Al-Salam Journal for Engineering and Technology
Journal Homepage: http://journal.alsalam.edu.iq/index.php/ajest
e-ISSN: 2790-4822 p-ISSN: 2958-0862

Enhanced Thyroid Disease Prediction Using Ensemble Machine Learning Techniques

Hadeel Alkhazzar[1]*

[1]Department of Computer and Communication, Collage of Engineering, American University of Science and Technology, Lebanon.

*Corresponding Author: Hadeel Alkhazzar

DOI: https://doi.org/10.55145/ajest.2025.04.01.018
Received September 2024; Accepted November 2024; Available online December 2024

ABSTRACT: There is an urgent need for precise and effective diagnosis techniques since thyroid diseases are becoming more common and have a major negative influence on public health. Conventional diagnostic methods frequently lack speed and accuracy, which causes therapy delays and unfavorable patient outcomes. The purpose of this study is to improve the predicted accuracy of thyroid illness detection by utilizing ensemble machine learning techniques. The dataset was preprocessed to ensure cleanliness and format compatibility with machine learning algorithms. Four classifiers were used to evaluate their predictive capabilities: Bagging Classifier, Support Vector Machine (SVM), AdaBoost, and Gaussian Naive Bayes. Bagging Classifier, utilizing 50 Decision Tree estimators, emerged as the most effective model. Bagging achieved the highest accuracy (99.87%), followed closely by AdaBoost (98.87%). SVM and Naïve Bayes performed comparatively lower, with accuracy scores of 94.58% and 93.69%, respectively. These findings highlight the effectiveness of ensemble methods for accurate and reliable thyroid disease prediction. These advancements could play a crucial role in developing automated tools for early detection and monitoring of thyroid disorders, ultimately improving patient outcomes and streamlining healthcare processes.

Keywords: Thyroid Disease Prediction, Decision Tree, Machine Learning, Bagging Classifier

1. INTRODUCTION

Thyroid disease is one of the most prevalent endocrine disorders globally [1]. According to [2], 4.6% of adults over the age of 12 have hypothyroidism, 1.2% of people in the USA by myself have hyperthyroidism, and almost 10% of humans in India have this situation, that's getting worse each day. The thyroid gland in the human body is responsible for producing essential hormones and maintaining metabolism. Thyroid disorders are linked to heart disease, joint discomfort, and infertility. Although some of the signs may additionally occur early, these conditions often develop later in life. The thyroid gland produces key hormones: levothyroxine (T4) and triiodothyronine (T3). T4 is the inactive form, which is transformed into the lively T3 in tissues, regulating metabolism, energy ranges, and growth. Both hormones are vital for physical capabilities. Irregularities within the thyroid gland can bring about big health troubles, with hyperthyroidism and hypothyroidism being of the most common situations [3]. Hyperthyroidism takes place while the thyroid produces too much hormone, main to symptoms like weight loss and fast heartbeat. Hypothyroidism, alternatively, entails inadequate hormone manufacturing, causing fatigue, weight benefit, and slowed metabolism. Both situations require well timed prognosis and treatment to control symptoms effectively.

Without early detection, thyroid disorders like hypothyroidism and hyperthyroidism can have severe negative impacts on health. Untreated instances of these issues, which effect the body's metabolism, can bring about severe side consequences like heart ailment, infertility, and mental health issues. Conventional diagnostic techniques rely on clinical evaluation and blood tests, which are frequently slow and prone to misunderstanding despite their effectiveness. More dependable and quick diagnostic tools are needed because of the increasing complexity of healthcare and the requirement for accuracy.

To facilitate prompt interventions, this study intends to improve the predictive accuracy of thyroid illness diagnosis by the application of machine learning (ML) techniques. Evaluation and comparison of many classifiers, with an emphasis on ensemble approaches, is our main objective in order to ascertain how well they handle intricate medical

*Corresponding author: *Hadeelw264@gmail.com* datasets. The objective is to create a trustworthy predictive model that can enhance thyroid disease early diagnosis and treatment, which will benefit patients and medical professionals.

