2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10723840

Performance Analysis of Machine Learning Algorithms for Thyroid Disease

Aashish Sekar V
Department of DSBS
SRMIST, Kattankulathur
Chennai, Tamil Nadu
aashishlakjay@gmail.com

Dwijavanthi J
Department of DSBS
SRMIST, Kattankulathur
Chennai, Tamil Nadu
dwija2005@gmail.com

Dr. V. Vijayalakshmi
Department of DSBS
SRMIST, Kattankulathur
Chennai, Tamil Nadu
Corresponding Author:
vivenan09@gmail.com

***Abstract*—This paper presents a comprehensive performance analysis of various machine learning algorithms for the prediction of thyroid disease, considering both binary and multi-class classification scenarios. We explore the effectiveness of Support Vector Machine, Decision Tree, Random Forest, Naive Bayes, K-nearest neighbor, XGBoost, and Artificial Neural Network. By leveraging these algorithms and incorporating pertinent risk factors from the dataset, our findings demonstrate stable accuracies across multiple classifiers, with the highest accuracy of 95.6% achieved with the XGBoost in binary classification. However, further analysis reveals that performs poorly in terms of true identification. To address this, rigorous analysis and evaluation have been conducted, and it has been found that through Artificial Neural Network, an accuracy of 94.92% can be achieved in the multi-class classification. In the multi-class classification scenario, similar methodologies were applied, leading to insightful results. We have found that it yields a recall score of 90.4%. Through visualization of training and testing accuracies, phenomena such as model overfitting and underfitting are investigated to understand the limitations of specific classifiers. The primary objective of this research is to identify optimal algorithms in terms of accuracy and computational efficiency for thyroid disease prediction, considering both binary and multi-class classification tasks.**

Keywords—Thyroid disease, Machine Learning, Support Vector Machine, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbor,XGBoost, ANN.

I. INTRODUCTION

The Thyroid gland produces hormones that impact every cell, tissue, and organ in the body. Thyroid Hormones control the body's metabolism. Iodine is required for the glands to secrete the hormones such as thyroxine and triiodothyronine which helps with protein synthesis and producing enzymes for the body. The International Council for Control of Iodine Deficiency Disorder and the World Health Organization (WHO) state that adults should consume 150–200 mcg of iodine per day, with a median excretion of less than 100 mcg per day.

Thyroid disease is a condition characterized by the thyroid gland overproducing or underproducing thyroid hormones. This leads to hormonal imbalance which affects the body. The frequency of thyroid diseases increases with age. This is due to an increase in levels of thyroid stimulatinghormone (TSH) increases with age along with T4 metabolism changes. However, the effects of these changes on physiology are unclear as there are contradictory reports published. Hypothyroidism is a condition that occurs when the glands do not produce the required thyroid hormone. Symptoms such as fatigue, depression, and weight gain are the most common ones. Hyperthyroidism is the opposite. It occurs when the gland overproduces thyroid hormones. Nervousness, irritability, weakness, and weight loss are some common symptoms.

In 2017, almost 200 million people were affected by thyroid disease with 40% of the world's population being at risk. This represents a substantial increase when compared to previous decades. In 1980, only 100 million people were affected by thyroid disease. The prevalence of thyroid diseases has increased significantly according to statistics.

Measuring thyroid stimulating hormones (TSH), thyroxine(T4), triiodothyronine (T3), and thyroid antibodies by blood tests, Ultrasounds, and thyroid scans can also be used to visualize the thyroid glands to detect abnormalities and diagnose thyroid disease.

The subsequent sections of the research paper are designed to systematically address the key elements of the study. The Literature Review section provides an overview of existing research related to the topic, identifying trends, gaps, and the foundation upon which the current study is built. The Methodology section details the research design, data collection methods, and analysis techniques used, ensuring the study's reproducibility and reliability. In the Results section, the findings of the study are presented in a clear and concise manner, often supported by tables, figures, and statistical analysis. The Discussion section interprets the results, linking them back to the literature review, highlighting the implications, and suggesting areas for future research. Finally, the Conclusion summarizes the study's contributions, reinforces its significance, and offers final thoughts on the research topic.

