrating the value of dimensionality reduction for boosting interpretability and performance. LR's linear assumptions limited its applicability in datasets with non-linear decision boundaries.

Verma et al. [6] applied Random Forest and SVM algorithms to an Iraqi medical laboratory dataset, attaining an accuracy of 94.50%. Saleh and Othman [14] combined SMOTE with SVM, improving classification fairness and achieving an accuracy of 91.00%. This work underscores the importance of addressing the imbalance to ensure equitable model performance. Sha [15] proposed an innovative hybrid model integrating Quantum Computing with SVM, reaching an impressive accuracy of 98.30% on the UCI dataset. While these variations of SVM exhibit high performance, they remain sensitive to kernel choices, and in cases like quantum-enhanced models, suffer from complexity and lack of practical scalability in current clinical environments.

Pal et al. [7] used the KEEL thyroid dataset and implemented an ensemble of Naïve Bayes, SVM, and K-NN classifiers, achieving accuracies ranging from 92.70% to 96.90%, reinforcing the power of hybrid modeling strategies. Ali and Browmi [13] introduced a novel approach by applying a Multi-Criteria Decision-Making (MCDM) framework on a Kaggle dataset, achieving 93.00% accuracy. Although these approaches exhibit strong adaptability and accuracy, they may involve complex integration logic and reduced model interpretability factors that can challenge clinical adoption.

In a broad comparative study, Deepika et al. [10] assessed Support Vector Machine (SVM), Decision Tree (DT), and Ar-

tificial Neural Network (ANN) on the UCI Dataset, reporting peak accuracies of 95.02%, 95.00%, and 98.60%, respectively, clearly demonstrating the superior predictive capability of neural networks when handling structured clinical data. Similarly, Tyagi et al. [11] evaluated ANN, KNN, and DT classifiers on the UCI dataset, with ANN achieving a high accuracy of 98.00%, further validating the robustness of the dataset as a benchmark for thyroid disorder classifications. However, deep models often suffer from a lack of explainability, increased risk of overfitting, and dependency on large, high-quality datasets, making them harder to deploy reliably in clinical settings.

Grouped analysis of classifier types reveals that while models like KNN and ANN achieve high accuracies, they are often limited by computational cost or lack of interpretability. SVM-based methods show higher precision but depend heavily on parameter tuning and balanced data. Ensemble and hybrid models generally perform better in complex settings but may require intricate configurations. These trade-offs highlight the need for well-balanced models that offer both predictive power and clinical transparency.

III. Methodology

This research project is centered on the classification of thyroid diseases through the application of machine learning algorithms, with particular emphasis on Ensemble Learning due to its proven efficacy in prior scholarly investigations. The adopted methodology follows a structured and systemic frameworks, comprising key phases including data acquisition, data preprocessing, feature selection, model development, and performance evaluation[17]. This comprehensive approach ensures methodological rigor and enhances the reliability and validity of the resulting predictive models. Fig. 1 illustrates the Machine Learning workflow employed for Predicting Thyroid Cancer.

