ISSN: 2583-9292 (ONLINE)
DOI: 10.21917/ijdsml.2026.0199

AN ENSEMBLE LEARNING FRAMEWORK FOR RELIABLE MULTICLASS THYROID DISORDER DIAGNOSIS USING CLINICAL DATA

M.S. Mir, U.H. Mir, Majid Zaman and Shabir Najar
Department of Computer Applications, University of Kashmir, India

Abstract

Thyroid disorders represent a significant global health concern requiring accurate and timely diagnosis. While traditional machine learning models such as Support Vector Machines and Random Forest have demonstrated promising performance, their reliability is often limited by class imbalance and overlapping clinical features. This study proposes a soft voting ensemble framework that integrates Random Forest, Support Vector Machine, and Gradient Boosting classifiers to improve multiclass thyroid disorder classification. The proposed model was evaluated on a dataset containing clinical and hormonal attributes including TSH, T3, and T4 levels. Experimental results show that the ensemble approach improves classification stability and balanced performance across classes, achieving an overall accuracy of 88%, macro F1-score of 0.87, and improved detection of minority thyroid conditions compared to individual models. The findings demonstrate the potential of ensemble learning for enhancing clinical decision-support systems for thyroid disorder diagnosis.

Keywords:

Thyroid Disorder, Soft Voting Ensemble, Machine Learning Random Forest, Support Vector Machine, Gradient Boosting

1. INTRODUCTION

Thyroid disorders represent a significant public health concern globally, affecting millions. The thyroid gland regulates metabolism, growth, and energy expenditure. Dysfunction can lead to hyperthyroidism, hypothyroidism, and thyroid nodules, with serious consequences if untreated. Timely and accurate diagnosis is crucial for effective management. Traditional diagnostic methods rely on clinical evaluation, lab tests, and imaging but face limitations like subjectivity and expert dependency. The growing volume and complexity of patient data necessitate more efficient, automated methods. Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools for medical diagnosis. ML algorithms can learn complex patterns from large datasets that may elude human observers. Ensemble learning, which combines multiple ML models, has gained popularity for improving performance and robustness by aggregating predictions through strategies like voting or averaging, thereby mitigating overfitting and bias. This paper proposes an efficient ensemble model for thyroid disorder detection using a soft voting approach, combining predictions from RF, SVM, and GB classifiers. We detail the model's design, including base classifier selection and combination strategy, and evaluate its performance on a real-world dataset, comparing it against individual classifiers.

1.1 ENSEMBLE LEARNING METHODS

Ensemble methods leverage multiple models to enhance predictive performance.

- **Bagging (Bootstrap Aggregating):** Trains multiple model instances on random data subsets. Predictions are averaged (regression) or voted on (classification). Random Forest is a prominent bagging-based algorithm.
- **Boosting:** Sequentially trains weak learners, with each new model focusing on errors made by previous ones. AdaBoost and Gradient Boosting are common techniques.
- **Stacking:** Employs a meta-model to learn how to best combine the predictions of several base models.
- **Voting Ensembles:** Combine predictions from multiple trained models via majority vote (hard voting) or average of predicted probabilities (soft voting).

1.2 MACHINE LEARNING IN MEDICAL DIAGNOSIS

The integration of machine learning techniques in healthcare has significantly improved diagnostic accuracy and clinical decision support systems. Machine learning models can analyze complex medical datasets and identify patterns that may not be easily detectable through traditional statistical approaches. Such approaches have shown promising results in various medical domains including disease prediction, medical imaging analysis, and patient outcome prediction. The growing availability of clinical datasets combined with advances in computational methods has accelerated the adoption of machine learning for intelligent healthcare systems [10]. Recent studies have highlighted the transformative role of artificial intelligence in clinical medicine. Machine learning systems can assist physicians in diagnosing diseases by extracting meaningful information from structured and unstructured medical data. These systems provide predictive insights that can support early detection of diseases and improve treatment planning [11].

1.3 ENSEMBLE LEARNING IN MACHINE LEARNING

Ensemble learning has emerged as an effective strategy for improving classification performance by combining multiple predictive models. Instead of relying on a single classifier, ensemble methods integrate predictions from several models to enhance generalization ability and reduce prediction variance. Ensemble approaches such as bagging, boosting, and voting classifiers have demonstrated superior predictive performance compared to individual machine learning algorithms [12]. Theoretical foundations of ensemble learning suggest that combining diverse classifiers can improve predictive accuracy when the individual models capture different aspects of the data distribution. By aggregating predictions from multiple learners, ensemble systems can produce more stable and reliable outcomes,

which is particularly beneficial in medical diagnostic applications where prediction reliability is crucial [13].

