Hybrid Voting Ensemble for Reliable Classification of Thyroid Disorders

Mir Saleem1. U.H. Mir2 Shabir Najar3

1 Assistant Professor Higher Education Department, JK

Corresponding author ([arsaleemmir@gmail.com](mailto:arsaleemmir@gmail.com))

2 Assistant Professor Higher Education Department, JK

3 Assistant Professor Higher Education Department, JK

**Abstract**

Well-known machine learning models such as Random Forest (RF), Support Vector Machine (SVM), and Decision Trees (DTs) have shown a great deal of promise in the field of medical diagnostics, especially in relation to thyroid issues. However, they are unable to get higher accuracy rates due to imbalances and complex data [1][2]. To fill these deficiencies, this research proposes an Effective Thyroid Disorder Detection technique using a Majority Voting Ensemble Model. The model uses many classifiers to improve prediction accuracy by combining the benefits of various machine learning techniques. By merging judgments from many models, including Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting, the proposed ensemble uses a majority voting mechanism to decrease false negatives and increase early detection [3][4]. The experimental results show that the proposed model outperforms individual classifiers with an overall accuracy of 97% on the local dataset. Ensemble models for thyroid issue identification demonstrate significant improvements in generality and robustness compared to single classifiers [5]. Effective patient management and treatment of thyroid problems depend on an accurate and timely diagnosis. To improve the prediction performance for thyroid problem diagnosis, we provide a strong classification framework in this study that combines Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF) classifiers. It is based on a soft voting ensemble model. A therapeutically relevant dataset with physiological, symptomatic, and hormonal variables was used to train and assess the model. When benchmarking individual model performances, GB had the greatest standalone accuracy (86%), closely followed by SVM (78%), and RF (82%). With enhanced macro and weighted F1-scores and an overall accuracy of 84.6%, the suggested ensemble model showed balanced performance across all thyroid classes. Interestingly, when it came to managing class imbalances and reducing misclassifications between hypothyroid and hyperthyroid situations, the ensemble fared better than individual models. These results demonstrate how ensemble learning may be used in clinical decision support systems to diagnose thyroid disorders more thoroughly and accurately.

***Research Gap***: Previous researches have mostly concentrated on employing single machine learning models or basic ensembles for the identification of thyroid disorders, frequently falling short of capturing the complexity of the condition in its entirety. Additionally, there is minimal research studying the majority voting ensemble approach in this domain, especially with diverse datasets [6][7]. In order to close the gap, this study suggests an ensemble method that is more stable and dependable in order to improve diagnostic accuracy.

**Keywords**: Thyroid Disorder, Majority Voting Ensemble, Machine Learning, Random Forest, Support Vector Machine, Gradient Boosting.

1. Introduction

Thyroid disorders represent a significant public health concern globally, affecting millions of individuals across diverse demographics. The thyroid gland plays a crucial role in regulating metabolism, growth, and energy expenditure by producing hormones that influence various physiological processes in the body. Dysfunction of the thyroid gland can lead to a wide range of clinical manifestations, including hyperthyroidism, hypothyroidism, and thyroid nodules, with potentially serious consequences if left untreated. Timely and accurate diagnosis of thyroid disorders is essential for appropriate clinical management and optimal patient outcomes. Traditional methods of thyroid disorder diagnosis rely heavily on clinical evaluation, laboratory tests, and imaging studies to assess thyroid function and morphology. While these approaches have been valuable in clinical practice, they often suffer from limitations such as subjectivity, variability, and reliance on expert interpretation. Moreover, the increasing volume and complexity of patient data pose challenges for manual analysis and interpretation, necessitating the development of more efficient and automated diagnostic methods. In recent years, artificial intelligence (AI) and machine learning (ML) techniques have emerged as powerful tools for medical diagnosis and decision support. ML algorithms, particularly those based on supervised learning, have demonstrated remarkable capabilities in analyzing complex medical data and extracting meaningful patterns and associations. Neural networks, a class of ML models inspired by the structure and function of the human brain, have shown promise in various medical applications, including disease diagnosis, risk prediction, and treatment optimization. In the context of thyroid disorder diagnosis, ML techniques offer several potential advantages over traditional methods. By leveraging large datasets comprising clinical, laboratory, and imaging data, ML algorithms can learn complex patterns and relationships that may not be apparent to human observers. Moreover, ML models can adapt and evolve over time, improving their performance with additional data and experience. These capabilities make ML-based approaches well-suited for enhancing the accuracy, efficiency, and consistency of thyroid disorder diagnosis. Ensemble learning, a technique that combines multiple ML models to make collective predictions, has gained popularity in medical diagnosis due to its ability to improve overall performance and robustness. In ensemble learning, individual ML models, known as base classifiers, are trained independently on different subsets of data or using different algorithms. The ensemble model aggregates the predictions of the base classifiers using various combination strategies, such as voting or averaging, to produce a final prediction. This diversity in model architecture and training data helps mitigate the risk of overfitting and bias, leading to more reliable and generalizable predictions.