The rest of the paper is organized into the following sections, where section 2 reviews related works in the diagnosis of thyroid diseases. Section 3 outlines the methodology, detailing data preprocessing and model development, with an emphasis on ensemble learning techniques. The results are discussed in section 4, where the performance of the models is compared. Finally, section Five concludes the study, summarizing the findings and suggesting avenues for future work.

2. RELATED WORK

Various methods for diagnosing thyroid disorders have been proposed in the literature. Proactive thyroid disease prediction is crucial for timely treatment, which can save lives and reduce medical costs. Technological developments in data processing and computation have led to the application of machine learning and deep learning (DL) approaches to identify thyroid illness categories, such as hypothyroidism and hyperthyroidism, and to detect thyroid diagnosis in its early stages. Numerous studies in the literature address the diagnosis of thyroid disorders using patient personal information, including age and sex, as well as hormonal measures. Notably, deep neural network models are used in some research, while ML models are used in others for classification and prediction.

2.1 MACHINE LEARNING APPROACHES

The use of levothyroxine sodium (LT4), a thyroid hormone, in the treatment of thyroid diseases was investigated in [4] using a machine learning approach. The patient's hypothyroidism was examined and an attempt was made to predict how LT4 would be treated. The “AOU Federico II” dataset was created using 27 features and using ten distinct classifiers and the tree classifier obtained a maximum accuracy of 84%. Many studies address the identification of thyroid diseases using hormonal parameters and patient personal data, such as age and gender with machine learning models. Several methods, including logistic regression, random forests, and others, were used in [5] to predict the three main thyroid types (normal, hyperthyroid, and hypothyroid) and achieved a maximum accuracy of 96.4%. Utilizing machine learning algorithms, [6] found that the prediction accuracy of hyperthyroidism and hypothyroidism was 90.9% and 93.8%, respectively, in order to differentiate between patients with these conditions. In [7], a multi-kernel SVM is suggested for the classification of thyroid disorders. The method performed with 97.49% accuracy on UCI thyroid datasets. Feature selection has been implemented to improve performance through enhanced gray wolf optimization.

2.2 ENSEMBLE LEARNING APPROACHES

Ensemble learning, which is a type of machine learning, ensures the reliability and effectiveness of the prediction model because a number of classifiers participate in decision making. A public thyroid disease dataset containing 29 clinical features from the UC Irvine ML repository was used for prediction purpose in [8]. Boosting approach has been used to achieve accurate detection, which can be used with real-time computer-aided diagnosis (CAD) systems to speed up diagnosis and encourage early treatment. An XGBoost model was applied in [9] to predict thyroid in a UC Irvin knowledge discovery dataset. XGBoost emerged as the top-performing model with the highest accuracy after a comparison of the suggested model's accuracy with that of k-nearest neighbors (KNN), logistic regression, and decision trees. In [10], various ML techniques, including artificial neural networks (ANN), KNN, random forest algorithm, and decision tree, are used to the dataset to produce a comparative analysis that improves disease prediction based on the parameters chosen from the dataset. The dataset was also improved to accurately predict the classification. After processing the data set, the highest accuracy of the random forest algorithm was obtained, 94.8%.

2.3 DEEP LEARNING APPROACHES

Other approaches use deep neural network models. For the classification of thyroid disorders, a multiple multi-layer perception (MMLP) method was put forth in [11]. Applying MMLP with a six-grid ensemble improves accuracy by 0.7% when compared to using just one MLP. Despite achieving up to 99% classification accuracy on big dataset samples, DL approaches such as MMLP require considerable computational resources and are costly to train. The study [12] focuses on the prediction of hypothyroid illness using a new hybrid DL artificial neural network prediction approach based on Long Short Term Bidirectional Associative Memory (LSTBAM). The dataset that was delivered has 29 qualities and factors that illustrate the characteristics of the condition. A thyroid nodule classification method based on feature fusion and deep learning approaches was suggested in another study [13]. The system used a combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to achieve high accuracy (95.2%), sensitivity (93.1%), and specificity (96.8%) on the dataset, which included 5,310 ultrasound pictures of thyroid nodules.

