Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 12, No. 4, 2024, 511-520.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Comparative Analysis of Tree-Based Machine Learning Algorithms on Thyroid Disease Prediction Using ROS Technique and Hyperparameter Optimization

Elahe Moradi*

Department of Electrical and Computer Engineering, Yadegar-e-Imam Khomeini (RAH) Shahre Rey Branch, Islamic Azad University, Tehran, Iran.

Article Info

Article History:
Received 27 August 2024
Revised 20 December 2024
Accepted 12 January 2025

DOI:10.22044/jadm.2025.14990.2598

Keywords:
Machine Learning, Thyroid Disease Prediction, Imbalanced Data, Optimization, Random Forest, XGB.

**Corresponding author: Elahe.Moradi@iau.ac.ir (E. Moradi).*

Abstract

Thyroid disease is common worldwide and early diagnosis plays an important role in effective treatment and management. Utilizing machine learning techniques is vital in thyroid disease diagnosis. This research proposes tree-based machine learning algorithms using hyperparameter optimization techniques to predict thyroid disease. The thyroid disease dataset from the UCI Repository is benchmarked to evaluate the performance of the proposed algorithms. After data preprocessing and normalization steps, data balancing has been applied to the data using the random oversampling (ROS) technique. Also, two methods of grid search (GS) and random search (RS) have been employed to optimize hyperparameters. Finally, employing Python software, various criteria were used to evaluate the performance of proposed algorithms such as decision tree, random forest, AdaBoost, and extreme gradient boosting. The results of the simulations indicate that the Extreme Gradient Boosting (XGB) algorithm with the grid search method outperforms all the other algorithms, obtaining an impressive accuracy, AUC, sensitivity, precision, and MCC of 99.39%, 99.97%, 98.85%, 99.40%, 98.79%, respectively. These results demonstrated the potential of the proposed method for accurately predicting thyroid disease.

1. Introduction

Thyroid disorders are one of the most common chronic diseases that affect a large population in the world. Various research studies show that a large number of people in different communities suffer from this disease; Therefore, early diagnosis of this disease has a vital role and importance [1]. Predicting and diagnosing diseases is a very complicated process for experts in the medical field. In recent years, utilizing artificial intelligence (AI) methods and machine learning (ML) algorithms to adopt the correct diagnosis and prediction of diseases in medical sciences has been the focus of researchers [2, 3].

Data-driven methods have emerged as useful instruments for analyzing dataset from a variety of fields, such as medicine and healthcare [4, 5]. Specifically, ML techniques can efficiently process amounts of medical data to enable accurate disease prediction and diagnosis [6, 7]. The authors in [8] used a neural network classifier with selecting features in association Synthetic Minority Oversampling Technique (SMOTE) for predicting of thyroid disease. The proposed classifier achieved an accuracy of 98.10%. The authors in [9] compared the variant classifiers for thyroid disease prediction including support vector machine (SVM), random forest (RF), XG-Boost, and ensemble method and the results indicated an enhanced accuracy of 99.10%. The research in [10] described five classifiers obtaining SVM, RF, AdaBoost, long short-term memory (LSTM) network, and CNN-LSTM for thyroid disease

prediction. The proposed algorithms obtained an accuracy of 99%.

The authors of [11] investigated a variety of machine learning algorithms in the experiment, such as artificial neural network (ANN), tree-based algorithm (TTA), and statistical models for predicting of thyroid disease. The artificial neural network had the highest score of all their proposed methods with the accuracy of 95.87%. The authors in [12] studied the multiple multilayer perceptron (MMLP) algorithm using back propagation error and adaptive learning approach for thyroid disease diagnosis. The accuracy of 99% had been obtained using their proposed method. The authors of [13] presented the ensemble classifiers using SMOTE method for early thyroid risk prediction. The authors discussed different machine learning algorithm obtaining SVM, decision tree, naïve bayes, and multiple linear regression for thyroid disease prediction. The results displayed the accuracy of 99.23% for predicting of thyroid disease [14].

The authors in [15] proposed two classical machine learning algorithms including SVM and decision tree with Bayesian graph for classification of thyroid disease. Also, they applied SMOTE approach for balancing thyroid dataset, and results showed avg F1-score of 98%. The research in [16] considered a modified extreme gradient boosting (MXGB) algorithm for predicting thyroid disease and the proposed method attained a notable 96.40 % accuracy. In addition, the authors of this paper evaluated the Matthews Correlation Coefficient criterion and the simulations showed that the value of this criterion is 86.10%. The authors of [17] explored different machine learning algorithms for classification thyroid texture and the proposed algorithm reached an accuracy of around 90%. The authors in [18] suggested Multitask Cascade Convolution Neural Network architecture for classification of thyroid nodule and their proposed algorithm achieved an impressive accuracy of 98.20%, which outperformed the average performance of common convolutional neural network by 5%. The research in [19] proposed the CNN method with a decision support system for classification medullary thyroid cancer and their result indicated a classification accuracy of 99.00%. The research in [20] introduced a convolutional neural network (CNN) with a new nodule feature method entitled feature fusion ResNet for diagnosing thyroid ultrasonography nodules. Their result showed an accuracy and F-score of 88.30% and 92.52%, respectively. In previous studies, the SMOTE technique has been widely employed for balancing and categorizing thyroid datasets. While SMOTE is an effective approach for data balancing, particularly for large datasets, it often underperforms when applied to smaller datasets due to its reliance on generating synthetic samples, which may deviate from the original data distribution. Additionally, prior research has heavily relied on trial-and-error methods for hyperparameter tuning, rather than structured optimization techniques like Grid Search and Random Search, which offer more systematic and effective solutions. Moreover, the evaluation metrics used in many studies, such as precision, accuracy, and F1-score, though common, do not adequately address the challenges posed by imbalanced datasets. Metrics like the Matthews Correlation Coefficient (MCC), which provides a more balanced assessment by considering all elements of the confusion matrix, have been underutilized.