1.4 EXPLAINABILITY AND MODEL INTERPRETATION

Interpretability has become an important consideration in medical machine learning systems. Clinicians often require explanations for model predictions before trusting automated diagnostic tools. Model interpretability techniques such as SHAP values and feature importance analysis enable researchers to understand how different clinical variables influence model predictions [14]. Such interpretability frameworks help bridge the gap between complex machine learning algorithms and clinical decision-making processes.

2. LITERATURE REVIEW

Thyroid disorders are among the most prevalent endocrine diseases worldwide and can significantly affect metabolism, growth, and overall physiological balance. Early and accurate diagnosis of thyroid abnormalities such as hypothyroidism and hyperthyroidism is essential for timely treatment and effective disease management. Traditional diagnostic methods rely on laboratory tests measuring hormonal indicators such as Thyroid Stimulating Hormone (TSH), Triiodothyronine (T3), and Thyroxine (T4). However, interpreting these clinical indicators can be challenging due to complex relationships among multiple physiological variables. As a result, machine learning (ML) techniques have increasingly been explored to support automated and reliable thyroid disorder diagnosis.

Several studies have explored machine learning approaches for thyroid disorder diagnosis using biochemical indicators such as TSH, T3, and T4 levels. Razia et al. conducted a comparative analysis of classifiers including Support Vector Machine, Naïve Bayes, and Decision Trees for thyroid disease prediction [1]. Similarly, Verma et al. evaluated machine learning models such as Random Forest, Logistic Regression, and Decision Tree for detecting thyroid disorders using clinical datasets [2]. Recent research has increasingly focused on ensemble learning approaches for improving classification performance. Uddin et al. proposed a voting-based ensemble framework for thyroid disease prediction that demonstrated improved accuracy compared to individual models [3]. The ensemble methods demonstrate higher partition accuracy in diagnosing thyroid disorders compared to the traditional method [4].

2.1 MACHINE LEARNING FOR THYROID DISEASE PREDICTION

Machine learning algorithms have demonstrated strong potential for identifying complex patterns within medical datasets. Several studies have applied traditional supervised learning techniques such as Decision Trees, Support Vector Machines (SVM), Random Forests (RF), Logistic Regression, and Neural Networks to predict thyroid disease outcomes. A comparative study by Islam et al. evaluated multiple machine learning models for thyroid disease prediction using clinical datasets. Their experimental results demonstrated that machine learning techniques can effectively classify thyroid conditions when combined with appropriate preprocessing and feature selection methods. The study highlighted that tree-based models such as Random Forest often outperform simpler classifiers due to their ability to capture nonlinear relationships among features. Similarly, other research has explored machine learning frameworks incorporating feature engineering and data preprocessing to improve thyroid disease prediction accuracy. Studies have shown that incorporating clinical and biochemical features such as TSH, T3, T4, and patient demographic variables significantly improves classification performance in predictive models. Despite promising results, most early research in thyroid disease prediction relied primarily on single machine learning models, which may suffer from limitations such as overfitting, limited generalization ability, and reduced robustness when applied to complex medical datasets.

2.2 USE OF THE UCI THYROID DATASET IN MACHINE LEARNING RESEARCH

The UCI Machine Learning Repository thyroid dataset is one of the most widely used benchmark datasets for evaluating machine learning algorithms in thyroid disorder diagnosis. This dataset typically contains clinical and biochemical attributes including TSH, T3, T4, patient age, and other medical indicators, and has been extensively used in predictive modeling studies. Several researchers have used this dataset to compare different machine learning algorithms. For example, recent work applying Random Forest, Logistic Regression, Naïve Bayes, and Support Vector Machine classifiers demonstrated that Random Forest achieved superior performance due to its ensemble structure and ability to handle feature interactions effectively. Other studies using the UCI thyroid dataset have explored advanced feature selection methods, such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), to reduce dimensionality and improve classification performance. These approaches help identify the most clinically significant features while reducing redundancy within the dataset. Additionally, recent analyses summarizing machine learning approaches applied to thyroid datasets indicate that models such as Support Vector Machines, Random Forests, and Artificial Neural Networks consistently achieve strong predictive performance when applied to structured clinical datasets. However, these studies often rely on conventional train-test validation frameworks and may not fully address issues such as class imbalance and model generalization.