II. LITERATURE SURVEY

The application of machine learning (ML) algorithms in healthcare, particularly in thyroid disease analysis, has garnered significant attention in recent years. [1] Bini et al. emphasize

the potential of AI, particularly machine learning, to revolutionize thyroid disease analysis and patient care. However, they acknowledge the need for further research to address challenges like data quality, interpretability of ML models, and responsible clinical integration. By leveraging artificial intelligence (AI) techniques, researchers have aimed to enhance diagnostic accuracy and streamline patient care processes. Laboratory testing remains a cornerstone in thyroid disease management, but some pitfalls and nuances must be considered for optimal clinical utility. [2] by Soh and Aw discusses laboratory testing in thyroid conditions. It focuses on the limitations and advantages, also known as pitfalls and clinical utility, of these tests in diagnosing thyroid problems. ultrasound appears to be a valuable tool for diagnosing thyroid nodules, but it may not be definitive in all cases. If you have any concerns about a thyroid nodule, consult a healthcare professional for proper diagnosis and treatment [3,4].

Various ML techniques, including selective feature analysis and hybrid classification systems, have been explored for thyroid disease prediction, showcasing diverse levels of accuracy and efficiency [5, 6] this research highlights the potential of machine learning with feature selection for accurate thyroid disease prediction, potentially offering a valuable tool for medical professionals. Researchers have investigated the effectiveness of different algorithms such as Naive Bayes, decision trees, support vector machines (SVM), and regression trees in classifying thyroid conditions [7, 8, 9, 10]. Traditionally, diagnosing thyroid conditions relies on physical exams, medical history, and blood tests. ML offers a potential tool to analyze data and predict thyroid disease, potentially aiding early detection and treatment.

Moreover, comparative studies evaluating the efficacy of different ML algorithms, such as K-nearest neighbors (KNN) and decision trees (C4.5), have shed light on the strengths and limitations of each approach [11]. In addition to diagnostic applications, ML techniques have been employed for data management and analysis in healthcare systems, facilitating efficient data exchange, archiving, evaluation, and decision-making processes [12].

Overall, the literature underscores the evolving landscape of ML applications in thyroid disease analysis, emphasizing the need for comprehensive performance analysis and methodological evaluations to optimize diagnostic accuracy and clinical outcomes. Further research in this area holds the potential to revolutionize thyroid disease management and improve patient care pathways.

III. Methodology

This title deals with the workflow diagram, the methodology, and the algorithm we used to classify thyroid diseases. Machine Learning Algorithms used are Support Vector Machine, Decision Tree, Random Forest, Naive Bayes, K-Nearest Neighbor, XGBoost, and Artificial Neural Networks. Figure 1 shows the workflow diagram.

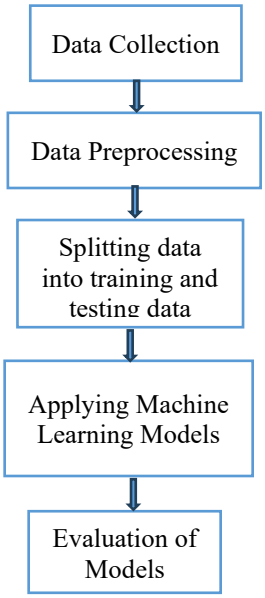

Fig. 1. Workflow Diagram

A. Procedure

The procedures as follows:

Step 1: Data Collection

Step 2: Data Preprocessing

Step 3: Exploratory Data Analysis

Step 4: Splitting the data into 20% for testing and 80% fortraining.

Step 5: Model Building using the training data.

Step 6: Model Evaluation with the testing data.

Step 7: Comparison of models result.

B. Data Collection

The Thyroid disease dataset used is from the Garvan Institute and presents a comprehensive array of approximately 2800 instances each. This dataset stands valuable for the detailed exploration of patterns and relationships in thyroid-related health data. Patient details, encompassing gender, medical treatments(e.g., on_thyroxine, thyroid_surgery), and measured values (TSH, T3, TT4, T4U, FTI, TBG), contribute to the dataset's comprehensiveness. The dataset used, incorporates both continuous and categorical variables, including binary indicators like 'on_thyroxine,' coveringa spectrum of negative and positive thyroid conditions. The inclusion of hypothyroid and hyperthyroid variants further expands the dataset's dimensions, providing a comprehensive representation of thyroid health states.