These limitations underscore the need for further research that incorporates more robust data balancing techniques, structured hyperparameter optimization, and comprehensive evaluation metrics tailored for imbalanced datasets. The proposed methodology in this study directly addresses these challenges by leveraging Random Oversampling (ROS) as an effective method for balancing small datasets, employing Grid Search and Random Search for systematic hyperparameter tuning, and utilizing MCC alongside other standard metrics to provide a thorough evaluation of the proposed tree-based machine learning models. The main contributions of this research are as follows:

- Enhancing accuracy through a comparative analysis of tree-based algorithms, including Random Forest, Decision Tree, Extreme Gradient Boosting, and AdaBoost, for thyroid disease prediction.
- Utilizing Grid Search and Random Search techniques to optimize hyperparameters and determine the most effective tuning approach for each model.
- Improving the performance of the proposed methods by employing the Random Oversampling (ROS) technique to address data imbalance effectively.
- Using the Matthews Correlation Coefficient (MCC) as a key metric to provide a balanced evaluation of model performance, particularly for imbalanced datasets, making it suitable for medical applications.

This research is particularly beneficial for healthcare professionals and decision-makers in the healthcare field. By leveraging optimized machine-learning models, such as those proposed in this study, medical practitioners can achieve

more accurate predictions for thyroid disease, facilitating early diagnosis and effective treatment planning. Additionally, metrics like MCC and comprehensive evaluation methods enhance the robustness of the proposed approaches for imbalanced datasets commonly encountered in medical applications. These contributions also provide policymakers with reliable tools for decision-making in clinical environments, improving patient care and outcomes.

This paper is organized as follows: Section 2 introduces the materials and the proposed tree-based machine learning algorithms, and procedures used under imbalanced dataset for this work. Section 3 presents the results and compares variant machine learning algorithms using classification criteria to predict thyroid disease. Section 4 discusses and compares two hyperparameter optimization techniques for the proposed algorithms. Lastly, Section 5 concludes this study.

2. Material and Methods

The following subsections address each background material and study methods.

2.1. Dataset Acquisition

The thyroid dataset used in this study was obtained from the UCI Machine Learning Repository. It contains 383 instances and 17 attributes. Throughout the fifteen years that the information for this dataset was being collected, every patient was tracked for a minimum of a decade. Attribute name, attribute description, and attribute role obtaining input or target are shown in Table 1. Figure 1 depicts the percentage of instants for each type of target. As shown in figure 1, 275 of the instants were patients with no recurred (71.80%), whereas 108 of the patients have recurred (28.20%).

Since most clinical datasets are imbalanced, it is necessary to balance them for improved algorithm performance. In the subsection, a technique known as ROS is employed to balance the imbalanced thyroid dataset.

2.2. Pre-processing

2.2.1. Feature Scaling

Feature scaling is a crucial data pre-processing technique that standardizes feature values by bringing them to a comparable scale. This ensures a consistent representation of all attributes, enhancing the performance and reliability of the proposed algorithm. Using raw data without preprocessing can significantly reduce the speed and accuracy of learning systems. To address this, the data was normalized before being applied to the system, ensuring uniformity and improved model performance.

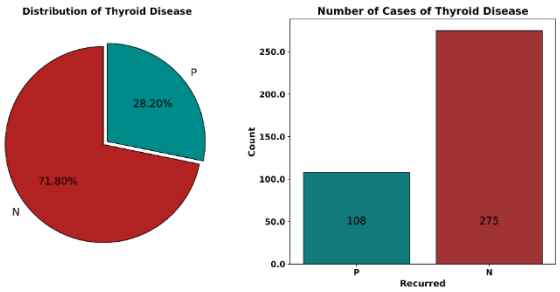

Figure 1. Thyroid dataset distribution.

It is essential for datasets exhibiting varying range, unit, or magnitude characteristics. Common techniques include standardization, normalization, and min-max scaling. This process enhances algorithm performance, convergence, and avoids biasing attributes with larger values. for datasets exhibiting varying range, unit, or magnitude characteristics.

This paper employs standard scaling as a data preprocessing technique. By centering the data at zero mean and scaling to unit standard deviation, this method normalizes the attributes and produces standard attributes. Its mathematical formulation is as follows:

$$x_{scalled} = \frac{x - \mu}{\sigma} \quad (1)$$

Where, $x_{scalled}$ is the value of feature after scaling, and x is the original value. In addition, μ is the mean, and σ is standard deviation of the feature values [21].