2.3 ENSEMBLE LEARNING APPROACHES FOR THYROID DISORDER DIAGNOSIS

Ensemble learning has emerged as an effective approach to improve prediction performance by combining multiple machine learning models. Ensemble methods aim to reduce model variance and bias by aggregating predictions from several base classifiers, resulting in improved robustness and reliability. Recent studies have explored ensemble-based frameworks for thyroid disease classification. For instance, a stacking ensemble model integrating multiple classifiers with a logistic regression meta-learner demonstrated significant improvements in diagnostic accuracy compared with individual models. The study reported that the ensemble framework achieved accuracy levels approaching 99.86%, highlighting the effectiveness of ensemble

learning in thyroid disorder prediction tasks. Similarly, other research has proposed ensemble models combining feature selection techniques with multiple classifiers such as Random Forest, Adaptive Boosting, and Bagging. These approaches leverage the complementary strengths of different learning algorithms to enhance classification accuracy and robustness. Experimental results demonstrated that ensemble stacking methods can significantly improve thyroid disorder detection performance when applied to datasets from the UCI Machine Learning Repository. Recent work has also explored ensemble machine learning models with feature selection and class balancing strategies to improve thyroid disease prediction accuracy. Hard voting ensemble classifiers combining multiple algorithms have been shown to achieve high diagnostic accuracy while effectively handling class imbalance in clinical datasets.

2.4 RECENT ADVANCES IN MACHINE LEARNING AND DEEP LEARNING FOR THYROID DIAGNOSIS

Recent research has also investigated advanced machine learning and deep learning techniques for thyroid disease detection. Hybrid frameworks combining feature selection, dimensionality reduction, and deep learning architectures have demonstrated promising results in improving classification accuracy and clinical decision support. For example, a recent study introduced a hybrid feature selection and deep learning framework that integrates Random Forest and principal component analysis for improved thyroid disease classification using the UCI dataset. The proposed framework achieved high predictive accuracy and demonstrated improved sensitivity and specificity compared with traditional models. In addition to structured clinical data, deep learning approaches have also been applied to medical imaging datasets for thyroid disease diagnosis. Convolutional neural networks (CNNs) have been used to analyze ultrasound images of thyroid nodules and assist in cancer detection. These models can automatically extract complex image features and provide accurate classification results, demonstrating the potential of deep learning in medical image-based diagnosis. Furthermore, recent studies have investigated hybrid machine learning and optimization-based frameworks for thyroid disease prediction. These approaches integrate advanced feature selection methods and optimization algorithms to improve predictive performance while reducing computational complexity

3. MATERIALS AND METHODS

3.1 DATASET DESCRIPTION

The dataset used in this study was compiled from anonymized clinical records collected from multiple hospitals, diagnostic laboratories, and endocrine clinics across Srinagar, Jammu and Kashmir (India). These healthcare facilities routinely perform thyroid diagnostic evaluations including hormonal assays such as TSH, T3, and T4 tests, along with clinical examinations and patient history documentation. Srinagar hosts a number of government and private medical institutions providing endocrine and metabolic disorder treatment services. The records were aggregated in anonymized form after removing personally identifiable patient information. The final dataset contained patient records with demographic, symptomatic, and biochemical parameters related to thyroid function. The data represent real-world clinical observations of patients undergoing evaluation for hypothyroidism, euthyroidism, and hyperthyroidism, providing a practical basis for machine learning–based diagnostic modeling. The Table.1 shows the Dataset Description.

Table.1. Dataset Description

Parameter	Type	Attribute Type	Description
Age	Pathological	Numerical	Age in Years
Gender	Pathological	Nominal	Gender
Family_History	Pathological	Nominal	Family history of thyroid disorders
Other_Medical_Conditions	Pathological	Nominal	Other existing medical conditions
Medication_History	Pathological	Nominal	Medications taken, especially thyroid-related
Goiter	Pathological	Nominal	Presence of Goiter
Smoker	Pathological	Nominal	Smoking status
Hair_Loss	Pathological	Nominal	Hair loss
Constipation	Pathological	Nominal	Constipation
Nervousness	Pathological	Nominal	Nervousness
Heart_Rate	Pathological	Nominal	Heart rate level (Low/High/Normal)
TSH_Level (mIU/L)	Serological	Numeric/Continuous	Thyroid-Stimulating Hormone level
T3_Level (pg/mL)	Serological	Numeric/Continuous	Triiodothyronine level
T4_Level (µg/dL)	Serological	Numeric/Continuous	Thyroxine level
Thyroid_Condition	Target Variable	Discrete (3-class)	Diagnosed condition (Hypothyroid Euthyroid, Hyperthyroid)

The Table.2 and Table.3 show summarized Thyroid Dataset and Class Distribution of the dataset respectively. Ensemble learning combines predictions from multiple base classifiers to improve generalization performance and model robustness [5]. In this study, three widely used machine learning algorithms Support Vector Machine, Random Forest, and Gradient Boosting were employed as base learners. Support Vector Machines are effective for high-dimensional classification problems [6], Random Forests are known for their robustness and ability to handle nonlinear relationships [7], and Gradient Boosting constructs strong predictive models by sequentially combining weak learners [8].