For real-time applications, high-performance deep learning models are less practical due to their high computing resource requirements. Although ensemble approaches and conventional machine learning models are computationally effective, they frequently fall short in terms of accuracy for critical diagnoses.

3. METHODOLOGY

Ensemble learning's basic concept is that combining predictions from various learners can produce more solid and trustworthy results [14]. The idea of ensemble learning mirrors real-life decision-making scenarios. For instance, when diagnosing a disease, a medical team may consult multiple specialists to pool their expertise, combining perspectives to reach a robust and reliable conclusion. Similarly, ensemble strategies aggregate the predictions of multiple fashions, making sure that the final choice is greater correct and less liable to individual biases or mistakes.

Three main categories can be used to group different ensemble learning techniques (Bagging, Boosting, Stacking):

• Bagging, or Bootstrap Aggregating, is used to reduce variance. It is much like strolling surveys among exceptional subsets of data to reduce uncertainty in the consequences by way of training a couple of times of the same algorithm on numerous subsets of the dataset. As an example of these algorithms, Random Forest, a well-known bagging technique, creates multiple decision trees using subsets of the data. Each tree predicts the final results independently, and the values are averaged (for regression) or determined through a majority vote (for classification). This aggregation reduces the variance making the model sturdy.

• Boosting works similarly to a teacher giving pupils who are having trouble with a certain subject greater attention. For example, a math instructor may decide to focus more on geometry in later courses if they observe that students are consistently struggling with it. Similar principles are followed by boosting algorithms like AdaBoost, which train models one after the other with an emphasis on fixing the mistakes of the earlier models.

• Because it increases prediction dependability and achieves superior accuracy across a variety of datasets, ensemble learning has shown itself to be very useful in classification issues, consistently outperforming single models [15].

Figure 1 shows the steps used to predict thyroid diseases. The methodology begins with collecting data containing useful information for detecting thyroid diseases and distinguishing patients from healthy people in the early stages. The data is then processed to become suitable for machine learning models. A number of models will be created and trained to choose the best. Essential libraries have been imported such as numpy and pandas to efficiently handle numerical data, StandardScaler to normalize data and enhance the performance of ML algorithms, and matplotlib to visualize results.