C. Data preprocessing

Data preprocessing is a crucial step in the data analysis pipeline where raw data is transformed, cleaned, and organized to make it suitable for further analysis. We firstread multiple CSV data files containing the data into the pandas data frame. We displayed and deleted multiple rows and columns

containing null and duplicate values. We eliminated rows and columns with insufficient data for analysis and Standardized categorical values for consistent Testing and Training of the Data. We mapped categorical variables to numerical values for model compatibility and interpolated missing values for a more complete dataset.

D. Data Exploration

We conducted a thorough exploration of the dataset to gain insights into the distributions of variables and their interrelationships. Initially, we employed bar graphs to visually represent the distribution of each variable withinthe dataset. This allowed us to observe the frequency or proportion of data points falling within distinct categoriesor ranges. Additionally, to understand the correlations between variables, we constructed a correlation matrix as shown in Figure 2. This matrix provided a comprehensive overview of the relationships between pairs of variables, highlighting potential dependencies or associations. The utilization of bar graphs and correlation matrices facilitated a comprehensive understanding of the dataset's characteristics, aiding in subsequent analysis and interpretation of results.

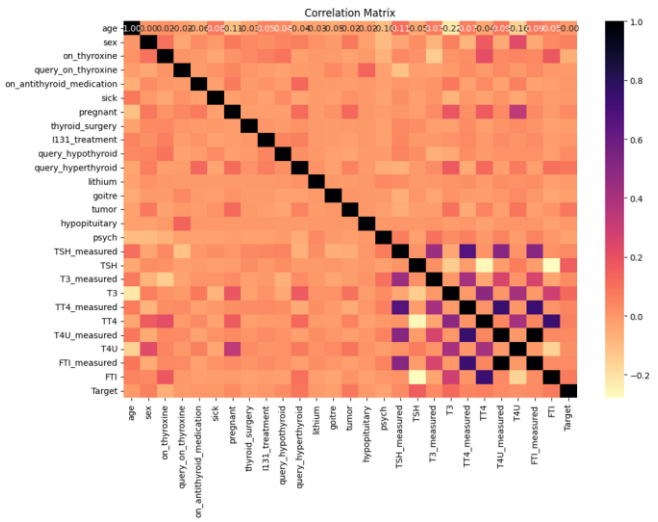

Fig. 2. Correlation Diagram

E. Model Training

To assess the generalization performance of the models, we partitioned the dataset into training and testing sets using a 20-80 split ratio, ensuring sufficient data for both training and evaluation. This approach helps assess the models' ability to generalize to unseen data. We applied machine learning algorithms for classification tasks and deep learning models for tasks requiring enhanced complexity. Both training and testing sets were utilized to train and evaluate machine learning models.

Support Vector Machine (SVM): SVM is a classification algorithm that uses hyperplanes to separate classes in a high-dimensional space.

Decision Tree: Decision trees divide the feature space into segments to enable sequential decisions, rendering them useful for tasks such as regression and classification. Random Forest: Random Forest is a method of ensemble learning that trains multiple decision trees and combines their predictions for enhanced accuracy and robustness.

Naive Bayes: Naive Bayes is a probabilistic machine learning algorithm that commonly uses Bayes' theorem to predict a given input's classification based on feature independence.

K-nearest neighbor: KNN is a non-parametric algorithm for classification that predicts a new data point's class or value by relying on the majority class or mean value of k nearest neighbours in the training dataset.

XGBoost: XGBoost is an optimized implementation of gradient boosting, incorporating regularization techniquesand parallel computing to improve speed and performance.

IV. Result Discussion

Binary Classification:

Binary classification is needed to check whether particular person is having thyroid disease or not. We carried out this experiment to minimize the false negative, as negligence could lead to major consequences. Through this experiment, we analyzed that the models could not accurately differentiate between Hyperthyroidism and Hypothyroidism. XGBoost showed exceptional accuracy and a great recall score, minimizing the falsest negative cases. Table 1 explains the binary class confusion matrix for all the models.