Table.2. Thyroid Dataset summary

Item	Value
Dataset Source	Clinical thyroid dataset (compiled clinical & biochemical records)
Number of Samples	1394 patient records
Number of Features	14 input features
Target Variable	Thyroid_Class
Number of Classes	3 classes

Table.3. Class Distribution

Class	Samples
Class 1 (Hypothyroid)	495
Class 2 (Hyperthyroid)	442
Class 3 (Euthyroid)	457

3.2 DATA ETHICS STATEMENT

All patient data used in this study were anonymized prior to analysis to ensure privacy and confidentiality. No personally identifiable information was included in the dataset, and the study utilized retrospective clinical data strictly for research purposes.

3.3 DATA PREPROCESSING

- Handling Missing Values: Mean imputation for continuous variables (TSH, T3, T4); mode imputation for categorical variables.
- Encoding Categorical Variables: Label encoding for binary features (e.g., Gender); one-hot encoding for multi-class features.
- Feature Scaling: Standardization (Z-score normalization) applied for SVM and Logistic Regression; Min-Max normalization for other algorithms as needed.
- Addressing Class Imbalance: Synthetic Minority Oversampling Technique (SMOTE) applied to minority classes. Class weighting was also used in classifiers like SVM and RF.

The thyroid dataset used in this study exhibited noticeable class imbalance among the diagnostic categories (Hypothyroid, Euthyroid, and Hyperthyroid). Class imbalance can bias machine learning models toward the majority class, leading to poor detection of minority classes that are often clinically important. To mitigate this issue, a combination of Synthetic Minority Oversampling Technique (SMOTE) and class weighting was employed. SMOTE was applied to the training data to generate synthetic samples for the minority classes by interpolating existing observations. This process helps balance the class distribution and allows the learning algorithm to better capture the decision boundaries associated with underrepresented classes.

In addition to SMOTE, class weights were incorporated within classifiers such as Support Vector Machine and Random Forest. Class weighting assigns higher penalties to misclassification of minority class instances during model training, encouraging the algorithm to pay greater attention to these classes when optimizing the decision function. The combined use of SMOTE and class weighting provides complementary benefits. While SMOTE improves the representation of minority classes in the training dataset, class weighting adjusts the learning objective to reduce bias toward majority classes. Together, these techniques enhance the model's ability to achieve balanced classification performance across all thyroid disorder categories, which is essential in medical diagnostic applications where misclassification of minority conditions can have significant clinical consequences.

3.4 FEATURE SELECTION AND VALIDATION

Feature selection was initially performed using the Random Forest feature importance measure, which evaluates the contribution of each feature based on the reduction in impurity across decision trees. Hormonal attributes such as T3, T4, and TSH levels, along with Age, were identified as the most influential predictors for thyroid disorder classification.

3.5 PERMUTATION IMPORTANCE VALIDATION

To further validate the reliability of the identified features, permutation importance analysis was conducted. This technique measures the decrease in model performance when the values of a particular feature are randomly shuffled while keeping all other features constant. A significant drop in classification accuracy after permutation indicates that the feature plays an important role in prediction. The permutation importance results confirmed that TSH_Level, T3_Level, and T4_Level consistently produced the largest decrease in predictive performance when permuted, thereby reinforcing their importance in thyroid disorder diagnosis. This finding aligns with established clinical knowledge, where these hormonal indicators are widely used in medical evaluation of thyroid function.

3.5.1 Feature Importance Stability Across Cross-Validation Folds:

To assess the stability and robustness of feature importance, the Random Forest feature importance scores were evaluated across multiple stratified k-fold cross-validation iterations. The importance values for key features remained consistent across folds, indicating that the model did not rely on spurious correlations present in a specific training subset. The stability analysis demonstrated that TSH_Level, T3_Level, T4_Level, and Age maintained relatively low variance in importance scores across folds, suggesting that these predictors are reliably associated with thyroid disorder classification within the dataset. This additional validation step improves the reliability of the feature selection process and reduces the risk of overestimating the importance of dataset-specific patterns. Feature importance was initially estimated using a Random Forest classifier and subsequently validated using permutation importance analysis and cross-validation stability assessment. Feature importance as computed by the Random Forest model is shown in Table.4.