2.2.2. Imbalanced Dataset

Due to the inherent imbalance in many clinical datasets, balancing them is crucial for optimizing algorithm performance. Numerous techniques exist to address imbalanced datasets, including methods SMOTE, ROS, and Adaptive Synthetic Sampling (ADASYN). Previous research on thyroid disease prediction and diagnosis has predominantly employed SMOTE technique for dataset balancing [8, 13, 15]. Given the relatively small sample size of the dataset in this study, ROS technique is deemed the most appropriate [22]. Consequently, Table 2 presents the class distribution of the thyroid disease dataset before and after the application of ROS technique. Furthermore, as illustrated in Figure 2, the dataset, following the application of ROS technique, displays a balanced distribution with 275 (50%) non-recurrent and 275 (50%) recurrent cases.

Table 1. Description of attributes from thyroid dataset.

Sl. No.	Attribute Name	Attribute Description	Attribute role
1	Age	The age of the patient at the time of diagnosis or treatment.	Input
2	Gender	The gender of the patient (male or female).	Input
3	Smoking	Whether the patient is a smoker or not.	Input
4	Hx Smoking	Smoking history of the patient.	Input
5	Thyroid Function	The status of thyroid function, possibly indicating if there are any abnormalities.	Input
6	Physical Examination	Findings from a physical examination of the patient, which may include palpation of the thyroid gland and surrounding structures.	Input
7	Adenopathy	Presence or absence of enlarged lymph nodes (adenopathy) in the neck region.	Input
8	Pathology	Specific types of thyroid cancer as determined by pathology examination of biopsy samples.	Input
9	Hx Radiotherapy	History of radiotherapy treatment for any condition.	Input
10	Risk	The risk category of the cancer based on various factors, such as tumor size, extent of spread, and histological type.	Input
11	Focality	Whether the cancer is unifocal (limited to one location) or multifocal (present in multiple locations).	Input
12	T	Tumor classification based on its size and extent of invasion into nearby structures.	Input
13	N	Nodal classification indicating the involvement of lymph nodes.	Input
14	M	Metastasis classification indicating the presence or absence of distant metastases.	Input
15	Stage	The overall stage of the cancer, typically determined by combining T, N, and M classifications.	Input
16	Response	Response to treatment, indicating whether the cancer responded positively, negatively, or remained stable after treatment.	Input
17	Recurred	Indicates whether the cancer has recurred after initial treatment.	Target

Table 2. Real and balanced thyroid dataset.

Samples	Total	Majority	Minority	Ratio of majority to minority
Real dataset	383	275	108	2.55
Balanced dataset	550	275	275	1

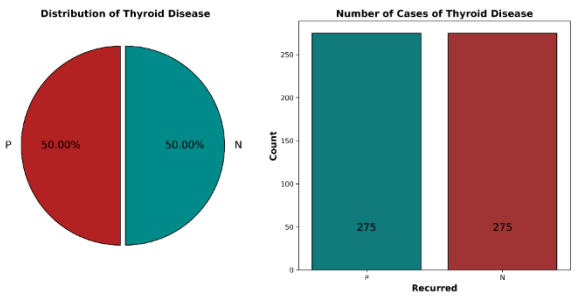

Figure 2. Balanced dataset using ROS technique.

2.2.3. Data Splitting

Employing a 70:30 split, the dataset was divided into training and testing sets, applying 70% data for training, and 30% for testing. The dataset was confirmed using 5-fold cross-validation on the thyroid set, displaying consistent performance. The Python frameworks including Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn were applied for performing the machine learning pre-processing. Furthermore, this paper utilized the ROS technique from the imbalanced-learn Python framework.

2. 3. Proposed Method

The following subsection presents tree-based machine learning algorithms such as decision tree (DT) classifier, random forest (RF) classifier, extreme gradient boosting (XGB) classifier, and AdaBoost (AB) classifier for thyroid disease prediction.

2.3.1. Decision Tree

Decision tree classifiers construct hierarchical algorithms where internal nodes indicate attributes upon which decisions are based, edges denote the possible outcomes of these decisions, and leaf nodes represent the predicted class. The classification of a new data instance is determined by traversing the tree from the root node, following the branches that correspond to the instance's attribute values, until a leaf node is reached [23,

24]. In this paper, DT hyperparameters included splitter, max depth, min samples leaf, max features, class weight, and criterion.

2.3.2. AdaBoost

AdaBoost is an ensemble classifier that sequentially combines multiple 'weak classifier' to construct a 'strong classifier'. Decision stumps, which are one-level decision trees, are a common choice for base learners in AdaBoost due to their simplicity and computational effectiveness [25]. In this search, AB hyperparameters included number of estimators, learning rate, algorithm, base estimator criterion, and base estimator max depth.

2.3.3. Extreme Gradient Boosting

Extreme Gradient Boosting is a highly effective tree-boosting algorithm that can utilize the datasets and achieve state-of-the-art results. By leveraging advanced methods like regularized learning, and approximate algorithms, the XGB excellently moderates overfitting and improves generalization [26]. In this study, XGB hyperparameters included max depth, learning rate, number of estimators, and subsamples.

2.3.4. Random Forest

Random forest algorithm builds numerous DT on training dataset to cause a classifier model. This algorithm decides on a tree-based on furthermost selections, which proposes high accuracy when working with especially huge dataset. This algorithm associations two feature selection approaches, bagging, and random selection, to produce a more effective ensemble algorithm. Employing variant trees with random forest algorithm reduces the problem of overfitting and time of training. [27, 28]. In this paper, RF hyperparameters included number of estimators, max depth, max features, min sample split, min sample leaf, bootstrap, class weight, and criterion. In this study, two criteria such as Gini index, and Entropy index were employed by the following formula:

$$Gini\ Index = 1 - \sum_{n=1} P_n^2 \tag{2}$$

$$Entropy\ Index = -\sum_{n=1} P_n \log_2(P_n) \tag{3}$$

where P_n denotes the probability of class n.

