Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 6 (2023)

Enhanced Accuracy in Thyroid Disease Classification: A Comparative Analysis of Random Forest and Decision Tree Methods

[1]Dr. V. Manimekalai, [2]Dr. S.Gomathi alias Rohini

[1]Assistant Professor, Department of Computer Technology, Dr. N.G.P. Arts and Science College, Coimbatore

[2]Associate Professor, Head, Department of Artificial Intelligence and Machine Learning, Kongunadu Arts and Science College, Coimbatore

Abstract

In rural areas where rapid diagnosis of lifestyle conditions is often unavailable, the development of intelligent prediction systems using modern computing techniques is imperative. This study aimed to enhance the accuracy of thyroid disorder identification by creating a more precise classification model. The study compared the performance of two machine learning techniques, namely Random Forest and Decision Tree, to determine the optimal training model for detecting thyroid disorders. Utilizing datasets from the UCI machine learning library, these classifiers were applied to differentiate between patients with hyperthyroidism and hypothyroidism. Performance measurements such as Accuracy and Precision were used to assess the models. The Decision Tree classifier achieved an 84 percent accuracy rate, while the Random Forest classifier demonstrated an 85 percent accuracy rate. Similarly, the precision rates for the Decision Tree and Random Forest models were calculated to be 82 percent and 84 percent, respectively. These findings suggest that both classifiers offer promising results in accurately identifying thyroid disorders, with the Random Forest model exhibiting slightly higher accuracy and precision.

Keywords: *Thyroiditis, Hyperthyroidism, Hypothyroidism, and Thyroid cancer, Machine Learning, Random Forest, Decision Tree*

1. Introduction

Thyroid diseases represent a diverse spectrum of conditions affecting the thyroid gland, a small butterfly-shaped organ located in the neck that plays a crucial role in regulating metabolism and other essential bodily functions. These disorders can range from benign nodules to life-threatening malignancies, impacting millions of individuals worldwide. Understanding thyroid diseases is paramount due to their prevalence, varied presentations, and significant implications for health and well-being.

The thyroid gland produces hormones, primarily thyroxine (T4) and triiodothyronine (T3), which influence metabolism, growth, and energy levels throughout the body. When the thyroid gland malfunctions, either by producing too much hormone (hyperthyroidism) or too little (hypothyroidism), it can lead to a wide array of symptoms affecting multiple organ systems. Moreover, thyroid disorders are not limited to dysfunction in hormone production. Inflammation of the thyroid, known as thyroiditis, can occur due to autoimmune processes, viral infections, or other causes, leading to pain, swelling, and alterations in hormone levels. Additionally, the development of thyroid nodules, which are abnormal growths within the gland, presents challenges in diagnosis and management, as they may be benign or malignant.

Thyroid cancer though less common than benign thyroid conditions, is a significant concern due to its potential for metastasis and adverse outcomes if not detected and treated early. Differentiated thyroid cancer, including papillary and follicular carcinomas, comprises the majority of cases, while rarer forms such as medullary and anaplastic thyroid carcinomas pose greater challenges in treatment and prognosis. Given the diverse nature of thyroid diseases and their impact on overall health, accurate diagnosis, timely intervention, and ongoing management are essential. This introduction aims to provide an overview of thyroid disorders, highlighting their complexity, clinical manifestations, and significance in healthcare. Subsequent sections will delve into specific aspects of thyroid diseases, including classification, etiology, diagnosis, and treatment modalities, to enhance understanding and inform clinical practice.

Machine learning techniques such as Decision Trees and Random Forests have shown promise in predicting thyroid disease outcomes. These methods utilize computational algorithms to analyze patient data and identify patterns that can aid in accurate disease classification. Decision Trees are simple yet powerful algorithms that construct a tree-like structure to model decisions based on input features. In the context of thyroid disease prediction, Decision Trees can analyze patient data such as thyroid hormone levels, thyroid function test results, imaging findings, and clinical symptoms to classify individuals as either having hyperthyroidism, hypothyroidism, or normal thyroid function.

Decision Trees partition the data based on feature values, iteratively splitting it into subsets that are more homogeneous with respect to the target variable (i.e., thyroid disease classification). Each split is determined by selecting the feature that maximizes the separation between different classes of thyroid disease. Once the tree is constructed, new patient data can be traversed through the tree to predict the likelihood of thyroid disease based on their feature values.