TABLE I. Binary Classification Confusion Matrix

Model Name		Predicted Values	
		Negative	Positive
Random Forest	Actual Values	1295	70
		67	6
Support Vector Classifier		1365	0
		73	0
KNearest Neighbors		1214	151
		7	66
XGBoost		**1359**	**63**
		57	**16**
Bernoulli Naive Bayes		1359	6
		72	1
Decision Trees		1297	68
		63	10
Artificial Neural Network		1365	0
		72	1

Best Model for Identifying Positives: KNearest Neighbors (KNN) due to its high recall (90.4%), indicating it correctly identifies most positive cases. Best Balanced Model: XGBoost with high accuracy (95.6%) and a better balance of precision (20.3%) and recall (21.9%) compared to other models. Models with High Accuracy but Poor Recall: SVC, Bernoulli Naive Bayes, and ANN have high accuracy but fail to identify positive cases effectively. Random Forest and Decision Trees have moderate performance with balanced but low precision and recall for the positive class.

Multi Classification:

We evaluated the performance of support vector machines, decision trees, random forests, naive Bayes, k-nearest neighbors, XGBoost, decision trees, and Artificial neural networks. We use a wide range of metrics to assess the effectiveness and outcomes of these algorithms, such as accuracy, precision, recall, and F1-score. We analyze the strengths and weaknesses of each approach, highlighting the advantages and potential limitations of each algorithm. To gain deeper insights and effectiveness of different machine learning models in predicting outcomes using confusion matrices. Table 2 shows the multi class confusion matrix.

TABLE II. MULTI CLASSIFICATION CONFUSION MATRIX

Model Name		**Predicted Values**		
		Negative	**Hypothyroid**	**Hyperthyroid**
Random Forest	**Actual Values**	1295	48	22
		49	5	0
		18	0	1
Support Vector Classifier		1365	0	0
		54	0	0
		19	0	0
K Nearest Neighbors		1345	13	7
		51	3	0
		19	0	0
XGBoost		1302	42	21
		39	15	0
		18	0	1
Bernoulli NaiveBayes		1359	0	6
		53	1	0
		19	0	0
Decision Tree		1297	45	23
		44	10	0
		19	0	0
Artificial Neural Network		**1365**	**0**	**0**
		54	**0**	**0**
		18	**1**	**0**

These matrices provide a simple breakdown of true positives, true negatives, false positives, and false negatives which helps us to understand how accurately each model distinguishes between different categories. The Artificial Neural Network (ANN) has the highest overall accuracy at 94.92%. However, it performs poorly in detecting Hypothyroid cases, similar to the Support Vector Classifier. The models show varying performance across different classes, indicating that a model's suitability may depend on which class(es) are of most interest.

ROC and PR Curve:

Examining the Receiver Operating Characteristic (ROC) curve aids in evaluating the balance between true positive and false positive rates for each model, while Precision-Recall curves differentiate precision and recall, offering a detailed insight into model performance and emphasizing classification imbalances.

We evaluate and compare the performance of the considered models using both Receiver Operating Characteristics (ROC) curves and Precision-Recall (PR) curves. This allows us to assess the trade-off between true positive rate and false positive rate (ROC) and between precision and recall (PR) across different classification thresholds. By analyzing both curves, we gain a comprehensive understanding of each model's ability to correctly identify positive and negative instances while considering the class imbalance.

Purpose of ROC Curves: These curves depict the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) for various classification thresholds. TPR represents the model's ability to correctly identify positive cases (individuals with thyroid disease), while FPR indicates the rate of incorrectly classifying healthy individuals as having the disease. Expected Findings in this Study: The study focused on minimizing false negatives (correctly identifying positive cases).

Analyzing the ROC curves for each model would likely reveal: KNN: The ROC curve for KNN might show a steeper rise in TPR compared to FPR at lower thresholds, reflecting its high recall (ability to identify positive cases) observed in the binary classification results. Bernoulli Naive Bayes: The ROC curve for Bernoulli Naive Bayes might show a good balance between TPR and FPR across various thresholds, aligning with its strong

overall performance in multi-class classification. Other Models: The curves for models like SVC, ANN, Random Forest, Decision Tree, XGBoost, and KNN might exhibit varying trade-offs between TPR and FPR, potentially explaining their differences in performance compared to Bernoulli Naive Bayes.