2.4. Hyperparameter Optimization

Hyperparameter tuning is a critical step in enhancing and optimizing the performance of machine learning algorithms. In this paper, two techniques including grid search (GS) and random search (RS) are utilized to tune and utilized to optimize and tune hyperparameters. In GS, a predefined set of values or ranges for each hyperparameter is specified. GS then exhaustively explores the algorithm's performance for all possible combinations of hyperparameters. By providing a comprehensive search of the hyperparameter space, it aims to identify the optimal hyperparameters. Nevertheless, this technique becomes computationally expensive when dealing with a large number of hyperparameters and a wide range of values. RS technique samples a specified number of hyperparameter combinations at random, rather than exhaustively evaluating all possible combinations. This procedure is predominantly effective and efficient when the hyperparameter space is vast and computationally expensive to explore exhaustively [5, 29, 30].

2.5. Performance Evaluations

The performance of the proposed algorithms is assessed based on confusion matrix displayed in Table 3. In this paper, the common criteria such as accuracy, sensitivity, precision, F1-score, and area under curve (AUC) were utilized to measure the results of the machine learning algorithms [31]. Also, considering that the thyroid dataset in this research is a binary classification, the Matthews Correlation Coefficient (MCC) criterion was used to evaluate the proposed algorithms. further, the confusion matrix and receiver operating characteristic (ROC) curve have also been used to assess the performance of the proposed algorithms [32]. In the following, the criteria formulas are presented:

$$Accuracy\ (\%) = \frac{TN + TP}{TN + TP + FP + FN} \times 100 \tag{4}$$

$$Sensitivity\ (\%) = \frac{TP}{TP + FN} \tag{5}$$

$$Precision\ (\%) = \frac{TP}{TP + FP} \times 100 \tag{6}$$

$$F1 - score\ (\%) = \frac{2TP}{2TP + FP + FN} \times 100 \tag{7}$$

$$MCC\ (\%) = \frac{(TP \times TN - FP \times FN\ ) \times 100}{\sqrt{(TP + FP)(FN + TP)(FP + TN)(FN + TN)}} \tag{8}$$

In Table 3, the confusion matrix depicts the binary classification results in terms of TP, TN, FP, and FN.

Table 3. Confusion Matrix

		Actual	
		+	-
Predicted	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

TP denotes the samples that were correctly predicted as the positive class, while TN denotes the sample that were correctly predicted as the negative class. FP denotes samples that were

incorrectly predicted as the positive class and FN denotes samples that were incorrectly predicted as the negative class [5].

3. Result

In this research, the thyroid disease dataset obtained from the UCI Machine Learning Repository was utilized. At first, significant pre-processing steps including missing value, outlier detection, and data scaling by standard scaling were applied to the dataset. Due to the imbalanced nature of the dataset and the limited number of samples in the thyroid dataset, the random oversampling (ROS) technique has been used to balance the dataset and improve the performance of the algorithms. Subsequently, employing the Python-based scikit-learn framework, the dataset was split into training and testing sets with a ratio of 70% to 30%. Finally, tree-based algorithms, including decision tree, random forest, XGB, and AdaBoost, were applied to the dataset. To optimize the hyperparameters of the proposed algorithms in this paper, GS and RS methods have been used and criteria such as accuracy, precision, sensitivity, F1-score, and AUC have been used to evaluate their performance.

Table 4 presents the comparative performance of the proposed algorithms, whose hyperparameters have been tuned and optimized through Grid Search (GS) method. As shown in Table 3, the XGB algorithm outperformed all the other algorithms using GS method, achieving the highest scores in terms of accuracy (Acc), sensitivity (Se), precision (Pr), F1-score (F1), and AUC, with values of 99.39%, 98.85%, 99.40%, 99.39%, and 99.97%, respectively.

Table 4. Comparative of tree-based algorithms using ROS technique and grid search method.

Algorithm	Grid Search method				
	Acc(%)	Se(%)	Pr(%)	F1(%)	AUC(%)
DT	96.36	98.85	96.46	96.36	99.71
AB	98.18	98.85	98.19	98.18	99.96
RF	98.18	98.85	98.19	98.18	99.97
XGB	99.39	98.85	99.40	99.39	99.97

Table 5 displays a comparative analysis of the proposed algorithms, whose hyperparameters were accurately fine-tuned using the Random Search (RS) method. The results, as depicted in Table 4, explicitly demonstrate the superiority of the RF algorithm, which achieved the highest scores across all evaluation metrics: accuracy (96.97%), sensitivity (93.75%), precision (97.14%), F1-score (96.96%), and AUC (99.76%).

Table 5. Comparative of tree-based algorithms using ROS technique and random search method

Algorithm	Random Search method				
	Acc(%)	Se(%)	Pr(%)	F1(%)	AUC(%)
DT	95.76	96.25	95.77	95.76	99.07
AB	96.97	93.75	97.14	96.96	99.66
RF	96.97	93.75	97.14	96.96	99.76
XGB	96.97	93.75	97.14	96.96	99.66

Figures 3-6 present the confusion matrices and receiver operating characteristic (ROC) curves for the proposed algorithms including decision tree (DT), AdaBoost (AB), random forest (RF), and extreme gradient boosting (XGB), respectively. The hyperparameters of these algorithms were tuned using grid search method. The results demonstrate the superiority of the extreme gradient boosting algorithm compared to the other proposed algorithms with grid search method.