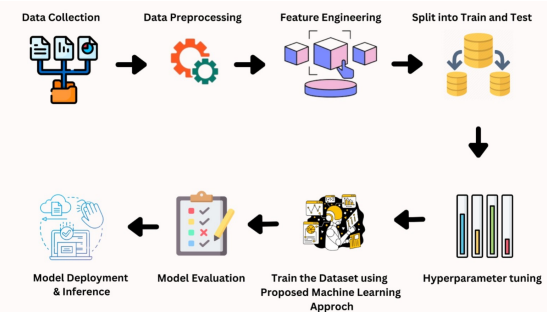

Fig. 1. Machine Learning Process for Predicting Thyroid Cancer

A. Data Collection, Pre-Processing, and Feature Development

The dataset employed in this study is sourced from the UCI Machine Learning Repository[16], specifically the Thyroid Disease dataset, which provides a solid foundation for predictive modeling in the medical domain. It comprises approximately 3,772 patient records encompassing 29 attributes, including both numerical and categorical variables relevant to the assessment of thyroid function. Each record corresponds to an individual patient and includes demographic details such as age and sex. It also contains comprehensive clinical and laboratory data. The dataset exhibits a significant class imbalance (P=3481, N=291), which was intentionally retained to mirror the real-world prevalence of thyroid conditions in clinical settings. Artificial resampling or balancing the dataset was avoided to maintain the biological authenticity of the data and preserve the clinical relevance of model training. Notably, the clinical attributes cover critical information on medication usage (e.g., thyroxine and antithyroid drugs), pregnancy status, and any prior history of thyroid-related treatments or surgeries. These factors provide critical context to the evaluation of thyroid dysfunction. Additionally, the dataset includes key diagnostic indicators that are pivotal for medical decision-making. Key biochemical markers such as Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), Total Thyroxine (TT4), Free Thyroxine Index (FTI), and Thyroxine Uptake (T4U) are present, all of which are widely recognized in clinical endocrinology for assessing thyroid activity.

Data preprocessing was systematically conducted to ensure the dataset's quality and readiness for model training. Upon completion of these preprocessing steps, the dataset retained its original structure, comprising 3,772 instances and 29 attributes, with the target variable suitably encoded for multi-class classification tasks. Following this stage, a comprehensive process of feature development and selection was implemented with the dual objectives of improving model interpretability and enhancing predictive performance. The importance of feature using the Random Forest Classifier is illustrated in Fig. 2.

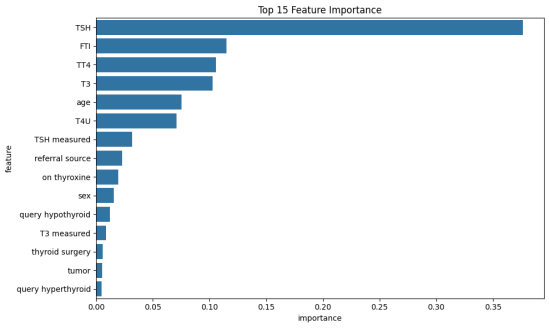

Fig. 2. Feature Importance Using Random Forest Classifier

The bar chart illustrates the top 15 most influential features as identified by a machine learning model, presumably a tree-based classifier, trained to predict thyroid-related conditions. The horizontal axis represents normalized feature importance scores, while the vertical axis lists the corresponding feature names. Among these, "TSH" (Thyroid Stimulating Hormone) emerges as the most significant predictor, with a score exceeding 0.35, substantially higher than all other variables. This finding aligns with established clinical knowledge, as TSH is a critical biomarker in the diagnosis and monitoring of thyroid function.

Other hormone-related features, such as "FTI" (Free Thy-

roxine Index), ”TT4” (Total Thyroxine), and ”T3” (Triiodothyronine) demonstrate moderate importance, with scores ranging approximately from 0.07 to 0.12. These indicators are directly involved in thyroid hormone regulation and, as such, hold considerable relevance in the predictive context. Additional variables, including ”age,” ”T4U” (Thyroxine Uptake), and ”TSH measured,” contribute modestly to the model 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”, exhibit minimal influence, with importance scores approaching zero. While these attributes may possess contextual or epidemiological value, their low predictive impact suggests limited utility in the direct classification process.

The selection of 15 features from the original 29 indicates the implementation of a feature selection strategy aimed at reducing dimensionality. This approach serves to improve model interpretability, decrease computational complexity, and enhance generalization performance.

Overall, the chart underscores the central role of biochemical markers, particularly TSH and other thyroid hormone metrics, as primary determinants in machine learning-based thyroid disease classification. In contrast, demographic and historical attributes appear to exert comparatively less influence on the predictive outcome. A marked class imbalance is present, with a disproportionately higher number of normal cases. This imbalance was rigorously addressed during model training and evaluation to mitigate bias and ensure equitable performance across all classes. Comprehensive data preprocessing was undertaken to prepare the dataset for analysis. These steps were essential to maintain data integrity, minimize bias from heterogeneous feature ranges, and support effective model learning.