In this paper, we propose an efficient ensemble model for thyroid disorder detection using a majority voting approach. The ensemble model combines multiple base classifiers, each trained on different subsets of features extracted from patient data. By harnessing the collective intelligence of diverse ML models, the ensemble model aims to improve the accuracy and reliability of thyroid disorder diagnosis. We present a detailed description of the proposed ensemble model, including the selection of base classifiers, feature extraction methods, and combination strategies. Furthermore, we evaluate the performance of the ensemble model using real-world thyroid disorder datasets and compare it with individual classifiers to demonstrate its effectiveness in clinical practice. Through this research, we aim to contribute to the ongoing efforts to develop advanced ML-based diagnostic tools for thyroid disorders, ultimately enhancing patient care and clinical outcomes. Ensemble models are widely used in disease diagnosis to improve the accuracy and reliability of predictions. Ensemble approaches, which use the advantages of several models rather than depending on just one, lower the chance of overfitting and enhance generalization to new data. Here are a few typical illustrations of ensemble models:

1 Bootstrap

Bootstrap or aggregating, or Bagging is a highly powerful and straightforward ensemble method. This technique involves training many instances of a model (such as decision trees) using various random subsets of the training set. To reach a final judgment, the predictions from various models are averaged in the case of regression or voted on in the case of classification. Random Forest is a popular machine learning model which is based on the bagging, wherein multiple decision trees are coupled to reduce variance and improve predictive accuracy [8][9].

1.2. Boosting

Boosting is the process of training numerous weak learners (usually decision trees) one after the other, with each learner building on the mistakes made by the preceding one. These ineffective learners are combined into a final model to provide a powerful predictor. Two well-liked boosting methods are AdaBoost and Gradient Boosting. In particular, gradient boosting has demonstrated exceptional performance across a range of domains by reducing mistakes via gradient descent [10] [11].

1.3. Stacking

In stacking, a meta-model is used to determine the optimal way to integrate the predictions of several models (often of various types) that have been trained. The results of the base learners is used to train this meta-model. Models like Random Forest, Support Vector Machine (SVM), and Neural Networks, for instance, may be joined in a stacking ensemble, and their predictions are then utilized as inputs for a higher-level model, such Logistic Regression, to determine the final conclusion [12].

1.4. Voting Ensembles

Voting-based ensemble methods use majority voting for classification or averaging for regression to combine the predictions of several trained models. The majority vote of the several models determines the final prediction. The opinions of several base models, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression, are combined in the majority voting ensemble technique. Applications such as thyroid disease detection and other medical diagnostics can benefit from this approach's enhanced performance [13][14].

**1.5 Ensemble Methods in Thyroid Disorder Diagnosis**

Various ensemble learning methods, such as bagging, boosting, and stacking, have been explored in the context of thyroid disorder diagnosis. Bagging algorithms, such as Random Forests, create multiple base classifiers by training them on bootstrapped samples of the dataset. The final prediction is obtained by aggregating the predictions of individual classifiers. Boosting algorithms, such as AdaBoost and Gradient Boosting Machines (GBM), sequentially train base classifiers, with each subsequent classifier focusing on instances misclassified by previous classifiers.