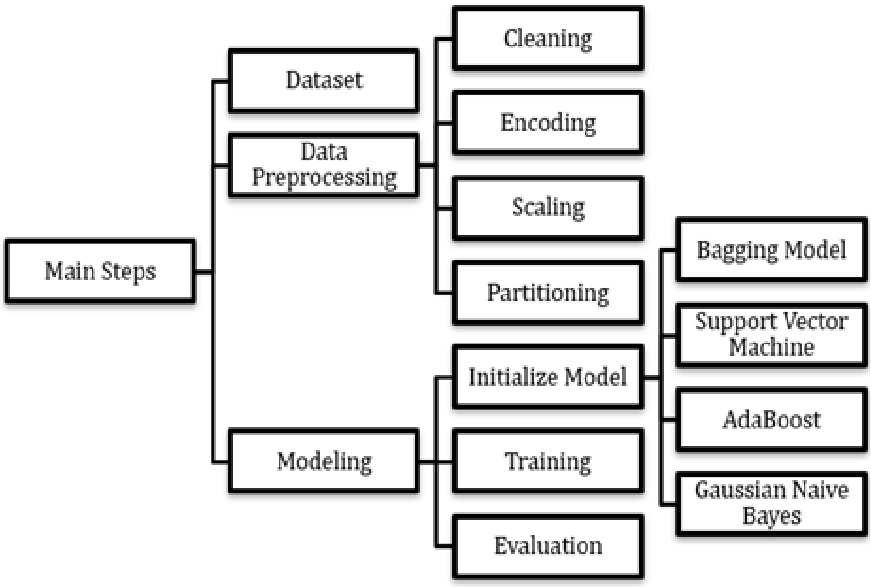
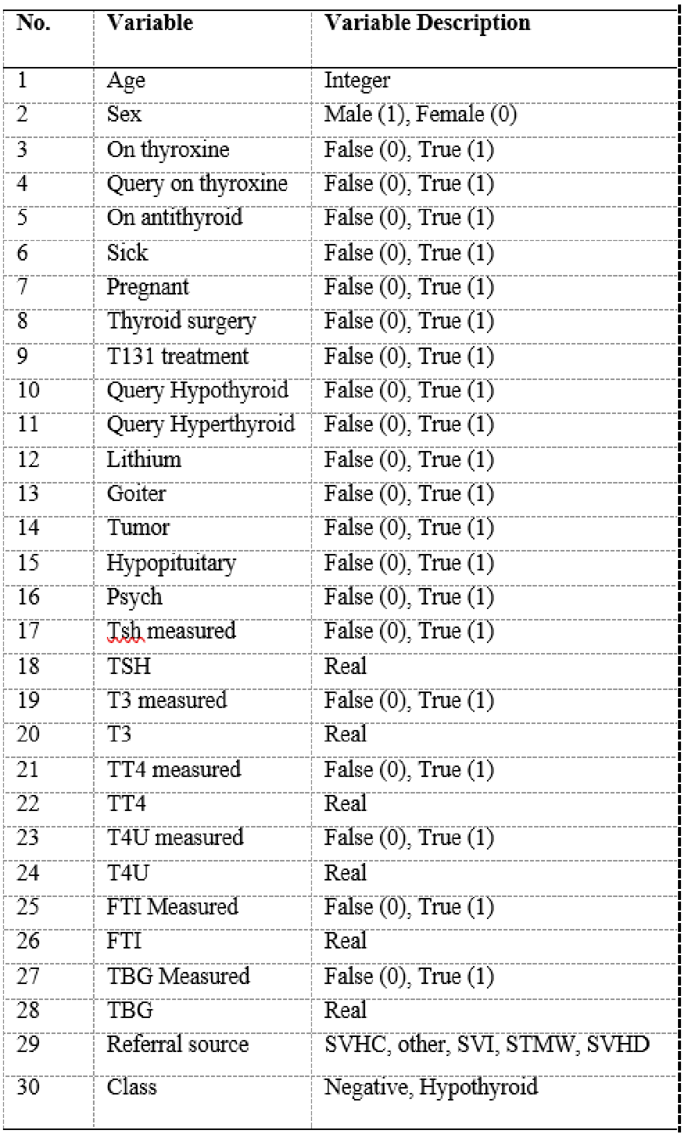


FIGURE 1. - **Proposed methodology for predicting thyroid diseases**

3.1 DATASET

"Thyroid Disease Dataset" dataset was collected from kaggle.com [16]. Of 3772 samples, 291 (7.7%) were normal (negative) and 3481 (92.3%) were hypothyroid (positive). The dataset's features are listed in Table 1.

Table 1. - Dataset features



No.	Variable	Variable Description
1	Age	Integer
2	Sex	Male (1), Female (0)
3	On thyroxine	False (0), True (1)
4	Query on thyroxine	False (0), True (1)
5	On antithyroid	False (0), True (1)
6	Sick	False (0), True (1)
7	Pregnant	False (0), True (1)
8	Thyroid surgery	False (0), True (1)
9	T131 treatment	False (0), True (1)
10	Query Hypothyroid	False (0), True (1)
11	Query Hyperthyroid	False (0), True (1)
12	Lithium	False (0), True (1)
13	Goiter	False (0), True (1)
14	Tumor	False (0), True (1)
15	Hypopituitary	False (0), True (1)
16	Psych	False (0), True (1)
17	Tsh measured	False (0), True (1)
18	TSH	Real
19	T3 measured	False (0), True (1)
20	T3	Real
21	TT4 measured	False (0), True (1)
22	TT4	Real
23	T4U measured	False (0), True (1)
24	T4U	Real
25	FTI Measured	False (0), True (1)
26	FTI	Real
27	TBG Measured	False (0), True (1)
28	TBG	Real
29	Referral source	SVHC, other, SVI, STMW, SVHD
30	Class	Negative, Hypothyroid