Purpose of PR Curves: These curves depict the relationship between Precision (correctly identified positive cases) and Recall (identifying all positive cases) for different classification thresholds. Expected Findings in this Study: The PR curves would likely complement the ROC curves and potentially reveal: KNN: The PR curve for KNN might show a high and stable precision value, reflecting its focus on minimizing false negatives. Bernoulli Naive Bayes: The PR curve for Bernoulli Naive Bayes might show a good balance between precision and recall across various thresholds, again aligning with its strong overall performance. Other Models: The curves for other models might exhibit variations in the trade-off between precision and recall, potentially offering insights into their strengths and weaknesses compared to Bernoulli Naive Bayes.

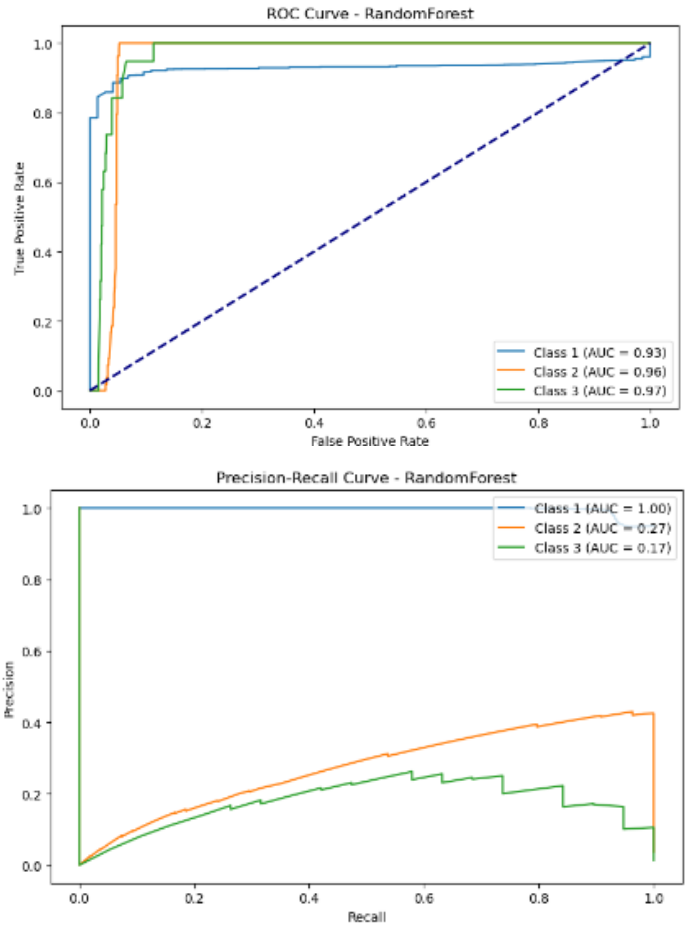

Fig. 3. ROC-PR Curve

Figure 3 shows the Receiver Operating Characteristics (ROC) curves and Precision-Recall (PR) curves for Random forest model. The results revealed distinct strengths and weaknesses among the evaluated models. The performance matrix for all the models given in table 3 with precision, recall, f1-score and accuracy.

TABLE III. MULTI CLASSIFICATION PERFORMANCE MATRIX

Model	Precision	Recall	F1-Score	Accuracy
Random Forest	90.67%	90.47%	90.57%	90.47%
Support Vector Classifier	90.10%	94.92%	92.45%	94.92%
K Nearest Neighbors	90.93%	93.74%	92.17%	93.74%
XGBoost	91.99%	91.66%	91.82%	91.66%
Bernoulli Naive Bayes	93.90%	94.58%	92.41%	94.58%
Decision Tree	91.21%	90.89%	91.05%	90.89%
Artificial Neural Network	90.17%	94.92%	92.48%	94.92%

Bernoulli Naive Bayes stands out with high precision, recall, F1-score, and accuracy, indicating strong performance overall. Support Vector Classifier and Artificial Neural Network also perform well, especially in terms of recall and accuracy. Random Forest and Decision Tree perform comparably to the other models but have slightly lower precision and recall. XGBoost and K Nearest Neighbors perform well but are slightly behind Bernoulli Naive Bayes, SVC, and ANN in terms of precision, recall, and F1-score.