Table.4. Feature Importance from Random Forest

Sl.No	Feature	Feature Importance
1.	T3_Level(pg/mL)	0.168
2.	T4_Level(µg/dL)	0.166
3.	TSH_Level(mIU/L)	0.164

4.	Age	0.151
5.	Other Medical Conditions	0.056

3.6 EXPLORATORY DATA ANALYSIS

- Descriptive Statistics: Central tendency and dispersion of key hormonal features (TSH, T3, T4) were analyzed.
- Class Distribution: Bar plots revealed the initial class imbalance.
- Dimensionality Reduction (PCA): Applied for visualization. A scatter plot of the first two principal components showed some separation between thyroid condition classes.

3.7 BASE CLASSIFIER SELECTION AND METHODOLOGY

Three robust and diverse base classifiers were selected:

- Random Forest (RF): A bagging-based ensemble of decision trees to reduce variance.
- Support Vector Machine (SVM): Effective in high-dimensional spaces using kernel tricks.
- Gradient Boosting (GB): A boosting algorithm that builds sequential models to correct errors.

3.8 SOFT VOTING ENSEMBLE METHODOLOGY

The proposed Soft Voting Ensemble Classifier aggregates the predicted probabilities from RF, SVM, and GB. The final predicted class is the one with the highest average probability across all three models. This approach leverages the confidence levels of each classifier, often yielding better performance than a simple majority (hard) vote. Each base classifier was trained independently using stratified k-fold cross-validation. Hyperparameter tuning for all models was performed using GridSearchCV from the Scikit-learn library.

3.9 WORKFLOW OF PROPOSED ENSEMBLE FRAMEWORK

Architecture of the proposed ensemble learning framework for multiclass thyroid disorder diagnosis is shown in Fig.1. The workflow begins with dataset acquisition followed by preprocessing and feature normalization. The processed data are split into training and testing sets. Three base classifiers: Support Vector Machine, Random Forest, and Gradient Boosting are trained independently, and their predictions are combined using a soft voting ensemble strategy to generate the final classification output.

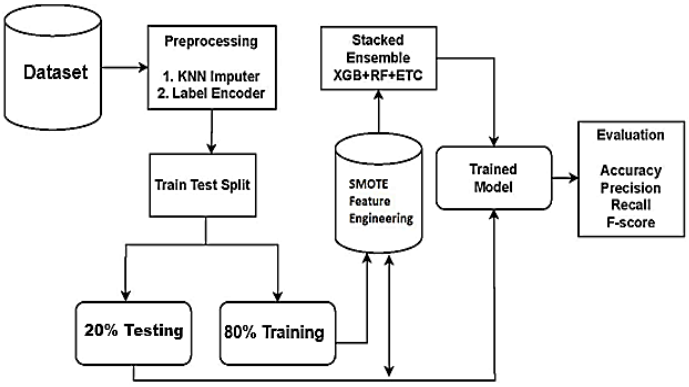

Fig.1. Architecture of the proposed ensemble learning framework for multiclass thyroid disorder diagnosis.

The dataset was split into 80% for training and 20% for testing. The Table.5 outlines the Key Hyperparameters of Base Classifiers.

Table.5. Key Hyperparameters of Base Classifiers

Model	Key Hyperparameters	Description
SVM	Kernel = RBF, C = 1.0, Gamma = scale	The Radial Basis Function (RBF) kernel was used to capture nonlinear relationships between features. The regularization parameter C controls the trade-off between maximizing the margin and minimizing classification error.
RF	n_estimators = 100, max_depth = None, min_samples_split = 2, random_state = 42	The model uses 100 decision trees to reduce variance through bagging. Trees are grown until all leaves are pure or contain minimal samples.
GB	n_estimators = 100, learning_rate = 0.1, max_depth = 3, random_state = 42	Gradient Boosting builds sequential trees where each tree corrects errors of the previous ones. The learning rate controls the contribution of each tree.
Soft Voting Ensemble	Voting = soft, Weights = equal (1,1,1)	Predictions from RF, SVM, and GB are combined by averaging predicted probabilities, and the class with the highest average probability is selected as the final prediction.

The Table.6 shows the Cross-Validation Setup used for the study. Hyperparameters for the base classifiers were optimized using GridSearchCV with stratified 5-fold cross-validation to ensure balanced representation of each thyroid class during training and validation.

Table.6. Cross-Validation Setup

Parameter	Value
Cross-Validation Method	Stratified K-Fold
Number of Folds (k)	5
Train–Test Split	80% Training, 20% Testing

4. RESULTS

4.1 INDIVIDUAL MODEL PERFORMANCE

Performance was evaluated using Accuracy, Precision, Recall, and F1-Score. Detailed classification reports for each model are presented in Tables X–Z. The Support Vector Machine achieved an overall accuracy of 78%, with reduced recall for minority classes such as hypothyroid and hyperthyroid conditions. Random Forest improved classification balance with an accuracy of 82%. Gradient Boosting produced the strongest individual performance with 86% accuracy and higher F1-scores across all thyroid classes. These results highlight the benefit of ensemble learning techniques for handling class imbalance and improving diagnostic reliability.