In summary, the UCI Thyroid Disease dataset provides a robust and clinically relevant foundation for machine learning applications in thyroid disorder classification. Its rich composition of demographic, clinical, and biochemical features renders it particularly suitable for addressing complex multi-class classification problems within the medical domain, thereby affirming the practical value of data-driven diagnostic methodologies.

B. Hyperparameter Tuning

This section delineates the comprehensive model development pipeline, encompassing algorithm selection, training, and strategic hyperparameter optimization across a diverse array of machine learning classifiers. The models evaluated include Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), and Decision Tree, each selected for their proven efficacy in classification tasks and capacity to model nonlinear and high-dimensional relationships. The Random Forest Classifier employs a bagging-based ensemble approach[18] that aggregates multiple decision trees to yield robust predictions while minimizing variance. Fine-tuning this model involved adjusting parameters such as the total number of estimators, the criteria for internal node division, and the minimum leaf size to balance learning depth and generalization. The K-Nearest Neighbors (KNN) algorithm, known for its simplicity and effectiveness in instance-based learning, was optimized by experimenting with neighbourhood sizes, distance functions, and weighting strategies. The best-performing configuration featured five neighbours, the Manhattan distance as the similarity measure, a power parameter value of 1, and a distance-based weighting scheme that places greater influence on closer neighbors.

Gradient Boosting, a sequential ensemble method that incrementally builds learners to rectify the errors of previous models[19], was meticulously calibrated using key parameters, such as learning rate, number of boosting stages, and tree complexity. The model achieved its most accurate predictions with a learning rate of 0.2, 200 boosting iterations, and trees of limited depth, striking a deliberate trade-off between model complexity and overfitting risk. In the part of Logistic Regression, optimal performance was achieved using L1 regularization with a low regularization strength(0.1) and the ‘liblinear solver’, an effective combination for handling sparse or imbalanced datasets like those found in thyroid classification. The tuning process explores various regularization strengths, penalty functions, and solvers to improve the model’s generalization performance on unseen data. Random Forest is renowned for its robustness in handling high-dimensional and noisy datasets. It builds multiple decision trees during training and outputs the mode of their predictions, thereby reducing variance and avoiding overfitting. Its hyperparameter involves selecting the number of estimators(trees), the maximum depth of each tree, the minimum number of samples required to split a node, and the criteria used for measuring the quality of spilts. These parameters are tuned to balance modelcomplexity, interpretability, and generalization performance.

Ultimately, the Decision Tree classifier was constructed by recursively dividing the dataset based on key feature thresholds, forming a structured hierarchy of decision nodes. To refine its performance, critical parameters such as splitting strategy, tree depth limits, and minimum sample thresholds for splits and leaves were carefully tuned. The fine-tuning procedure encompasses careful calibration of splitting criteria, constraints on tree depths, minimum sample thresholds for both node splits and leaf nodes, alongside a feature selection technique. Such a deliberate and structured approach guarantees the construction of models that are not only robust but also precisely aligned with the unique requirements of the task at hand.

Each model underwent a detailed and systematic hyperparameter tuning process to identify the most effective configurations for optimal performance. The Random Forest model achieved its best results with unrestricted tree depth, a minimum of one sample per leaf(min_samples_leaf: 1), at least two samples needed to splits nodes(min_samples_split: 2), and a total of 300 trees(n_estimators: 300) . For KNN, the ideal parameters included the Manhattan distance metric

(metric: 'manhattan'), five neighbors (n_neighbors: 5), a power parameter of 1 (p: 1), and distance-weighted voting (weights: 'distance'). The Gradient Boosting performed optimally with a learning rate of 0.2, a maximum tree depth of 3, and 200 boosting rounds. Logistic Regression reached peak effectiveness using L1 regularization strength of 0.1(C: 0.1), and the 'liblinear' solver, which is particularly suited for sparse or imbalanced data. The SVM model excelled with an RBF kernel (kernel: 'rbf'), automatic gamma determination, and a penalty coefficient of 10 (C: 10).