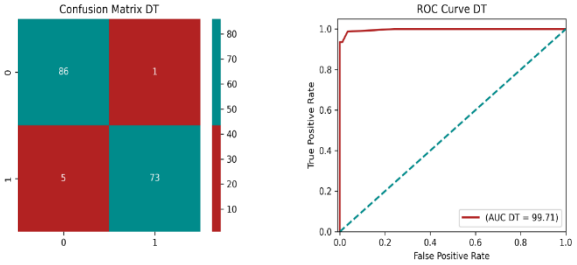

Figure 3. Confusion matrix and ROC curve of decision tree algorithm using grid search method.

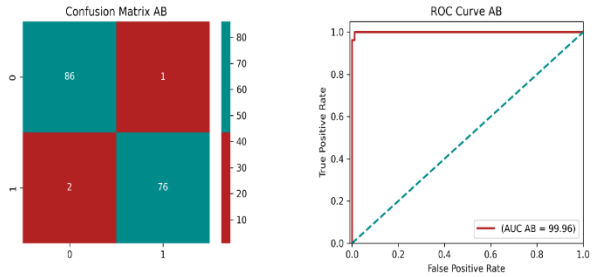

Figure 4. Confusion matrix and ROC curve of AdaBoost algorithm using grid search method.

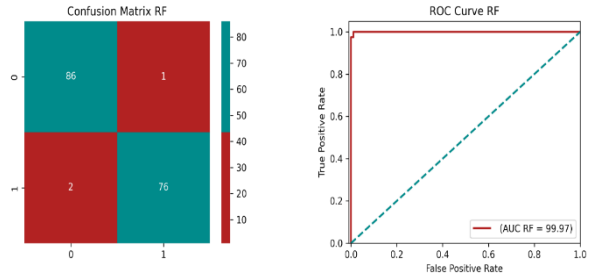

Figure 5. Confusion matrix and ROC curve of random forest algorithm using grid search method.

The comparative performance of the proposed decision tree, AdaBoost, random forest, and extreme gradient boosting algorithms, with hyperparameters optimized via random search, is

visualized in Figures 7-10 through confusion matrices and ROC curves. The random forest and extreme gradient boosting algorithms outperform the other algorithms using random search method, underscoring their superior performance.

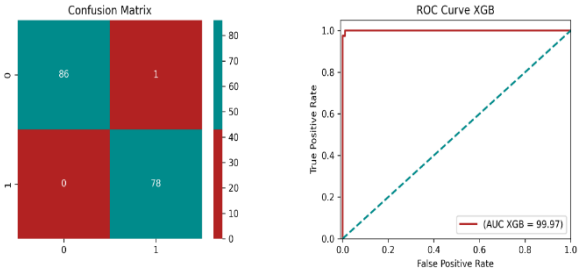

Figure 6. Confusion matrix and ROC curve of extreme gradient boosting algorithm using grid search method.

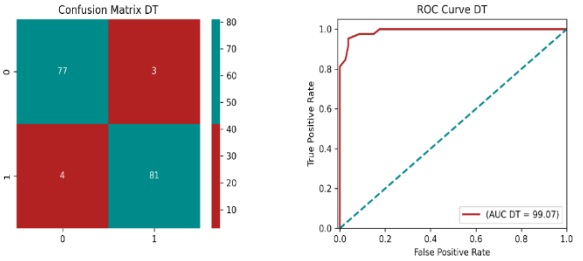

Figure 7. Confusion matrix and ROC curve of decision tree algorithm using random search method.

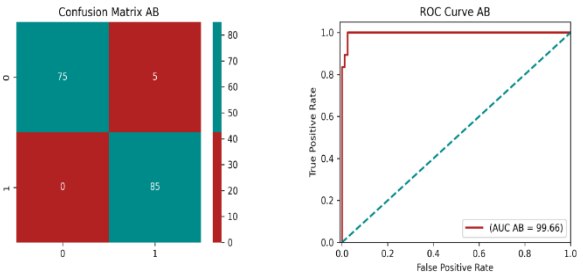

Figure 8. Confusion matrix and ROC curve of AdaBoost algorithm using random search method.

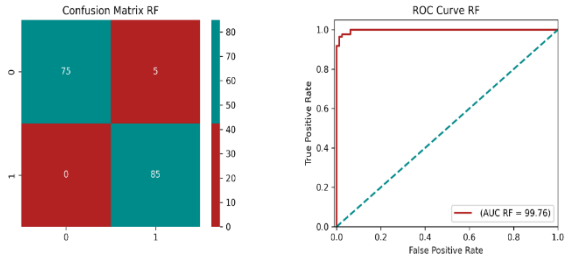

Figure 9. Confusion matrix and ROC curve of random forest algorithm using random search method.

