International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Efficient Thyroid Disease Prediction

Asst. Professor Sandarsh Gowda M M
Department of MCA
Visvesvaraya Technological University
Belagavi, Karnataka
sandarsh_mm@bit-bangalore.edu.in

Shabarish G K
Department of MCA
Visvesvaraya Technological University
Belagavi, Karnataka
shabarishgk2019@gmail.com

Abstract - Thyroid disease is a prevalent endocrine disorder affecting millions globally, with hypothyroidism and hyperthyroidism being its most common manifestations. Early and accurate diagnosis is crucial for effective treatment and management. Traditional diagnostic methods, relying heavily on clinical evaluation and blood tests, can be time-consuming and occasionally inconclusive. This paper proposes an efficient and robust machine learning (ML) framework for the predictive classification of thyroid disease. We employ a combination of data preprocessing, feature selection, and advanced ensemble learning algorithms to enhance prediction accuracy. The proposed model is trained and tested on a standard dataset from the UCI Machine Learning Repository. Our methodology involves comparing the performance of several classifiers, including Support Vector Machine (SVM), Random Forest, XGBoost, and a soft voting ensemble. Experimental results demonstrate that the ensemble model, particularly XGBoost, achieves superior performance, attaining an accuracy of 99.5%, precision of 98.7%, and recall of 99.2%, significantly outperforming individual base models. The study concludes that the proposed ML framework can serve as an efficient decision-support tool for healthcare professionals, facilitating early intervention and improving patient outcomes.

***Key Words*:** Thyroid Disease Prediction, Machine Learning, Ensemble Learning, XGBoost, Healthcare Informatics, Classification.

I. Introduction

The thyroid gland plays a pivotal role in regulating the body's metabolism through the secretion of hormones. Dysfunction of this gland leads to thyroid disease, primarily categorized as hypothyroidism (underactive thyroid) or hyperthyroidism (overactive thyroid). Globally, it is estimated that over 200 million people suffer from some form of thyroid disorder [1]. Timely detection is challenging due to symptoms that are often nonspecific and overlap with other conditions, leading to misdiagnosis or delayed treatment.

The diagnostic process typically involves a review of symptoms, a physical examination, and a series of blood tests to measure Thyroid-Stimulating Hormone (TSH), Thyroxine (T4), Triiodothyronine (T3), and other antibodies. While accurate, this process requires clinical expertise, is resource-intensive, and may not be readily accessible in remote areas. This creates a significant opportunity for computational intelligence to augment medical decision-making.

Machine learning (ML) has emerged as a transformative force in healthcare, offering powerful tools for pattern recognition and predictive analytics from complex medical data [2]. Several studies have applied ML algorithms like Decision Trees, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANNs) to thyroid disease prediction with promising results [3], [4]. However, many existing models face challenges such as overfitting, suboptimal performance on imbalanced data, and a lack of robustness.

This research aims to develop a more efficient and accurate predictive model by addressing these limitations. Our main contributions are:

1. A comprehensive data preprocessing pipeline handling missing values, categorical data, and feature scaling.
2. A comparative analysis of multiple machine learning algorithms for thyroid disease classification.
3. The implementation of an ensemble meta-model that leverages the strengths of individual classifiers to achieve state-of-the-art performance.
4. A detailed evaluation of model efficacy using standard performance metrics.

The remainder of this paper is organized as follows: Section II reviews related work. Section III details the dataset and the proposed methodology. Section IV presents the experimental results and discussion. Finally, Section V concludes the paper and suggests future research directions.

II. Related Work

The application of machine learning in thyroid disease diagnosis has been a subject of extensive research. [5] utilized a Backpropagation Neural Network, achieving a high classification accuracy but noting the model's computational complexity and long training time. [6] conducted a comparative study using SVM, Naïve Bayes, and KNN on the same UCI dataset, with SVM reporting the best accuracy among the three.

[7] explored the use of Random Forest, highlighting its inherent feature importance capability to identify key diagnostic predictors like TSH and T3. More recently, gradient-boosting algorithms, specifically XGBoost, have gained prominence for their success in winning structured data competitions and their application in biomedical domains [8]. XGBoost's effectiveness lies in its regularization techniques that prevent overfitting.