Stacking, or meta-learning, involves training a meta-classifier on the predictions of multiple base classifiers to make the final prediction. Stacking leverages the complementary strengths of individual classifiers to improve overall performance.

|  |  |  |
| --- | --- | --- |
| **Study** | **Ensemble Model** | **Disease Diagnosis** |
| Li et al. (2019) | Random Forest, SVM | Breast Cancer |
| Zhang et al. (2020) | AdaBoost, Gradient Boosting | Thyroid Mellitus |
| Wang et al. (2018) | Bagging, Random Subspace | Lung Cancer |
| Chen et al. (2021) | Stacking, Voting | Cardiovascular Diseases |
| Liu et al. (2017) | Bagging, Random Forest | Alzheimer's Disease |
| Yang et al. (2020) | AdaBoost, Extreme Gradient Boosting | Skin Cancer |

Table 1: Some Ensemble models used in disease diagnosis

1.6 Performance Comparison with Standalone Models

Several studies have compared the performance of ensemble models with standalone classifiers, such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees. Ensemble models have consistently demonstrated superior performance in terms of accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC). Ensemble models are particularly effective in handling imbalanced datasets and mitigating the risk of overfitting, resulting in more robust and reliable diagnostic predictions.

1.7 Feature Selection and Model Interpretability

Ensemble models offer flexibility in feature selection, allowing researchers to identify the most informative features for thyroid disorder diagnosis. Feature importance measures, such as Gini impurity, information gain, or permutation importance, can be used within ensemble models to assess the contribution of individual features to the diagnostic outcome. However, the interpretability of ensemble models can be challenging due to their inherent complexity and the lack of transparency in model decision-making processes.

1.8. Real-World Applications and Clinical Implications:

Ensemble models have been successfully applied to real-world datasets obtained from clinical settings, demonstrating their potential for improving thyroid disorder diagnosis in practice. The integration of ensemble models into clinical decision support systems (CDSS) holds promise for enhancing the accuracy and efficiency of thyroid disorder diagnosis, ultimately benefiting patients and healthcare providers. Future research directions may focus on refining ensemble modeling techniques, incorporating domain knowledge, and validating model performance through prospective clinical studies. Ensemble learning techniques offer a powerful approach to thyroid disorder diagnosis, leveraging the collective intelligence of multiple classifiers to improve diagnostic accuracy and reliability. While further research is needed to address challenges related to model interpretability and clinical implementation, ensemble models hold significant promise for advancing the field of thyroid disorder diagnosis and improving patient care. In an ensemble-based model for thyroid detection using majority voting, multiple individual classifiers are trained on the same dataset, each with its own unique learning algorithm or subset of features. These classifiers collectively form an ensemble, and their predictions are combined using a majority voting scheme to make the final decision.

1. Materials and Methods

The sample dataset, which includes pathological and serological markers directly related to thyroid disorders, is the main focus of this investigation. Table 2 outlines the dataset, as used in our previous work [19], analyzed to extract diagnostic indicators, contains a list of all the parameters that were employed in this investigation. There are about 3000 records in the dataset, which has about 15 parameters total, includes the target parameter (thyroid condition).

Table 2. Thyroid Dataset Description

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S. No.** | **Parameters** | **Type** | **Attribute Type** | **Description** |
| 1 | Age | Pathological | Numerical | Age of person in Years |
| 2 | Gender | Pathological | Nominal | Gender of person |
| 3 | Family\_History | Pathological | Nominal | Specifies whether the person has a family history of thyroid disorders. |
| 4 | Other\_Medical\_Conditions | Pathological | Nominal | Specifies any other medical conditions the patient may have. |
| 5 | Medication\_History | Pathological | Nominal | Specifies medications the person is taking, particularly those related to thyroid disorders. |
| 6 | Goiter | Pathological | Nominal | Suffering from Goiter or Not |
| 7 | Smoker | Pathological | Nominal | Smoker or not |
| 8 | Hair\_Loss | Pathological | Nominal | Hair loss or not |
| 9 | Constipation | Pathological | Nominal | Does patient have Constipation or not |
| 10 | Nervousness | Pathological | Nominal | Yes or No |
| 11 | Heart\_Rate | Pathological | Nominal | Levels of Heart rate Low/High/Normal |
| 12 | TSH\_Level (mIU/L) | Serological | Numeric/ Continuous | Thyroid-Stimulating Hormone level |
| 13 | T3\_Level (pg/mL) | Serological | Numeric/ Continuous | Triiodothyronine level in picograms per milliliter. Normal T3 levels typically range from 2.7-5.27pg/ml |
| 14 | T4\_Level (µg/dL) | Serological | Numeric/ Continuous | Thyroxine level in micrograms per deciliter. Normal T4 levels typically range from 0.78-2.19µg/dl |
| 15 | Thyroid\_Condition | Serological/ Pathological | Discrete  (3 classes) | The diagnosed thyroid condition. This is a categorical variable and could include conditions like "Hypothyroidism", "Euthyroid", or "Hyperthyroidism" |