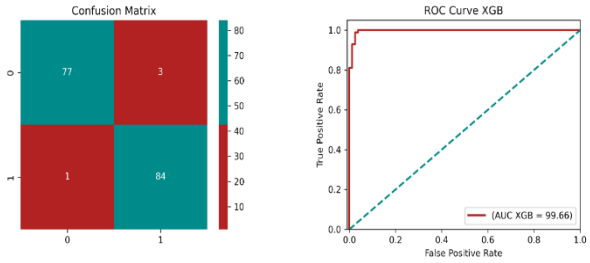

Figure 10. Confusion matrix and ROC curve of extreme gradient boosting using random search method

4. Discussion

In the results section, we surveyed four tree-based machine-learning algorithms with different criteria for predicting thyroid disease. In Figure 11, the MCC criterion for the proposed algorithms has been evaluated and compared with two grid search and random search methods. As shown in Figure 11, the MCC index for all the proposed algorithms, including decision tree, random forest, XGB, and AdaBoost with the grid search method is more efficient in the hyperparameter optimization than the random search method.

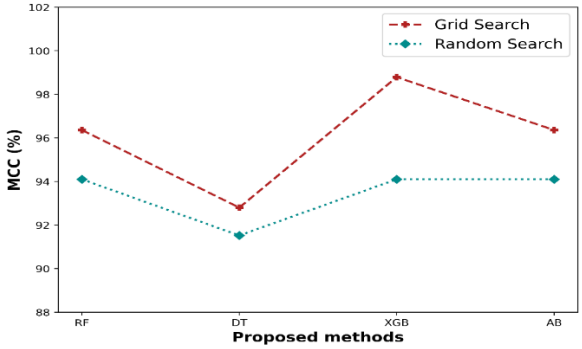

Figure 11. Comparison of MCC criterion using grid search and random search methods.

Figures 12-16 present a comparative analysis of accuracy, sensitivity, precision, F1-score, and AUC for two hyperparameter optimization techniques: grid search and random search. These metrics were employed to evaluate the performance of the proposed algorithms including DT, RF, XGB, and AB in thyroid disease prediction.

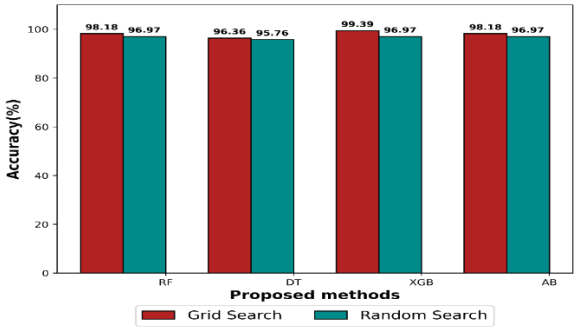

Figure 12. Comparison of accuracy criterion using grid search and random search methods.

Figure 12 illustrates the performance of both methods for accuracy criterion. The findings indicate that grid search is a more effective strategy for hyperparameter tuning.

A comparative evaluation of grid search and random search for hyperparameter optimization is depicted in Figures 13-16. The proposed algorithms for thyroid disease prediction were assessed using sensitivity, precision, F1-score, and AUC in these figures. The results support the superiority of grid search over random search in achieving optimal performance.

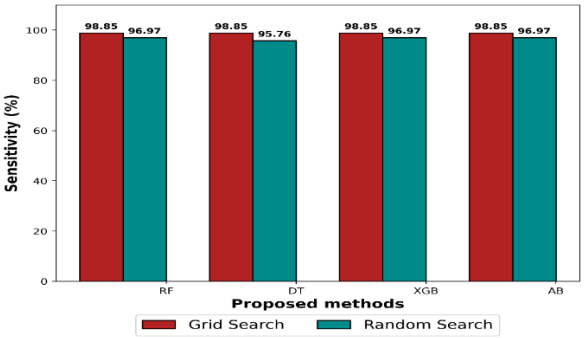

Figure 13. Comparison of sensitivity criterion using grid search and random search methods.

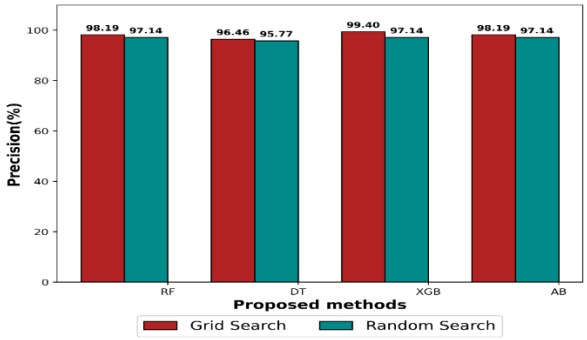

Figure 14. Comparison of precision criterion using grid search and random search methods.

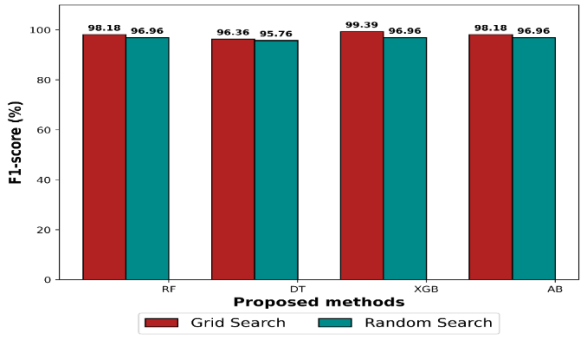

Figure 15. Comparison of F1-score criterion using grid search and random search methods.