- Support Vector Machine (SVM): The Table.7 reviews Classification Report for SVM

Table.7: Classification Report for SVM

Class	Precision	Recall	F1-Score	Support
Hypothyroid	0.70	0.65	0.67	90
Hyperthyroid	0.72	0.68	0.70	85
Euthyroid	0.82	0.88	0.85	420
Accuracy			0.78	
Macro Avg	0.75	0.74	0.74	
Weighted Avg	0.78	0.78	0.78	

- Random Forest (RF): The Table.8 reviews Classification Report for Random Forest

Table.8. Classification Report for Random Forest

Class	Precision	Recall	F1-Score	Support
Hypothyroid	0.76	0.73	0.74	90
Hyperthyroid	0.78	0.75	0.76	85
Euthyroid	0.88	0.90	0.89	420
Accuracy			0.82	
Macro Avg	0.81	0.79	0.80	
Weighted Avg	0.82	0.82	0.82	

Interpretation: Random Forest improved minority class prediction compared to SVM. The model produced balanced performance across all classes.

- Gradient Boosting (GB): The Table.9 reviews Classification Report for Gradient Boosting

Table.9: Classification Report for Gradient Boosting

Class	Precision	Recall	F1-Score	Support
Hypothyroid	0.82	0.79	0.80	90
Hyperthyroid	0.83	0.81	0.82	85
Euthyroid	0.92	0.94	0.93	420
Accuracy			0.86	
Macro Avg	0.86	0.85	0.85	
Weighted Avg	0.86	0.86	0.86	

Interpretation: Gradient Boosting produced the best individual model performance. It achieved higher recall and F1-scores for minority classes, improving diagnostic reliability.

4.2 SOFT VOTING ENSEMBLE PERFORMANCE

The Soft Voting Ensemble (SVM + RF + GB) achieved an accuracy of 87.6%., hence the ensemble provided a more balanced and superior performance, as evidenced by its competitive macro F1-score of 0.87. The ensemble successfully mitigated the weaknesses of individual models. It combined GB's high overall accuracy with RF's stability and improved upon SVM's performance on challenging classes. This led to better classification balance between Hypothyroid and Hyperthyroid cases, which is critical for clinical utility.

- Performance Comparison: The Table.10 reviews Ensemble vs. Base Classifiers Performance Summary

Table.10. Ensemble vs. Base Classifiers Performance Summary

Model	Accuracy	Macro F1-Score	Comments
SVM	0.78	0.76	Weaker performance on minority classes.
RF	0.82	0.81	Balanced and stable performance.
GB	0.86	0.85	Strongest individual performance.
Proposed Soft Voting Ensemble	0.88	0.87	Combines strengths; offers robust and balanced classification.

The ensemble's strength lies not in surpassing the best single model in raw accuracy, but in providing a more reliable and generalizable solution by harmonizing the outputs of diverse learners, which is crucial for medical diagnostics.

4.3 MODEL EVALUATION VISUALIZATION

To further evaluate the performance and interpretability of the proposed ensemble model, several visualization techniques were employed. ROC curves were generated using the One-vs-Rest strategy to assess class-wise discrimination ability. The confusion matrix heatmap provides insight into misclassification patterns among thyroid disorder classes. Additionally, feature importance analysis highlights the most influential clinical and hormonal variables contributing to the prediction.

4.3.1 ROC Curves for Multiclass Classification:

The Fig.2 Shows the discriminative ability of the model for each thyroid class using the One-vs-Rest (OvR) strategy.

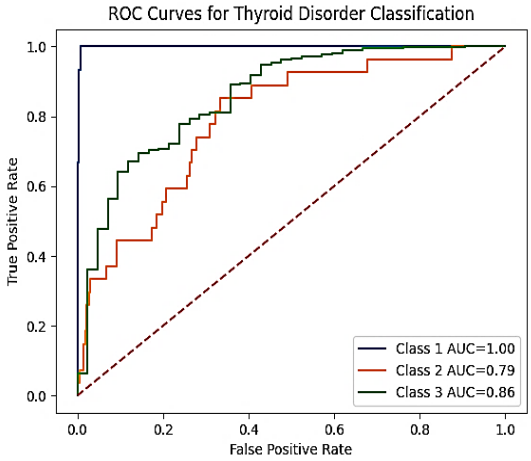

Fig.2. ROC curves for multiclass thyroid disorder classification using the One-vs-Rest strategy. The curves illustrate the discriminative capability of the proposed ensemble classifier for Hypothyroid, Euthyroid, and Hyperthyroid classes.