Hypothyroid Disease Data Set consists of 29 features and one outcome variable used to predict thyroid disease. The features include both categorical and numerical variables, with descriptions indicating patient attributes and medical history. Key features are demographic, such as age (integer), sex (binary), and several binary medical indicators like on thyroxine, query hypothyroid, and goiter. Clinical measurements such as TSH, T3, TT4, T4U, and FTI are recorded as real values. Categorical variables, such as the referral source, and binary indicators like psych and pregnant, also play significant roles in the analysis. The outcome variable, Class, differentiates between hypothyroid (92.3% of cases) and negative cases, which account for 7.7%.

3.2 DATA PREPROCESSING

The preprocessing phase involves preparing the data set to ensure it is clean, organized, and suitable for machine learning models:

Cleaning: There were initially missing values in a number of the dataset's columns, most notably the TBG (Thyroxine-Binding Globulin) column. As a result of its many missing values, TBG was eliminated from the dataset. In order to preserve data integrity, rows possessing missing values in crucial columns like TSH, T3, TT4, and T4U were also located and eliminated. Furthermore, columns containing only one unique value are dropped because they do not contribute to the prediction process. After this cleaning, the dataset was reduced to 2643 rows and 23 columns, ensuring that it only contained complete and meaningful data.

Encoding: Categorical variables need to be converted into numerical formats that models can process effectively. In one-hot encoding, binary dummy variables are created from categorical variables, where each category is represented by a distinct column with a binary value (0 or 1) to ensure that the model can handle categorical data.

Columns with fewer than 10 unique values are subject to encryption, while columns with more than these values are subject to scaling.

Scaling: This process converts numeric values to a mean of 0 and a standard deviation of 1, improving model performance and ensuring features are on a comparable scale. Numerical features are standardized using the StandardScaler library. After scaling, the dataset retains 2643 rows and 25 columns.

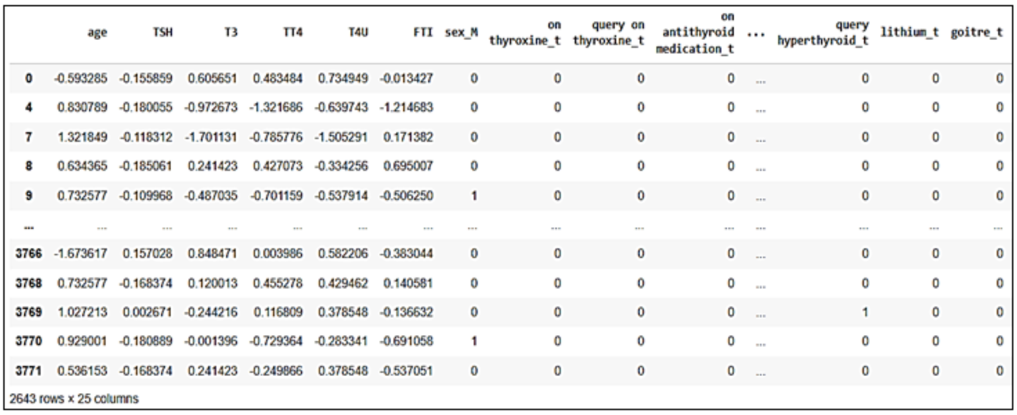
Partitioning: The target column, which contains the values "P" and "N," is transformed into binary values where "P" is replaced with 1 (positive) and "N" with 0 (negative), making it suitable for binary classification. Following that, the dataset was divided into training and testing sets, with 30% going toward testing and 70% going toward training. To guarantee that the instances were distributed randomly, the data was scrambled before splitting. The training set contains 1850 rows and 25 columns, while the test set contains 793 rows and 25 columns.