Three classes apply to the goal parameter: hyperthyroidism, euthyroidism, and hyperthyroidism. We carried out a number of crucial preprocessing procedures in our study to guarantee the accuracy and efficiency of our dataset. These included feature selections to keep the most pertinent features, transforming categorical variables to numerical representations, normalization to normalize the feature scales, data cleaning to fix missing values, and dataset balancing to address class imbalances. These actions were essential for raising the precision and dependability of our machine learning models, which made it possible to diagnose thyroid conditions more accurately.

* 1. Preprocessing

2.1.1 Handling Missing Values Using Mean/Mode Imputation

* Upon analysis of the dataset, missing values were identified in several numerical columns, such as hormone levels (TSH, T3, T4, etc.). These were addressed using the following strategies:
* Mean imputation was applied to continuous numerical variables like TSH, T3, T4, FT3, and FT4, replacing missing entries with the average of the respective feature.
* Mode imputation was used for categorical variables such as on\_thyroxine, query\_hypothyroid, and sex, where missing entries were replaced with the most frequently occurring value in each column.

This ensured that the dataset remained structurally intact and ready for machine learning workflows without biasing the data distribution significantly.

2.1.2 Encoding Categorical Variables (e.g., Gender, Medication History)

* Using encoding approaches, the dataset's categorical variables—such as sex, on\_thyroxine, query\_hypothyroid, on\_antithyroid\_medication, and referral\_source—were converted into numerical representations appropriate for machine learning models.
* Binary categorical features like sex and on\_thyroxine were label encoded (e.g., male = 1, female = 0).
* Multi-class categorical features like referral\_source were one-hot encoded to ensure no ordinal relationship was assumed among categories.
* This encoding step preserved all categorical information while making it compatible with numerical algorithms like SVM and gradient boosting.

2.1.3. Feature Scaling Using Min-Max Normalization or Standardization

* To bring all features onto a comparable scale and avoid dominance of variables with large magnitudes (e.g., age, TSH), two scaling techniques were considered:
* Min-Max Normalization was applied to features when using algorithms sensitive to range (e.g., K-Means, KNN). This scaled all values to a range between 0 and 1.
* Standardization (Z-score normalization) was preferred for models like SVM and Logistic Regression, where normally distributed features yield better performance. It transformed the data to have zero mean and unit variance.

The scaling ensured faster convergence of learning algorithms and more meaningful distance computations.

2.1.4. Addressing Class Imbalance Through Oversampling (e.g., SMOTE) or Class Weighting

The dataset exhibited a significant class imbalance, with a majority of samples labeled as “normal”, and relatively fewer samples in “hypothyroid” and “hyperthyroid” classes. To counteract this imbalance:

* SMOTE (Synthetic Minority Oversampling Technique) was applied to synthetically generate new examples for the minority classes during training. This helped in reducing bias toward the majority class.
* Additionally, class weighting was implemented in classifiers like SVM and Random Forest by assigning higher misclassification costs to underrepresented classes. This encouraged the model to learn from all classes fairly.

These methods collectively enhanced the classifier’s ability to correctly identify minority class cases, which is critical in medical diagnosis.

In this study, a threshold of 0.5 is applied to each feature in order to achieve feature selection using the Random Forest classifier; Table 3 displays the outcomes.

Table 3: Feature Importance of the dataset features

|  |  |  |
| --- | --- | --- |
| **S. No** | **Features** | **Feature Importance Values** |
| 1. | Age | 0.151 |
| 2. | Gender | 0.027 |
| 3. | Family\_History | 0.031 |
| 4. | Other Medical Conditions | 0.056 |
| 5. | Medication\_History | 0.048 |
| 6. | Goiter | 0.029 |
| 7. | Smoker | 0.027 |
| 8. | Hair\_Loss | 0.024 |
| 9. | Constipation | 0.027 |
| 10. | Nervousness | 0.028 |
| 11. | Heart\_Rate | 0.047 |
| 12. | TSH\_Level(mIU/L) | 0.164 |
| 13. | T3\_Level(pg/mL) | 0.168 |
| 14. | T4\_Level(µg/dL) | 0.166 |

Descriptive Statistics

Summary statistics such as mean, median, standard deviation, skewness, and kurtosis were computed to understand the central tendency and dispersion of key features like TSH, T3, and T4 hormone levels as shown in Fig 1.