V. CONCLUSION

Our aim of this study was to identify the most reliable algorithm for thyroid classification. Our focus was to minimize false positive cases to prevent severe consequences. After implementing the thyroid prediction using the various models. Our results underscore the importance of selecting appropriate machine learning algorithms for medical classification tasks, particularly in scenarios where false negatives can have significant repercussions. The findings from this study provide valuable insights for clinicians and researchers seeking to implement reliable models for thyroid disease diagnosis. Our analysis demonstrates the potential of machine learning for classifying thyroid diseases. Different algorithms offer varying strengths and weaknesses. KNN excelled at identifying positive cases in binary classification, while XGBoost offered a balanced performance. In multi-class classification, Bernoulli Naive Bayes stood out with the best overall performance. Future work could explore hyperparameter tuning for further performance optimization and investigate the integration of feature engineering techniques to potentially improve model performance. It's crucial to remember that these models are for predictive purposes and should not replace medical professional judgment.

REFERENCES

[1] Bini F, Pica A, Azzimonti L, Giusti A, Ruinelli L,Marinozzi F, Trimboli P. Artificial Intelligence in Thyroid Field—A Comprehensive Review. Cancers. 2021; 13(19):4740.

[2] Soh S, Aw T. Laboratory Testing in Thyroid Conditions-Pitfalls and Clinical Utility. Ann Lab Med 2019;39:3-14.

[3] Zhao, C.K., Ren, T.T., Yin, Y.F., Shi, H., Wang, H.X., Zhou, B.Y., Wang, X.R., Li, X., Zhang, Y.F., Liu, C. andXu, H.X., 2021. A comparative analysis of two machinelearning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate. Thyroid, 31(3), pp.470-481.

[4] Shi M, Nong D, Xin M, Lin L. Accuracy of Ultrasound Diagnosis of Benign and Malignant Thyroid Nodules: A Systematic Review and Meta-Analysis. Int J Clin Pract. 2022 Sep 13;2022:5056082.

[5] Chaganti R, Rustam F, De La Torre Díez I, Mazón JLV, Rodríguez CL, Ashraf I. Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques. Cancers (Basel). 2022 Aug 13;14(16):3914.

[6] D. Bhende and G. Sakarkar, "Performance Evaluation of Machine Learning Methods for Thyroid Prediction," 2023 11th International Conference on Emerging Trendsin Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6

[7] Ur Rehman, Abbad & Lin, Chyi-Yeu & Mushtaq, Zohaib & Su, Shun-Feng. (2021). Performance Analysis of Machine Learning Algorithms for Thyroid Disease. Arabian Journal for Science and Engineering.

[8] Razia, Shaik & SwathiPrathyusha, P. & Krishna, N & Sumana, N. (2018). A Comparative study of machine learning algorithms on thyroid disease prediction. International Journal of Engineering & Technology.

[9] Z. J. Peya, M. Shymon Islam and M. K. Naher Chumki, "Thyroid Disease Prediction based on Feature Selection and Machine Learning," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 495-500,

[10] M. Pal, S. Parija and G. Panda, "Enhanced Prediction ofThyroid Disease Using Machine Learning Method," 2022 IEEE VLSI Device Circuit and System (VLSIDCS), Kolkata, India, 2022, pp. 199-204

[11] Khalid Salman and Emrullah Sonuç, “Thyroid Disease Classification Using Machine Learning Algorithms”2021 J. Phys.: Conf. Ser. 1963 012140.

[12] Islam SS, Haque MS, Miah MSU, Sarwar TB, Nugraha R. Application of machine learning algorithms to predictthe thyroid disease risk: an experimental comparative study. Peerj Comput Sci.2022 Mar 3;8e.