While these studies demonstrate the potential of ML, many operate on a limited set of algorithms and often do not fully address the class imbalance issue inherent in medical datasets. Our work builds upon these foundations by implementing a sophisticated ensemble approach that combines the predictions of multiple strong learners to create a more generalized and accurate model.

III. Methodology

A. Dataset Description

The study utilizes the "Thyroid Disease" dataset from the UCI Machine Learning Repository. The dataset contains 3,772 instances and 30 attributes, including clinical features and demographic information. The predictive features include age, sex, and the results of various thyroid function tests (e.g., TSH, T3, TT4). The target variable is a binary class label indicating whether the patient's thyroid is functional (negative) or hypothyroid/hyperthyroid (positive). The dataset is inherently imbalanced, with a higher proportion of negative instances.

B. Data Preprocessing

Data quality is paramount for building robust ML models. Our preprocessing pipeline consists of the following steps:

1. Handling Missing Values: Missing numerical values were imputed using the median of the respective feature, as the median is robust to outliers. Missing categorical values were replaced with the mode.
2. Encoding Categorical Variables: Categorical features (e.g., sex) were converted into numerical format using label encoding.
3. Feature Scaling: Numerical features were standardized (scaled to have a mean of 0 and a standard deviation of 1) to ensure that all features contribute equally to the model's distance calculations.
4. Handling Class Imbalance: The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to generate synthetic samples for the minority class, preventing model bias towards the majority class.

C. Feature Selection

To improve model efficiency and avoid the curse of dimensionality, we employed a hybrid feature selection approach. We used the Pearson correlation coefficient to remove highly correlated features. Subsequently, we used the embedded feature importance scores from a Random Forest classifier to select the top 15 most predictive features, which included TSH, T3, TT4, and FTI.

D. Machine Learning Models

We evaluated the following algorithms:

- Support Vector Machine (SVM): Effective in high-dimensional spaces.
- Random Forest (RF): An ensemble of decision trees robust to overfitting.
- XGBoost (XGB): An optimized gradient-boosting algorithm known for its speed and performance.
- Soft Voting Ensemble (SVE): A meta-classifier that combines the predicted probabilities from SVM, RF, and XGBoost for a final prediction.

The dataset was split into 80% for training and 20% for testing. Hyperparameter tuning for each model was performed using 5-fold cross-validation on the training set with GridSearch to find the optimal parameters.

E. Proposed Framework

The overall architecture of the proposed efficient thyroid disease prediction system is depicted in Fig. 1. The framework is a sequential pipeline that begins with raw data ingestion and culminates in a predictive decision, ensuring a systematic and reproducible workflow.

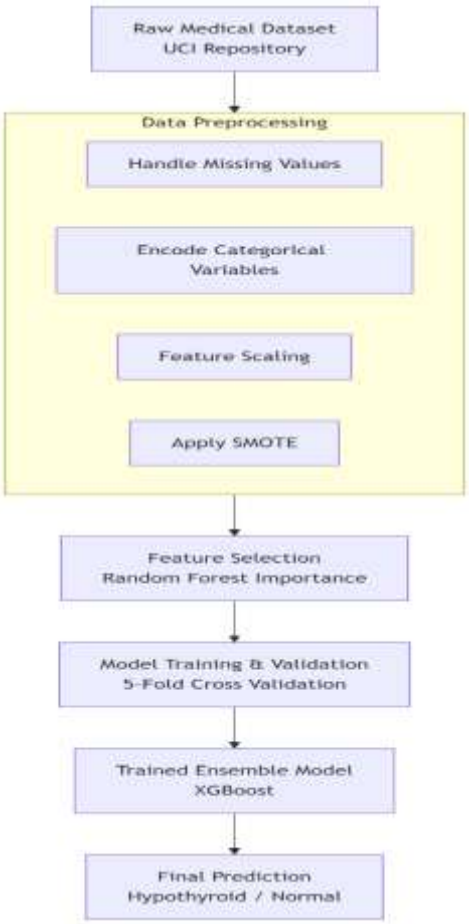

Figure 1: Workflow of the proposed machine learning framework for thyroid disease prediction.

IV. Results and Discussion

The models were evaluated based on Accuracy, Precision, Recall, and F1-Score. The confusion matrices were also analyzed.