Random Forest is an ensemble learning technique that combines multiple Decision Trees to improve predictive accuracy and robustness. Instead of relying on a single Decision Tree, Random Forest builds a multitude of trees using random subsets of the data and features. Each tree in the forest independently predicts the class of thyroid disease, and the final prediction is determined by aggregating the results of all trees, typically through a majority voting scheme.

Random Forests are particularly effective in handling high-dimensional data and mitigating overfitting, a common issue with individual Decision Trees. By incorporating randomness into the model-building process, Random Forests produce more generalized and stable predictions, making them well-suited for thyroid disease prediction tasks where data variability and complexity are present

2. Literature Review

Arvind Selwal et al. (2020) utilized the multilayer perceptron (MLP) machine learning method to develop an enhanced thyroid disorder prediction system. Their approach categorized individuals into normal, hyperthyroid, and hypothyroid groups based on 7 to 11 characteristics. Training the model with data from 120 individuals gathered from a hospital in Jammu and Kashmir, they employed the gradient descent backpropagation technique. Their predictive model demonstrated high accuracy across 11 criteria, particularly with a larger training dataset.

In a retrospective study, Min Hu et al. (2022) introduced a data mining technique aimed at identifying individuals with hyperthyroidism and hypothyroidism who would benefit from prompt medical attention. They employed four machine learning algorithms to construct a classifier prototype capable of distinguishing between individuals with thyroid disorders and control participants using standard laboratory tests. Performance evaluation involved metrics such as specificity, sensitivity, and the area under the receiver operating characteristic curve (AUROC). Additionally, they utilized feature importance analysis to understand the contribution of each attribute to the model's output, offering rapid and accurate diagnostic assistance using routine laboratory procedures.

Dhyan Chandra Yadav et al. (2020) explored the results of three data mining classifiers: Generalized Linear Model, Boosted Tree, and Neural Network, along with ensemble approaches like Stacking and Random Forest. Their study focused on monitoring performance using different parts of the thyroid dataset, including Thyroxine, Triiodothyronine, and TSH, each exhibiting varied levels. They emphasized the ensemble model as the primary classification tool, with Random Forest yielding the most favorable outcomes. Additionally, they proposed a novel technique employing the Diagnose Odd Ratio (DOR) to detect patterns and determine the necessity of additional therapy based on the patient's DOR levels, streamlining treatment decision-making processes.

In their study, Hafiz Abbad Ur Rehmana et al. (2020) explore the utilization of K-Nearest Neighbor (KNN) and its variants in identifying thyroid disorders. Their research encompasses three distinct phases: KNN without feature selection, KNN with feature selection based on L1 and chi-squared methods. Employing thyroid databases from the KEEL database collection and a registered hospital, they introduced three additional features to enhance the latest database's discriminative capability. Various distance measures were employed to evaluate the KNN model's performance on both datasets, with the Euclidean and Cosine distance functions demonstrating the highest accuracy when utilizing chi-square-based feature selection.

Tehseen Akhtar et al. (2021) focused on integrating three different attribute selection strategies with a homogeneous ensemble voting approach. Pre-processing and exception identification were completed before classification, with bagging and boosting ensemble techniques contributing two algorithms to the first ensemble. The bagging ensembles utilized methods such as random forest and bagging meta estimator (BME), while boosting ensembles employed AdaBoost and XGBoost. The BME outperformed other ensemble techniques in achieving high accuracy with minimal training and prediction time. Additionally, a voting ensemble incorporating hard and soft voting achieved the highest accuracy at a low computational cost, demonstrating the effectiveness of combining feature selection algorithms with various estimation methods and ensemble approaches.

Lerina Aversanoa et al. (2021) aimed to predict the trend of LT4 therapy for hypothyroidism patients. They compiled medical data from patients receiving treatment at a Naples clinic into a specialized dataset, allowing for therapy duration prediction based on hormonal markers and other patient characteristics. Multiple machine learning algorithms were employed, with the Extra-Tree Classifier showing promising results among ten different classifiers.

Md Riajuliislam et al. (2021) sought to predict hypothyroidism in its early stages using three feature selection strategies and various classification techniques. They found that recursive feature selection (RFE) consistently maintained high accuracy across most classification methods, outperforming other feature selection approaches.

Suman Pandey et al. (2015) aimed to improve classification accuracy by distinguishing thyroid from non-thyroid data using various classification approaches and an ensemble model. They demonstrated that feature selection techniques could enhance accuracy and performance, with an ensemble of C4.5 and Random Forest achieving superior accuracy with five features.