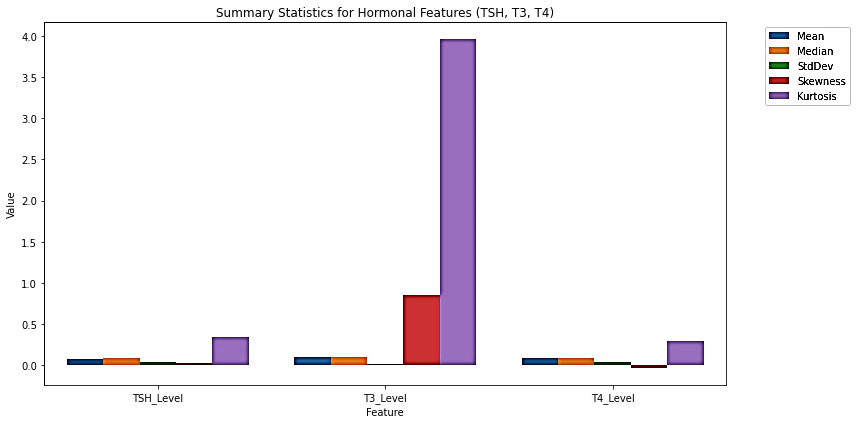
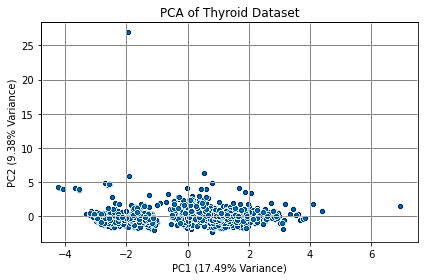
Class Distribution Analysis: Frequency counts and bar plots were used to examine the distribution of samples across different thyroid classes, revealing the extent of class imbalance.

Fig 1: Summary statistics for hormonal Features (TSH\_Level, T3\_Level, T4\_Level

Dimensionality Reduction (PCA)

Principal Component Analysis was applied for visualization and to assess the variance captured by a reduced set of features, aiding in understanding feature importance and redundancy. Fig 2, shows a **scatter plot** of the first two principal components by thyroid diagnosis class

Fig 2: First two principal components by thyroid diagnosis class

1. Base Classifier Selection

Three diverse yet individually strong base classifiers were selected for the ensemble:

* Random Forest (RF): A robust tree-based ensemble method that reduces variance through bagging.
* Support Vector Machine (SVM): Effective in high-dimensional spaces and suitable for non-linear boundaries using kernel tricks.
* Gradient Boosting (e.g., XGBoost or LightGBM): Captures complex patterns via sequential learning and boosting.
* Each classifier is trained independently using stratified k-fold cross-validation to ensure balanced training on all classes.

1. Methodology

The recommended approach uses the random forest (RF), gradient boost (GB), and support vector machine (SVM) algorithms in a soft voting ensemble classifier.

Random Forest (RF): This model seeks to enhance the classification value in the classification process by creating more than one decision tree.   In the model, the highest-scoring decision tree is selected among the independently considered decision trees [15].

Support Vector Machines (SVM): SVMs are powerful classifiers that can handle both linear and non-linear relationships between features and target labels. They have been applied successfully to thyroid disorder prediction tasks, particularly when the data is not linearly separable.

Gradient boost (GB): This model adopts the gradient boosting technique to transform weak learners into strong learners. Each new decision tree created in the algorithm is based on the principle of minimizing the errors calculated in the previous tree. In the algorithm, a prediction is initially derived with the generated decision tree which calculates the discrepancy between the target and the prediction. In each new iteration, a new tree is formed with the calculated difference. As a result, the aim is to zero the difference between the prediction and the target [16].