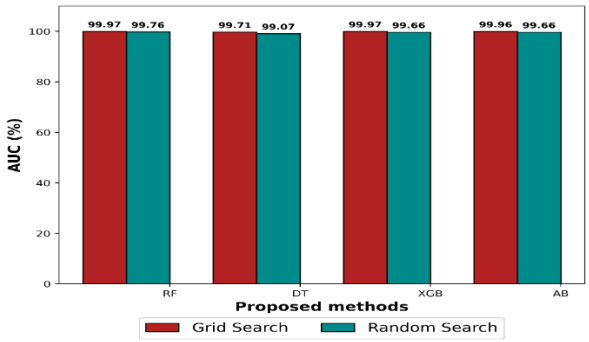

Figure 16. Comparison of AUC criterion using grid search and random search methods.

5. Conclusion

This paper employed tree-based machine learning algorithms, including decision trees, random forests, AdaBoost, and extreme gradient boosting, to predict thyroid disease. Hyperparameter optimization methods, such as grid search and random search, were utilized to enhance algorithm performance. The imbalanced binary classification dataset was sourced from the UCI Machine Learning Repository. To address class balance and enhance proposed algorithms performance, random oversampling (ROS) was implemented.

A comprehensive evaluation of the proposed algorithms was conducted using various criteria, including accuracy, sensitivity, precision, F1-score, MCC, AUC, confusion matrix, and ROC curve analysis. The XGB algorithm, when optimized using grid search, demonstrated superior performance in predicting thyroid disease compared to other proposed algorithms. Furthermore, random forest and XGB algorithms, trained with random search, exhibited competitive results. The findings indicate that in all the proposed algorithms, the grid search method demonstrated significant performance compared to the random search method in predicting thyroid disease.

References

[1] K. Geetha and S. Baboo, "An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayseian Prediction Method," *Global Journal of Computer Science and Technology*, vol. 16, no. 1, pp. 1-10, 2016.

[2] M. Pal, S. Parija, and G. Panda, "Enhanced Prediction of Thyroid Disease Using Machine Learning Method," *IEEE VLSI Device Circuit and System*, pp. 199-204, Feb. 2022.

[3] R. H. Agarwal, S. Degadwala, and D. Vyas, "Predictive Modeling for Thyroid Disease Diagnosis using Machine Learning," International Conference on Inventive Computation Technologies (ICICT), pp.227-231, Apr. 2024.

[4] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare – A Review," *Procedia Computer Science*, vol. 72, pp. 306–313, Jan. 2015.

[5] D. Asif, M. Bibi, M. S. Arif, and A. Mukheimer, "Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization," *Algorithms*, vol. 16, no. 6, p. 308, Jun. 2023.

[6] M. Ramya and P. V. S. Kumar, "PREDICTION AND PROVIDING MEDICATION FOR THYROID DISEASE USING MACHINE LEARNING TECHNIQUE (SVM)," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 11, no. 3, pp. 1099–1107, Jan. 2020.

[7] V. Prasad, T. S. Rao, and M. S. Babu, "Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms," Soft Computing, vol. 20, no. 3, pp. 1179–1189, Jan. 2015.

[8] T. Khan, "Application of Two-Class Neural of Thyroid Disease,", 11th International Conference on Cloud Computing, Data Science & Engineering, Jan. 2021.

[9] P. Kumari et al., "Explainable artificial intelligence and machine learning algorithms for classification of thyroid disease," *Deleted Journal*, vol. 6, no. 7, Jul. 2024.

[10] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," *Cancers*, vol. 14, no. 16, p. 3914, Aug. 2022.

[11] S. S. Islam, Md. S. Haque, M. S. U. Miah, T. B. Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," *PeerJ Computer Science*, vol. 8, p. e898, Mar. 2022.

[12] M. Hosseinzadeh et al., "A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things," *The Journal of Supercomputing*, vol. 77, no. 4, pp. 3616–3637, Aug. 2020.

[13] M. H. Alshayeji, "Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers," Machine Learning and Knowledge Extraction, vol. 5, no. 3, pp. 1195–1213, Sep. 2023.

[14] S. Razia, P. SwathiPrathyusha, N. V. Krishna, and N. S. Sumana, "A Comparative study of machine learning algorithms on thyroid disease prediction," *International Journal of Engineering & Technology*, vol. 7, no. 2.8, p. 315, Mar. 2018.

[15] G. Mollica et al., "Classification of Thyroid Diseases Using Machine Learning and Bayesian Graph Algorithms," *IFAC-PapersOnLine*, vol. 55, no. 40, pp. 67–72, Jan. 2022.

[16] S. Dalal et al., "Enhancing thyroid disease prediction with improved XGBoost model and bias management techniques," *Multimedia Tools and Applications*, Jul. 2024.

[17] P. Poudel, A. Illanes, E. J. G. Ataide, N. Esmaeili, S. Balakrishnan, and M. Friebe, "Thyroid Ultrasound Texture Classification Using Autoregressive Features in Conjunction with Machine Learning Approaches," *IEEE Access*, vol. 7, pp. 79354–79365, Jan. 2019.

[18] W. Song et al., "Multitask Cascade Convolution Neural Networks for Automatic Thyroid Nodule Detection and Recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1215–1224, May 2019.

[19] J. A. Chandio, G. A. Mallah, and N. A. Shaikh, "Decision Support System for Classification Medullary Thyroid Cancer," *IEEE Access*, vol. 8, pp. 145216–145226, Jan. 2020.