Dan Chen et al. (2020) focused on identifying ultrasound characteristics strongly associated with thyroid cancer and designing a scoring system to aid ultrasound physicians in correctly identifying thyroid nodules. Their study utilized logistic regression (LR) analysis and the least absolute shrinkage and selection operator (LASSO) technique to select ultrasound traits highly correlated with cancer. They evaluated the accuracy of different classification techniques using the area under the receiver operating characteristic curve (AUROC).

Nandhinidevi et al. (2020) aimed to improve classification efficacy by identifying relevant attributes before categorization. They utilized a random forest (RF) model to find relevant features and applied the KNN approach for multi-class classification after feature selection, resulting in improved prediction accuracy.

3. Methodology

Feature selection plays a pivotal role in model development, particularly in enhancing the efficiency of classifiers. Adopting feature selection strategies offers several key advantages, including mitigating overfitting, improving accuracy, and reducing training time. Subsequently, Machine Learning classifiers are employed to predict thyroid disorders using these selected features. In this study, the Random Forest technique is introduced as a classifier for forecasting a patient's thyroid condition. Additionally, the Decision Tree classifier is utilized for comparison with the proposed approach. Evaluation and comparison of the best approach for thyroid disease identification are conducted using multiple parameters derived from the confusion matrix at the conclusion of the experiment.

3.1 Dataset

For its extensive use in classification, the data set was acquired from the UCI machine learning repository (UCI, 2014) database. A total of 315 cases are included in the data collection, with 250 being normal, 45 being hyperthyroidism, and 20 being hypothyroidism.

3.2 Decision Tree

Decision Trees represent one of the most commonly utilized architectures in data mining due to their simplicity and interpretability. These structures employ a divide-and-conquer approach to partition the instance space into decision zones. Initially, a test is employed to identify a root node, and subsequently, the dataset is recursively divided based on the value of a selected test attribute until a specified termination condition is met. At the terminal nodes, known as leaf nodes, the class label is assigned. Each branch of the tree represents a decision rule-defined path leading to a leaf node. Subsequently, these decision rules are applied to new samples to classify them.

The utilization of Decision Trees typically involves three primary stages:

1. Learning Process: The initial step involves building the model using the training data, resulting in the formulation of categorization rules that define the decision tree structure.
2. Model Evaluation: In the second stage, the accuracy of the model is assessed by selecting a test, and the model's performance is evaluated based on the outcomes of this test.
3. Model Application: Finally, in the third stage, the trained model is utilized to classify or predict new data samples, providing valuable insights for decision-making.

3.3 Random Forest

The Random Forest (RF) is an ensemble classifier that leverages decision trees, with each tree serving as an individual classifier for classification tasks. It constructs trees randomly using the best fit method and adapts as the number of trees in the forest and the number of attributes per tree vary. Each of the N trees generates a classification output to classify an input sample. The RF then predicts outcomes by averaging the outputs of all trees in the forest. Subsequently, it selects the class with the highest vote count as the final outcome after aggregating all voting results. During each tree split, a randomized set of features is selected, and the tree is exclusively permitted to split based on those features' directions.

The Random Forest algorithm offers several advantages:

1. Enhanced Interpretability: The Random Forest algorithm provides improved interpretability and can effectively handle a large number of predictors.
2. Parameter Significance Ranking: Random Forests can efficiently rank the significance of parameters in regression or classification challenges, aiding in understanding their contribution to the model's predictive performance.
3. Variable Importance Assessment: The algorithm calculates which variables are most significant in the classification process, enabling insights into the factors driving the classification outcomes

4. Results and Discussions

In machine learning, the Confusion Matrix (CM) is a fundamental tool for comprehensive analysis in classification tasks. It provides a structured representation of the actual and predicted classes, with each row and column corresponding to these classes.

The accuracy of a classifier is assessed through metrics such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive and True Negative indicate correct predictions, while False Positive and False Negative signify incorrect predictions.