Table 2 summarizes the key preprocessing steps taken to prepare the thyroid dataset for machine learning models. Each step outlines the actions performed, along with their impact on the number of rows and columns in the dataset, ensuring data integrity and suitability for classification tasks.

Table 2. - Basic preprocessing procedures and dataset dimension changes

	Object	Actions	Rows	Columns
Initial Dataset			3772	30
Handling Missing Values	Focus on complete and relevant records	Drop TBG column Remove rows with missing data Remove features with only one unique value	2643	23
Encoding	Encode categorical features	Apply one-hot encoding to categorical columns	2643	25
Scaling	Ensure uniform contribution of numerical features	Apply standard scaler to numeric columns	2643	25
Partitioning	Create separate sets for training and unbiased evaluation	Split dataset into training (70%) and testing (30%)	1850 (Train) / 793 (Test)	25

This thorough preparation method made that the data was reliable, organized, and prepared for the creation of precise and effective machine learning models. Figure 2 shows samples from the dataset after processing.



	age	TSH	T3	TT4	T4U	FTI	sex_M	on thyroxine_t	query on thyroxine_t	on antithyroid medication_t	...	query hyperthyroid_t	lithium_t	goitre_t
0	-0.593285	-0.155859	0.605651	0.483484	0.734949	-0.013427	0	0	0	0	...	0	0	0
4	0.830789	-0.180055	-0.972673	-1.321686	-0.639743	-1.214683	0	0	0	0	...	0	0	0
7	1.321849	-0.118312	-1.701131	-0.785776	-1.505291	0.171382	0	0	0	0	...	0	0	0
8	0.634365	-0.185061	0.241423	0.427073	-0.334256	0.695007	0	0	0	0	...	0	0	0
9	0.732577	-0.109968	-0.487035	-0.701159	-0.537914	-0.506250	1	0	0	0	...	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3766	-1.673617	0.157028	0.848471	0.003986	0.582206	-0.383044	0	0	0	0	...	0	0	0
3768	0.732577	-0.168374	0.120013	0.455278	0.429462	0.140581	0	0	0	0	...	0	0	0
3769	1.027213	0.002671	-0.244216	0.116809	0.378548	-0.136632	0	0	0	0	...	1	0	0
3770	0.929001	-0.180889	-0.001396	-0.729364	-0.283341	-0.691058	1	0	0	0	...	0	0	0
3771	0.536153	-0.168374	0.241423	-0.249866	0.378548	-0.537051	0	0	0	0	...	0	0	0

2643 rows × 25 columns

FIGURE 2. - Dataset features after preprocessing

Four metrics were used to evaluate the performance of the models along with a confusion matrix that provides a breakdown of actual versus expected results across categories. When evaluating models on unbalanced datasets such as the set used, metrics of precision, recall, and F1 score are used because relying solely on accuracy (ratio of correct predictions) in these situations can be deceptive because the model can predict the majority class with great accuracy. Precision lessens the impact of false positives, which can result in needless medical procedures, by measuring the percentage of accurately detected positive instances out of all projected positives. Conversely, recall assesses the model's capacity to identify every true positive instance while reducing false negatives, which are crucial for medical diagnosis since they indicate missed disease detections. When the cost of false positives and false negatives varies much, the F1-score—which is the harmonic mean of precision and recall—offers a fair assessment. For instance, a false positive could unnecessarily alarm a patient without a condition, while a false negative could fail to identify a patient who needs urgent treatment. These metrics collectively ensure a nuanced assessment of the model’s reliability in medical decision-making.

3.3 BAGGING MODEL

Thyroid illness categorization can benefit from the application of ML techniques, including DT algorithms, which are becoming more and more common in medical diagnosis. By improving the performance of classifiers that are deemed weak in classification, Bagging Classifier seems to direct the work of classifiers and elevate the level of their training. It does this by training several classifiers and obtaining the best results from them. Figure 3 illustrates the suggested model.