[20] X. Zhao et al., "Automatic Thyroid Ultrasound Image Classification Using Feature Fusion Network," *IEEE Access*, vol. 10, pp. 27917–27924, Jan. 2022.

[21] E. Moradi, "A Data-Driven based Robust Multilayer Perceptron Approach for Fault Diagnosis of Power Transformers," 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), Feb. 2024.

[22] F. Kamalov, H.-H. Leung, and A. K. Cherukuri, "Keep it simple: random oversampling for imbalanced data," Advances in Science and Engineering Technology International Conferences (ASET), Feb. 2023.

[23] L. Aversano et al., "Thyroid Disease Treatment prediction with machine learning approaches," *Procedia Computer Science*, vol. 192, pp. 1031–1040, Jan. 2021.

[24] R. Jha, V. Bhattacharjee, and A. Mustafi, "Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society," *Wireless Personal Communications*, vol. 122, no. 2, pp. 1921–1938, Aug. 2021.

[25] T. R. Mahesh et al., "AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, Apr. 2022.

[26] Y. Chen, D. Li, X. Zhang, J. Jin, and Y. Shen, "Computer aided diagnosis of thyroid nodules based on the devised small-datasets multi-view ensemble learning," *Medical Image Analysis*, vol. 67, Jan. 2021.

[27] K. Sumwiza, C. Twizere, G. Rushingabigwi, P. Bakunzibake, and P. Bamurigire, "Enhanced cardiovascular disease prediction model using random forest algorithm," *Informatics in Medicine Unlocked*, vol. 41, Jan. 2023

[28] P. T. Noi and M. Kappas, "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector

Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery," *Sensors*, vol. 18, no. 2, Dec. 2017.

[29] T. Agrawal, Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient. Apress: Berkeley, CA, USA, pp. 31–51, 2021.
Agrawal, T.; Agrawal, T. Hyperparameter optimization using scikit-learn. In Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning models More Efficient; Apress: *Berkeley, CA, USA*, 2021; pp. 31–51.

[30] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," vol. 13, no. 1, pp. 281–305, Mar. 2012.

[31] S. M. Rostami and M. Ahmadzadeh, "Extracting Predictor Variables to Construct Breast Cancer Survivability Model with Class Imbalance Problem," *Journal of AI and Data Mining*, vol. 6, no. 2, pp. 263-276, 2018.

[32] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, no. 1, Feb. 2023.

تحلیل مقایسه‌ای الگوریتم‌های یادگیری ماشین مبتنی بر درخت برای پیش‌بینی بیماری تیروئید با استفاده از تکنیک ROS و بهینه‌سازی هایپرپارامترها

الهه مرادی*

گروه مهندسی برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی واحد یادگار امام خمینی (ره) شهرری، تهران، ایران.

ارسال ۲۰۲۴/۰۸/۲۷؛ بازنگری ۲۰۲۴/۱۲/۲۰؛ پذیرش ۲۰۲۵/۰۱/۱۲

چکیده:

بیماری تیروئید در سراسر جهان شایع است و تشخیص زودهنگام آن نقش مهمی در درمان و مدیریت موثر دارد. استفاده از تکنیک‌های یادگیری ماشین در تشخیص بیماری تیروئید حیاتی است. این پژوهش الگوریتم های یادگیری ماشین مبتنی بر درخت را با استفاده از تکنیک های بهینه‌سازی هایپرپارامترها برای پیش‌بینی بیماری تیروئید پیشنهاد می‌کند. مجموعه داده بیماری تیروئید از مخزن یادگیری ماشین UCI به عنوان مبنا برای ارزیابی عملکرد الگوریتم‌های پیشنهادی استفاده شده است. پس از مراحل پیش‌پردازش و نرمال‌سازی داده‌ها، تعادل‌دهی داده‌ها با استفاده از تکنیک نمونه‌برداری تصادفی افزایشی (ROS) بر روی داده‌ها اعمال شده است. همچنین، دو روش جستجوی شبکه‌ای (GS) و جستجوی تصادفی (RS) برای بهینه‌سازی هایپرپارامترها به کار گرفته شده‌اند. در نهایت، با استفاده از نرم‌افزار پایتون، معیارهای مختلفی برای ارزیابی عملکرد الگوریتم‌های پیشنهادی مانند درخت تصمیم، جنگل تصادفی، آدابوست و تقویت گرادیان شدید (XGB) مورد استفاده قرار گرفت. نتایج شبیه‌سازی‌ها نشان می‌دهد که الگوریتم تقویت گرادیان شدید با روش جستجوی شبکه‌ای نسبت به سایر الگوریتم‌ها عملکرد بهتری داشته و به درستی، AUC، حساسیت، دقت و MCC به ترتیب ۹۹,۳۹٪، ۹۹,۹۷٪، ۹۸,۸۵٪، ۹۹,۴۰٪، ۹۸,۷۹٪ دست یافته است. این نتایج نشان‌دهنده پتانسیل بالای روش پیشنهادی برای پیش‌بینی دقیق بیماری تیروئید است.

کلمات کلیدی: یادگیری ماشین، پیش‌بینی بیماری تیروئید، داده‌های نامتعادل، بهینه‌سازی، جنگل تصادفی، تقویت گرادیان شدید (XGB).