Finally, the Decision Tree classifier delivered its best performance using the entropy criterion, no limits on tree depth or feature splits, a minimum of two samples per leaf (min_samples_leaf set to 2), and 15 samples required to split nodes(min_samples_split set to 15). This thorough calibration process ensures that each algorithm is fine-tuned to deliver maximum predictive accuracy and robust generalization.

C. Model Prediction and Evolution

To evaluate the predictive capabilities of various machine learning algorithms for thyroid disease classifications, a comprehensive performance analysis was undertaken. This included the use of five-fold cross-validation, training and testing accuracy evaluations, confusion matrices, and detailed classification reports. The objective was to determine which models exhibit the greatest reliability and generalizability when applied to real-world diagnostic scenarios.

The study explored a variety of supervised learning models, including both classical algorithms and ensemble methods. Each model was trained on a preprocessed dataset and assessed to measure its ability to generalize to unseen data. Cross-validation was employed to mitigate the influence of dataset partitioning and to provide a robust estimate of the model performance.

The evaluation framework encompassed a range of performance indicators, including the mean and standard deviation of cross-validation accuracy, on both training and testing datasets, confusion matrices, and precision-recall metrics.

To address the class imbalance without altering the dataset, this study emphasized class-wise performance metrics, particularly precision, recall, and F1-score, to ensure fair evaluation of the minority class predictions. Traditional accuracy alone can be misleading under imbalanced conditions, as it may overlook a model's failure to detect rare but clinically significant instances. To improve predictive performance in such scenarios, a soft voting ensemble was developed, combining the strengths of three diverse classifiers, Gradient Boosting, Random Forest, and Decision Tree, to enhance predictive accuracy for thyroid disease classification. By integrating these models using a soft voting mechanism, the ensemble achieved a well-balanced trade-off between bias and variance. Fig. 3 illustrates the architectural workflow of the proposed ensemble model, highlighting the parallel use of base classifiers and their combination through soft voting to generate the final prediction.

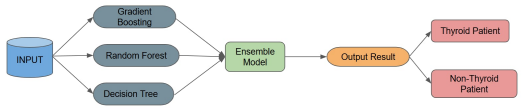

Fig. 3. Architecture of the soft voting ensemble model for thyroid disease classification.

In contrast, K-Nearest Neighbors, though incorporated as a baseline due to its conceptual simplicity and ease of interpretability, demonstrated relatively lower testing accuracy and a higher rate of false positives. The classification report for KNN indicated diminished recall, particularly in identifying minority or borderline cases. Additionally, the notable gap between its high training accuracy and reduced performance on the test set suggested a tendency toward overfitting and an increased sensitivity to local data distributions.

Collectively, these observations affirm the strength of ensemble methods in clinical classification tasks, particularly for complex conditions such as thyroid disorders, where both precision and consistency are essential. The ensemble model classification is in Fig. 4.

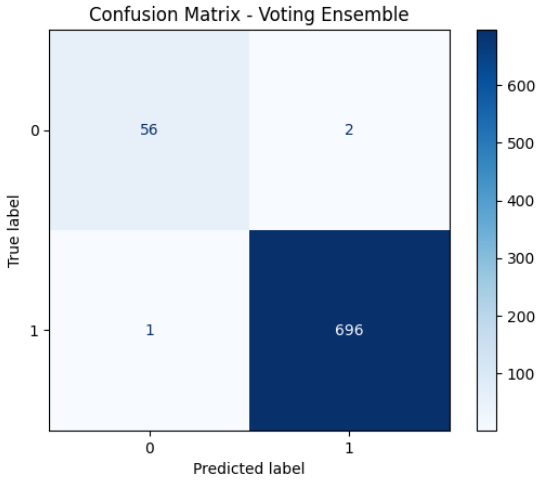

Fig. 4. Confusion Matrix Of Ensemble Method

To further evaluate the discriminative power of the model across varying classification thresholds, the ROC curve was also plotted, as shown in Fig. 5. The ROC curve highlights the relationship between the true positive rate and the false negative rate, providing a comprehensive view of the model's performance.