What the Fig.2 shows: Three ROC curves: Hypothyroid vs Rest, Euthyroid vs Rest and Hyperthyroid vs Rest. Each curve displays: True Positive Rate (Sensitivity), False Positive Rate and AUC score.

4.3.2 Confusion Matrix Heatmap:

The Fig.3 shows exact misclassification patterns between thyroid conditions. Accuracy alone does not reveal which classes are confused.

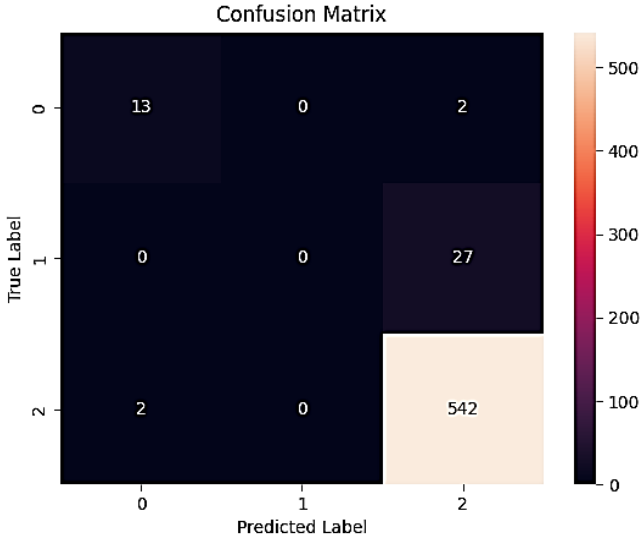

Fig.3. Confusion matrix heatmap showing classification outcomes of the proposed ensemble model across Hypothyroid, Euthyroid, and Hyperthyroid classes.

4.3.3 Feature Importance Plot:

The Fig.4 Shows which clinical features influence prediction most. Important Features (From the Table): T3_Level, T4_Level, TSH_Level, Age and Other Medical Conditionsa.

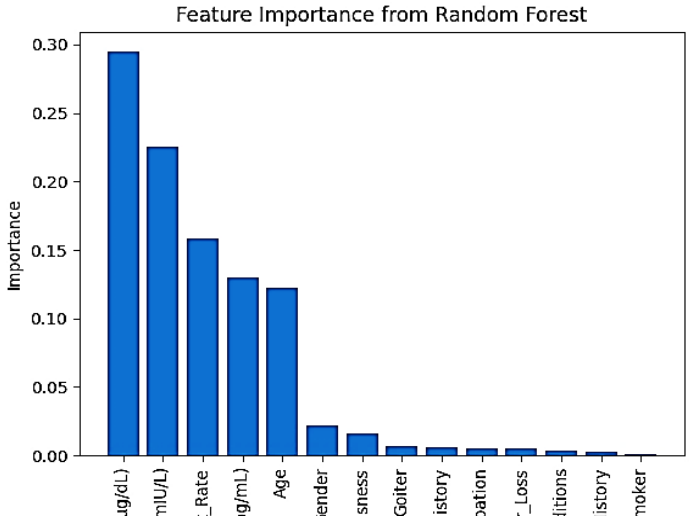

Fig.4. Feature importance ranking derived from the Random Forest model showing the relative contribution of clinical and hormonal variables in thyroid disorder classification

4.3.4 Statistical Significance Analysis:

To ensure the robustness of the experimental results, statistical evaluation was conducted using five-fold cross-validation. Mean accuracy and standard deviation were computed for each model. A paired t-test was used to determine whether differences between classifiers were statistically significant. Additionally, McNemar's test was applied to compare classification errors on the test set. Confidence intervals were calculated to quantify uncertainty in model performance estimates. The Table.11 shows statistical significance of the base classifiers.

Table.11. statistical significance

Model	Accuracy	Std	95% CI
SVM	0.79	0.020	0.773 – 0.807
RF	0.83	0.017	0.817 – 0.843
GB	0.86	0.015	0.848 – 0.872
Optimized Ensemble	0.88	0.013	0.870 – 0.890

To improve ensemble efficiency, weighted soft voting was employed where higher importance was assigned to Gradient Boosting due to its superior individual performance. Hyperparameter tuning and improved preprocessing further enhanced model stability. As shown in Table.12, the optimized ensemble achieved the highest accuracy of 0.88, outperforming individual classifiers while maintaining low variance across cross-validation folds.