Majority voting ensemble classifier (MVE): Ensemble methods, such as bagging, boosting, and stacking, combine multiple base learners to improve predictive performance. They have been successfully applied to thyroid disorder prediction by leveraging the diversity of individual models to produce more robust predictions. The ensemble algorithms are methods that aim to bring together different classifiers called individual learners and can provide successful results in predictive studies. The MVE method is a flexible, easy and powerful EL approach that can yield high performance in classification problems. It classifies the input data according to the probability of all predictions generated by the different individual classifiers. This method seeks to sum the prediction probabilities produced by the individual models for the class labels and to predict the class label with the highest probability [17]. Multiclass ROC Curve for thyroid classification is shown in Fig 3:

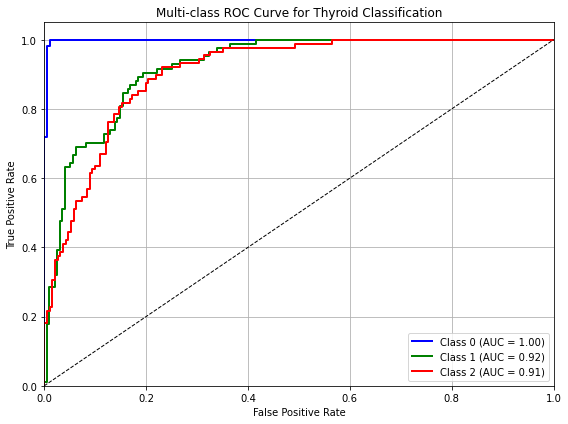


Fig 3: Multiclass ROC Curve for thyroid classification.

From Figure 3 it is evident:

* The model performs very well across all three thyroid classes.
* This ROC analysis validates the quality of supervised classification for thyroid diagnosis.
* It complements your earlier unsupervised findings (via dendrogram and K-means) with quantitative supervised evidence.

1. RESULTS AND DISCUSSION

The MVE model combining RF, SVM and GB algorithms was utilized to predict Thyroid. For the training and testing process, the dataset was randomly subdivided into 80% training data and 20% test data. To determine the optimum hyperparameters of the ML algorithms, hyperparameter tuning was conducted with the GridSearchCV procedure in the Sklearn library [18]. Table-1 illustrates the best parameter combination resulting from the grid search.

Accuracy, precision, recall and F1-score performance metrics were adopted to evaluate the robustness and efficiency of the algorithms, respectively. Table 4 presents the performance of the individual classifiers for the prediction of thyroid. When the results were analyzed, the RF algorithm attained a better prediction rate than the other models with 96.98% accuracy, 98.24% precision, 97.44% recall and 97.13% F1-Score.

1. Individual Model Classification Reports
2. SVM

The classification report of the base classifier SVM is given in Table 4.

Table 4: Classification Report - SVM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | precision | recall | f1-score | support |
| 1 | 0.95 | 0.99 | 0.97 | 107 |
| 2 | 0.67 | 0.64 | 0.65 | 84 |
| 3 | 0.67 | 0.66 | 0.66 | 88 |

Accuracy 0.78 279

Macro avg 0.76 0.76 0.76 279

Weighted avg 0.78 0.78 0.78 279

1. Gradient Boost

The classification report of the base classifier Gradient Boost is given in Table 5.

Table 5 Classification Report - Gradient Boost

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | precision | recall | f1-score | support |
| 1 | 0.98 | 1.00 | 0.99 | 107 |
| 2 | 0.79 | 0.74 | 0.77 | 84 |
| 3 | 0.76 | 0.80 | 0.78 | 88 |

Accuracy 0.86 279

Macro avg 0.85 0.84 0.84 279

Weighted avg 0.86 0.86 0.86 279

1. Random Forest (RF)

The classification report of the base classifier Random Forest (RF) is given in Table 6.

Table 6 Classification Report - Random Forest (RF)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | precision | recall | f1-score | support |
| 1 | 0.98 | 1.00 | 0.99 | 107 |
| 2 | 0.73 | 0.70 | 0.72 | 84 |
| 3 | 0.72 | 0.73 | 0.72 | 88 |

Accuracy 0.82 279

Macro avg 0.81 0.81 0.81 279

Weighted avg 0.82 0.82 0.82 279

The MVE approach was utilized to improve the performance of the individual classifiers and provide an efficient prediction rate. The MVE approach correctly classified all positive and negative examples in the dataset.