IV. Experiment and Result Analysis

The proposed method begins with a relatively imbalanced dataset, followed by essential data preprocessing steps. Categorical variables were encoded using LabelEncoder to convert them into numerical form. For feature engineering, RandomForestClassifiers were utilized to select the top 15 most relevant features out of 29, thereby enhancing model

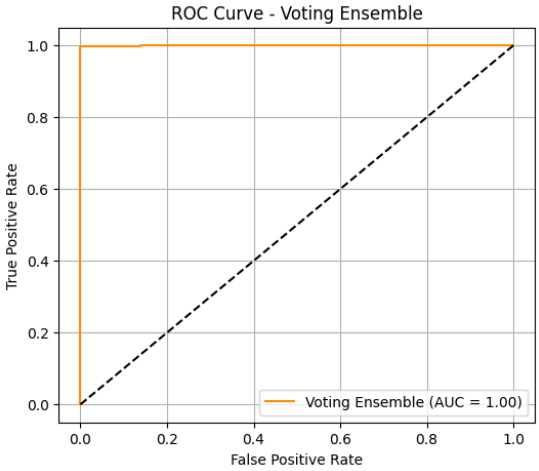

Fig. 5. ROC Curve of Soft Voting Ensemble Model

efficiency and reducing complexity. The dataset has 3,772 patient records, 29 attributes per instance, and was then split into training and testing sets in an 80:20 ratio (80% in training, 20% in testing). To ensure consistent feature scaling, StandardScaler was applied to normalize the input data. Subsequently, GridSearchCV was employed for hyperparameter tuning, optimizing each model's performance. A range of supervised learning models was trained and evaluated as summarized in the corresponding results table. To further boost predictive performance, a voting ensemble model was constructed using three diverse Gradient Boosting, Random Forest, and Decision Tree model.

This combination leveraged the strengths of each model, GB's sequential learning, RFs robust averaging over multiple trees[20], and DT's interpretability. By integrating these models using a soft voting mechanism, the ensemble achieved a well-balanced trade-off between bias and variance. Rather than relying on a single predictive approach, this model synergizes multiple perspectives, including GB's resilience to overfitting, RF's ensemble stability and robustness to noise, and DT's clarity in decision flow, thereby enabling more nuanced and accurate predictions.

What sets this ensemble apart is its capacity to adapt to imbalanced clinical datasets. While traditional classifiers tend to favor majority classes, the ensemble model maintained strong F1-scores across all thyroid categories, including underrepresented ones. It showed marked improvement in detecting subtle deviations from normal thyroid function, which are often difficult to capture in early-stage disease. Moreover, the ensemble's minimal false positive rate is especially vital in clinical diagnostics, as it reduces the likelihood of misdiagnosing healthy individuals, thereby preventing unnecessary anxiety, testing, and treatment.

The performance of the proposed ensemble model, constructed through soft voting, was evaluated using cross-validation scores, accuracy metrics, confusion metrics, and classification reports. The result shown in Fig. 6 demonstrates that it outperformed all individual classifiers, achieving the highest accuracy(99.60%), along with class 0 performance, precision (0.98), recall (0.97), and F1-score (0.97). Among the standalone models, GB alone reached similar accuracy but had slightly lower precision(0.97).The DT also performed well(F1-score 0.96), though with reduced recall(0.93). In contrast, RFs(Random Forest) recall dropped to 0.62, and KNN performed the worst, with a recall of 0.26 and an F1-score of 0.39 for class 0. These results highlight the ensemble's superior ability to detect minority class instances with high reliability, making it better suited for clinical diagnostics where such detection is critical.