The accuracy and precision criteria used to evaluate the proposed model are as follows:

Accuracy Prediction:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100\% - - - -(1)$$

Count of iterations	Random Forest %	Decision Tree%
10	85.21	79.28
20	84.22	80.21
30	83.32	80
40	83.21	80.13
50	85.71	80.27
60	86.32	80.37

Table 1: Accuracy Predictions of RF and DT

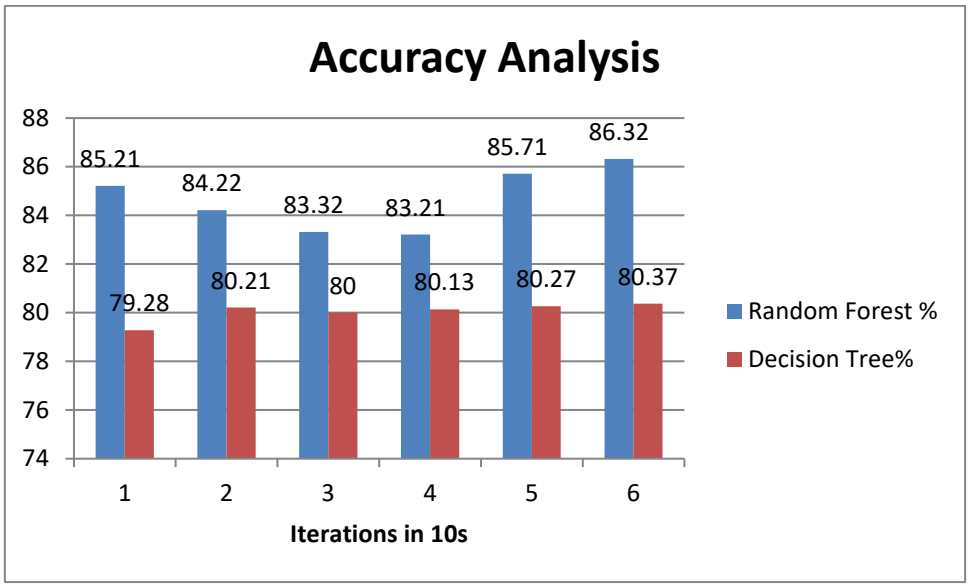

Figure 1: Accuracy Analysis of RF and DT

Table 1 and Figure 1 present the accuracy values obtained from the dataset using both Decision Tree and Random Forest algorithms. The results unmistakably demonstrate that Random Forest outperforms Decision Tree in terms of accuracy.

Precision Prediction:

$$.Precision = \frac{TP}{TP + FP} * 100\% - - - - - - - (2)$$

Count of iterations	Random Forest %	Decision Tree%
10	82.21	76.21
20	86.32	81.02
30	84.64	79.26
40	83.45	84.23
50	85.14	81.54
60	85.12	81.24

Table 2: Precision Predictions of RF and DT

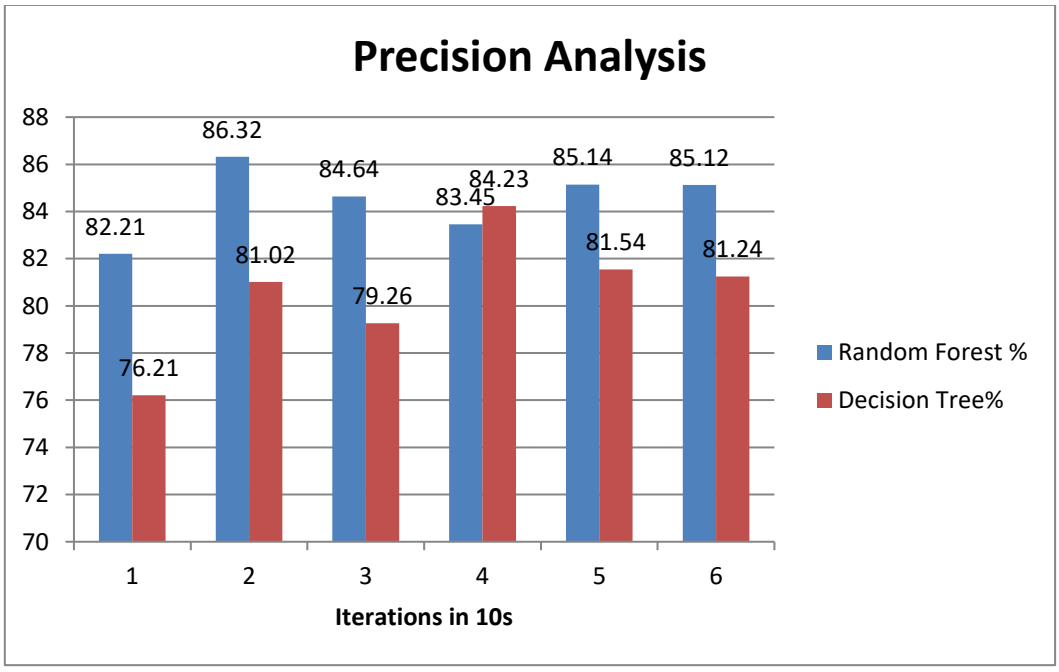

Figure 2: Precision Analysis of RF and DT

Table 2 and Figure 2 depict the precision values derived from the dataset using both Decision Tree and Random Forest algorithms. The results distinctly indicate that Random Forest surpasses Decision Tree in terms of precision.