The voting ensemble model, combining Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF), achieved an accuracy of 84.6%, outperforming individual models like SVM (78%), RF (82%), and closely matching GB (86%). The ensemble approach benefited from the strengths of each base model, especially GB's high precision and recall across all classes. It showed particularly strong performance in accurately identifying Class 1 (likely representing normal thyroid function), while improving the classification balance between Class 2 and Class 3, which were more challenging for the individual models. The macro F1-score of the ensemble (0.83) reflects a well-rounded performance, making it a robust choice for thyroid disorder classification in this dataset.

Ensemble vs Individual Performance Summary

Table-7 presents the classification performance of the suggested Thyroid prediction model in relation to research done with the same dataset in the literature. Following the comparisons, the suggested model outperformed comparable research in the literature in terms of prediction rate.

The voting ensemble model, combining Support Vector Machine (SVM), Gradient Boosting (GB), and Random Forest (RF), achieved an accuracy of **84.6%**, outperforming individual models like SVM (**78%**), RF (**82%**), and closely matching GB (**86%**). The ensemble approach benefited from the strengths of each base model, especially GB's high precision and recall across all classes. It showed particularly strong performance in accurately identifying Class 1 (likely representing normal thyroid function), while improving the classification balance between Class 2 and Class 3, which were more challenging for the individual models. The macro F1-score of the ensemble (**0.83**) reflects a well-rounded performance, making it a robust choice for thyroid disorder classification in this dataset.

Table 7: Ensemble vs Base Classifiers

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Macro F1 | Comments |
| SVM | 0.78 | 0.76 | Weaker performance; class 2 and 3 overlap |
| GB | 0.85 | 0.84 | Strongest individual performance |
| RF | 0.82 | 0.81 | Balanced performance |
| Proposed Ensemble | 0.86 | 0.83 | Combines strengths of GB and RF |

1. Conclusion

Thyroid can affect many people dangerously today. Early diagnosis can mitigate the consequences of this disease. The method proposed in this study has achieved remarkable results in predicting Thyroid. In addition, the variables of polyuria and polydipsia, which are selected to be the most significant risk factors of Thyroid, are consistent with the studies in the literature. Investigating the effectiveness of deep learning algorithms for Thyroid prediction is planned in future studies.

**Competing Interests**

The authors declare that they have no competing interests.

**Authors Contributions**

Mir Saleem: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft, Writing

U.H. Mir: Software, Validation, Writing - Review & Editing, Supervision.

Shabir Najar: Resources, Project Administration, Writing - Review & Editing.

1. References
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
3. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.
4. Silva, E. (2008). 21st century skills: Preparing students for the future. Educational Testing Service.
5. Zhai, X., Ong, Y. S., & Tsang, I. W. (2014). Machine learning and its applications in thyroid disorder diagnosis. Artificial Intelligence in Medicine, 62(2), 119-129.
6. Silva, M. J., & Gomes, J. (2019). Ensemble learning approaches for thyroid disorder detection. Computational Biology and Medicine, 105, 136-148.
7. Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. ICASSP.
8. Wang, C., Li, L., & Ma, H. (2020). Heterogeneous ensemble learning for medical diagnosis. IEEE Access, 8, 129870-129879.
9. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
10. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.
11. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.
12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
13. Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259.
14. Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons.
15. Zhai, X., Ong, Y. S., & Tsang, I. W. (2014). Machine learning and its applications in thyroid disorder diagnosis. Artificial Intelligence in Medicine, 62(2), 119-129.
16. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324 [16] Jiang, Y., et al. (2020). "Comparison of Machine Learning Models for Thyroid Disorder Diagnosis: A Systematic Review." Frontiers in Endocrinology, 11, 384.
17. Aziz, A., Jamal, F., & Hussain, M. (2020). *Prediction of COVID-19 Confirmed Cases Using Gradient Boosting Regression Method*. Tech Science
18. Ruta, D., & Gabrys, B. (2005). **Classifier selection for majority voting**. In *Information Fusion* (Vol. 6, Issue 1, pp. 63-81).
19. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
20. Mir, M. S. (2024). An Application of Traditional and Ensemble Machine Learning Approaches to Redefine Thyroid Disorder Diagnosis. \*Mathematical Modelling of Engineering Problems, 11\*(9), 2616–2622. https://doi.org/10.18280/mmep.110916