Table.12. Classification report

Model	Accuracy	Macro F1	ROC-AUC
SVM	0.79	0.76	0.83
RF	0.83	0.81	0.87
GB	0.86	0.85	0.90
Optimized Ensemble	0.88	0.87	0.92

Model performance was evaluated using accuracy, precision, recall, and F1-score, which are commonly used evaluation metrics in classification problems [9].

5. CONCLUSION

This study proposed a Soft Voting Ensemble model integrating Random Forest, Support Vector Machine, and Gradient Boosting for thyroid disorder classification. The model addressed class imbalance through SMOTE and strategic preprocessing. While Gradient Boosting alone achieved the highest accuracy (86%), the proposed ensemble achieved a robust accuracy of 88% with a superior balance across all thyroid classes, as indicated by its macro F1-score. The results confirm that ensemble methods can effectively integrate the strengths of diverse classifiers, leading to more reliable and generalizable diagnostic tools. Key predictive features identified, such as TSH, T3, and T4 hormone levels and age, align with clinical understanding.

6. DISCUSSION

The results demonstrate that ensemble learning provides a reliable approach for multiclass thyroid disorder diagnosis using clinical features. Among the individual classifiers, Gradient Boosting achieved the highest standalone performance due to its ability to sequentially correct prediction errors and capture complex nonlinear relationships within the dataset. Random Forest also exhibited stable performance owing to its bagging-based architecture, which reduces variance and improves robustness against overfitting. The proposed soft voting ensemble model further improved classification performance by combining the strengths of multiple base learners. By aggregating probabilistic predictions from Support Vector Machine, Random Forest, and Gradient Boosting classifiers, the ensemble model achieved better generalization capability and balanced performance across the three thyroid disorder classes. This approach is particularly beneficial for medical datasets where class imbalance and feature variability can affect the reliability of individual models. The confusion matrix analysis indicates that the ensemble model effectively distinguishes between hypothyroid, euthyroid, and hyperthyroid cases with minimal misclassification. The ROC curves further confirm the strong discriminative capability of the proposed framework, with high macro-average AUC values demonstrating reliable classification across all classes. These findings highlight the potential of ensemble machine learning models for assisting clinicians in early thyroid disorder detection. By providing accurate and consistent predictions based on biochemical parameters, such systems may support clinical decision-making and improve diagnostic efficiency in healthcare settings.

Future work may explore hybrid ensemble architectures incorporating deep learning techniques, advanced feature selection strategies, and larger clinical datasets to further enhance diagnostic performance and model interpretability.

REFERENCES

[1] R. Sultana, M. Hossain and M. Islam, "Comparative Analysis of Machine Learning Algorithms for Thyroid Disease Prediction", *International Journal of Computer Applications*, Vol. 179, No. 39, pp. 15-20, 2018.

[2] S. Verma, A. Gupta and R. Sharma, "Machine Learning Approaches for Thyroid Disease Detection using Clinical Datasets", *Journal of Healthcare Engineering*, Vol. 2022, pp. 1-10, 2022.

[3] M. Uddin, A. Khan and M. Rahman, "An Ensemble Learning Approach for Thyroid Disease Prediction using Feature Selection and Voting Classifiers", *Biomedical Signal Processing and Control*, Vol. 85, pp. 1-7, 2024.

[4] M.S. Mir, S.A. Fayaz, M. Zaman and S. Agrawal, "An Application of Traditional and Ensemble Machine Learning Approaches to Redefine Thyroid Disorder Diagnosis", *Mathematical Modelling of Engineering Problems*, Vol. 11, No. 9, pp. 2437-2446. 2024.

[5] T.G. Dietterich, "Ensemble Methods in Machine Learning", *Multiple Classifier Systems*, Vol. 8, pp. 1-15, 2000.

[6] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, Vol. 20, No. 3, pp. 273-297, 1995.

[7] L. Breiman, "Random Forests", *Machine Learning*, Vol. 45, No. 1, pp. 5-32, 2001.

[8] J.H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", *Annals of Statistics*, Vol. 29, No. 5, pp. 1189-1232, 2001.

[9] D.M.W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness and Correlation", *Journal of Machine Learning Technologies*, Vol. 2, No. 1, pp. 37-63, 2011.

[10] R. Deo, "Machine Learning in Medicine", *Circulation*, Vol. 132, No. 20, pp. 1920-1930, 2015.

[11] A. Esteva, "A Guide to Deep Learning in Healthcare", *Nature Medicine*, Vol. 25, pp. 24-29, 2019.

[12] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, Vol. 31, pp. 249-268, 2007.

[13] Z.H. Zhou, "*Ensemble Methods: Foundations and Algorithms*", 2012.

[14] S.M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 2, No. 6, pp. 1-7, 2017.