5. Conclusion

Thyroid hormones play a crucial role in regulating metabolic functions in humans. Leveraging machine learning technologies and clinical investigations has become commonplace in identifying thyroid disorders. This paper proposes the utilization of a Random Forest Classifier for the classification of thyroid gland problems. Comparative analysis with the Decision Tree technique reveals that Random Forest demonstrates superior classification accuracy and precision for the dataset. The thyroid dataset sourced from the UCI machine learning library yielded impressive scores of 84 percent classification accuracy and 85 percent precision. These results

affirm that the Random Forest classifier surpasses the Decision Tree method in prediction accuracy, as demonstrated in this study.

6. References

1. Selwal, Arvind &Raoof, Ifrah. (2020). A multi-layer perceptron based improved thyroid disease prediction system. Indonesian Journal of Electrical Engineering and Computer Science. 17. 524. 10.11591/ijeecs.v17.i1.pp524-532.

2. Hu, Min &Asami, Chikashi&Iwakura, Hiroshi & Nakajima, Yasuyo&Sema, Ryousuke& Kikuchi, Tsuyoshi & Miyata, Tsuyoshi &Sakamaki, Koji & Kudo, Takumi & Yamada, Masanobu &Akamizu, Takashi &Sakakibara, Yasubumi. (2022). Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests. Communications Medicine. 2. 10.1038/s43856-022-00071-1.

3. Yadav, Dhyan& Pal, Saurabh. (2020). Calculating Diagnose Odd Ratio for Thyroid Patients using Different Data Mining Classifiers and Ensemble Techniques. International Journal of Advanced Trends in Computer Science and Engineering. 9. 5463-70. 10.30534/ijatcse/2020/186942020.

4. Ur Rehman, Abbad& Lin, Chyi-Yeu& Mushtaq, Zohaib. (2020). Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. Journal of the Chinese Institute of Engineers. 44. 1-11. 10.1080/02533839.2020.1831967.

5. Akhtar, Tehseen& Gilani, Syed & Mushtaq, Zohaib &Arif, Saad & Jamil, Mohsin & Ayaz, Yasar & Butt, Shahid &Waris, Muhammad. (2021). Effective Voting Ensemble of Homogenous Ensembling with Multiple Attribute-Selection Approaches for Improved Identification of Thyroid Disorder. Electronics. 10. 3026. 10.3390/electronics10233026.

6. Aversano, L. &Bernardi, Mario &Cimitile, Marta &Iammarino, Martina & Macchia, Paolo &Nettore, Immacolata&Verdone, Chiara. (2021). Thyroid Disease Treatment prediction with machine learning approaches. Procedia Computer Science. 192. 1031-1040. 10.1016/j.procs.2021.08.106.

7. M. Riajuliislam, K. Z. Rahim and A. Mahmud, "Prediction of Thyroid Disease(Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 60-64, doi: 10.1109/ICICT4SD50815.2021.9397052.

8. Suman Pandey, AnshuTiwari , Akhilesh Kumar Shrivas and Vivek Sharma. (2015). Thyroid Classification using Ensemble Model with Feature Selection. International Journal of Computer Science and Information Technologies, Vol. 6 (3). pp. 2395-2398. ISSN: 0975-9646

9. Chen, Dan & Hu, Jun & Zhu, Mei & Tang, Niansheng& Yang, Yang & Feng, Yuran. (2020). Diagnosis of thyroid nodules for ultrasonographic characteristics indicative of malignancy using random forest. BioData Mining. 13. 14. 10.1186/s13040-020-00223-w.

10. Nandhinidevi, S. Poorani, P. GokilaBrindha (2020). Machine Learning Models for Relevant Feature Identification and Classification of Thyroid Data. International Journal of Innovative Technology and Exploring Engineering. ISSN: 2278-3075, Volume-9 Issue-5.