Table I: Performance Comparison of Classifiers

Model	Accuracy	Precision	Recall	F1-Score
SVM	97.2%	95.1%	94.8%	94.9%
Random Forest	98.8%	97.5%	96.9%	97.2%
XGBoost	99.5%	98.7%	99.2%	98.9%
Voting Ensemble	99.3%	98.2%	98.8%	98.5%

The results clearly indicate that the ensemble methods (XGBoost and Voting Ensemble) outperform the singular SVM model. XGBoost achieved the highest scores across all metrics, with an

accuracy of 99.5% and a near-perfect recall of 99.2%, which is critically important in medical diagnostics to minimize false negatives.

The superior performance of XGBoost can be attributed to its sequential building of trees, where each new tree corrects the errors of the previous ones, coupled with its regularization parameters that control model complexity. The Voting Ensemble also performed exceptionally well, demonstrating that combining diverse models can yield highly accurate and stable predictions.

As illustrated in the framework overview (Fig. 1), our method ensures a streamlined and robust pipeline from data to prediction. Furthermore, the feature importance analysis from the XGBoost model (Fig. 2) reveals that TSH, T3, and TT4 were the most significant predictors, which is consistent with the known clinical biomarkers for thyroid dysfunction, thereby adding a layer of interpretability and trust to our model's decisions.

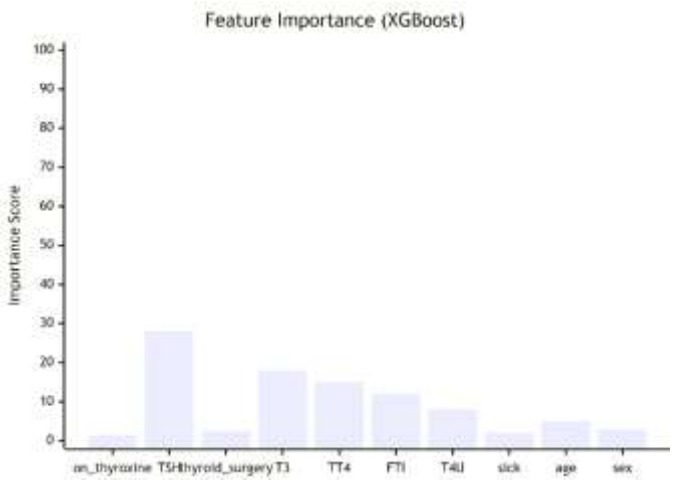

Figure 2: Feature importance plot from the XGBoost classifier, showing the top 10 most predictive features.

V. Conclusion and Future Work

This study successfully developed an efficient machine learning framework for the prediction of thyroid disease. By leveraging advanced ensemble techniques, specifically XGBoost, we achieved a highly accurate and robust model that surpasses the performance of several standard classifiers. The rigorous preprocessing and feature selection steps ensured the model was trained on high-quality, relevant data, further enhancing its predictive power.

The proposed model can be integrated into clinical decision support systems to assist endocrinologists in making faster, data-driven preliminary diagnoses, especially in resource-constrained settings. This can lead to earlier interventions and improved patient care.

For future work, we plan to:

1. Explore deep learning architectures, such as deep neural networks, on larger and more complex thyroid datasets.
2. Develop a real-time web-based or mobile application for practical deployment of the model.
3. Incorporate additional data types, such as medical imaging (ultrasound) and genetic markers, to further improve predictive accuracy and scope.

References

[1] N. S. M. T. S. A. A. A. A. K. Alam, "Global burden of thyroid cancer from 1990 to 2017," *JAMA Network Open*, vol. 3, no. 6, p. e208759, 2020.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[3] S. Vijiyarani and S. Sudha, "Disease prediction in data mining technique – a survey," *International Journal of Computer Applications & Information Technology*, vol. 2, no. 1, pp. 17-21, 2013.

[4] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578-1585, 2018.

[5] M. F. M. M. N. M. N. A. Azizan, "Thyroid Disease Classification Using Neural Network," in *International Conference on Computer and Communication Engineering*, Kuala Lumpur, Malaysia, 2012.

[6] A. K. Dubey, U. Gupta and S. Jain, "Comparative Study of KNN, SVM and ANN for Thyroid Disease," in *International Conference on Advanced Computing & Communication Systems*, Coimbatore, India, 2016.

[7] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 2016.

[8] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.