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

Rather than relying on a single predictive approach, this model synergizes multiple perspectives, including GB's resilience to overfitting, SVM's ability to separate classes with high margin, and Decision Trees's clarity in decision flow, thereby enabling more nuanced and accurate predictions. The overall comparative performance of multiple machine learning models is shown in TABLE I.

TABLE I
COMPARATIVE PERFORMANCE OF MODELS

Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Ensemble(Soft Voting)	99.60%	0.98	0.97	0.97
Gradient Boosting	99.60%	0.97	0.98	0.97
Decision Tree	99.34%	0.98	0.93	0.96
Random Forest	96.56%	0.90	0.62	0.73
SVM	94.83%	0.81	0.43	0.56
Logistic Regression	94.44%	0.72	0.45	0.55
K-NN	93.77%	0.79	0.26	0.39

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

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

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

$$F_1 - score = \frac{2TP}{2TP + FP + FN} \tag{4}$$

The Ensemble model demonstrates clear superiority by consistently achieving high precision and recall across all classes.

It notably records a substantially lower false positive rate and elevated F1-scores for the negative class, highlighting its robustness and reliability. These attributes make it especially well-suited for clinical applications, where accurate identification of negative cases is crucial to minimizing diagnostic errors and enhancing the overall effectiveness of medical decision-making.

V. Discussion and Conclusion

The comprehensive evaluation of the implemented models identified the Soft Voting Ensemble model as the most effective and reliable, achieving a remarkable test accuracy of 99.60%. This outstanding result highlights the model's superior ability to generalize across unseen data, thereby outperforming all standalone classifiers-particularly in detecting the minority class with high precision and recall. The comparative analysis between the existing methods and our proposed approach are in TABLE II.

TABLE II
Comparison of Existing Methods on Thyroid Disease Detection

Ref.	**Dataset Source**	**Method(s)**	**Accuracy (%)**
Abbad et al. [1]	DHQ Teaching Hospital	K-NN	97.84
Deepika et al. [2]	UCI Repository	SVM, DT, ANN	95.62, 95.00, 98.60
Pal et al. [3]	KEEL repository	Naïve Bayes, SVM, K-NN	94.70, 92.70, 96.90
Chandel et al. [4]	KEEL repository	K-NN, 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	K-NN	97.00
Tyagi et al. [7]	UCI Repository	ANN, K-NN, DT	97.50, 98.00, 75.00
Sharma et al. [8]	UCI Repository	RFE + Logistic Regression	92.70
Verma et al. [9]	Iraqi hospital lab	Random Forest, SVM	94.50
Sen et al. [10]	UCI Repository	RF + Gradient Boosting	95.73
Chaganti et al. [11]	UCI Repository	FS + Gradient Boosting	91.30
Chaubey et al. [12]	UC Irvine KDD	K-NN, DT	89.00
Ali and Broumi [13]	Kaggle Dataset	MCDM	93.00
Saleh and Othman [14]	Local Data	SMOTE + SVM	91.00
Sha [15]	UCI respiratory	Quantum + SVM	96.30
Proposed Model	UCI Repository	Ensemble (GB, SVM, DT)	**99.60**

Ensemble methods are inherently designed to combine the predictive strengths of multiple diverse classifiers. By aggregating the outputs of diverse weak learners, these models effectively reduce variance, enhance stability, and mitigate overfitting. The high accuracy achieved in this case reflects not only the robustness of the ensemble approach but also the efficacy of feature selection methods and hyperparameter optimization strategies employed. The model's capacity to handle high-dimensional and heterogeneous data with precision underscores its practical applicability, particularly in clinical and diagnostic settings where predictive reliability is paramount. The ensemble's consistent performance across multiple evaluation metrics further affirms its suitability for complex classification tasks, where both interpretability and accuracy are essential.

In subsequent research, we intend to expand the study by incorporating a larger and more diverse clinical dataset. This enhancement is expected to improve the model's generalizability, increase robustness across varied patient populations, and further validate its applicability in a real-world clinical environment.