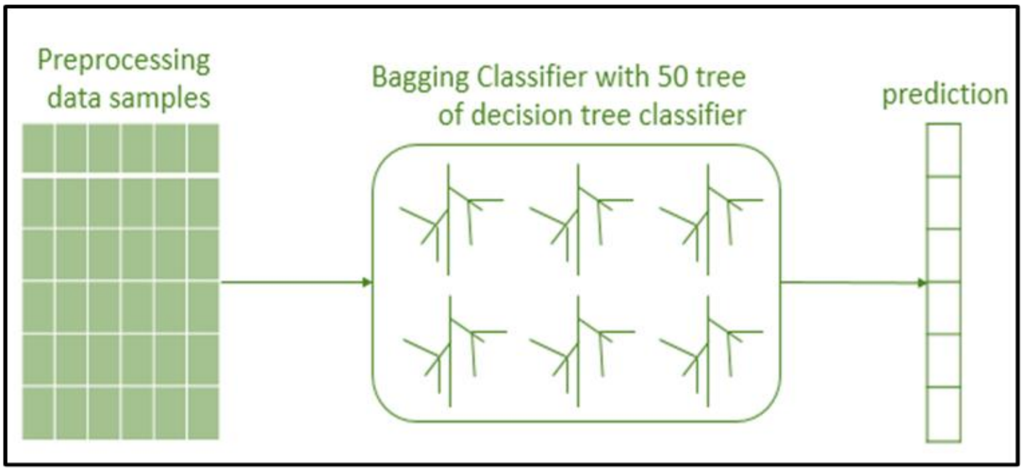
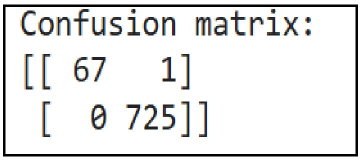


FIGURE 3. - Proposed bagging model

Base Model: A supervised learning approach called a Decision Tree creates a tree-like structure of decisions by recursively dividing the input according to feature values. The leaf nodes show the final classification results, but each internal node reflects a choice based on a characteristic.

Bagging Classifier: Despite their interpretability and visual appeal, decision trees can overfit, particularly when trained on small or noisy datasets. Their ability to generalize to unknown data may be limited by this overfitting. A bagging classifier can be used to solve this. By lowering variance, bagging enhances the model's performance. It accomplishes this by using randomly selected portions of the data, frequently with replacement (bootstrap sampling), to train multiple instances of the base classifier (Decision Tree). Fifty Decision Trees, each trained on a distinct subset of the data, make up the present model. All trees' forecasts are averaged to provide the final prediction (or, in the case of classification, a majority vote). In addition to improving Decision Trees' predictive abilities, this ensemble approach strengthens the model, reducing the possibility of overfitting and raising overall accuracy and stability.

The training data was fitted to the Bagging Classifier, which is made up of 50 decision tree classifiers. In this procedure, bootstrapping is used to train each tree in the ensemble on a randomly selected subset of the training data. The model builds decision boundaries inside each tree to learn how to translate input features to target labels. Multiple tree variety enhances generalization by enabling the model to identify various patterns in the data. To visualize the quantity of true positives, true negatives, false positives, and false negatives, a confusion matrix was computed as shown in Figure 4. The confusion matrix shows only one misclassification in both positive and negative categories, demonstrating the model's robustness and effectiveness.



```
Confusion matrix:
[[ 67   1]
 [  0 725]]
```

FIGURE 4. - Confusion matrix of Bagging Classifier

3.4 SUPPORT VECTOR MACHINE

SVM is a supervised ML technique that is applied to tasks involving regression and classification. Finding the best hyperplane to divide data elements into classes while optimizing the margin between them is how it operates. SVM is known for its strong performance in high-dimensional spaces where it has been widely used to analyze medical data, such as gene expression profiles or imaging datasets, to differentiate between healthy and diseased states. An SVM classifier with an RBF kernel maps data into a higher-dimensional space, enabling it to model non-linear relationships effectively. This configuration generates flexible decision boundaries that can handle complex patterns in the data.

3.5 GAUSSIAN NAIVE BAYES

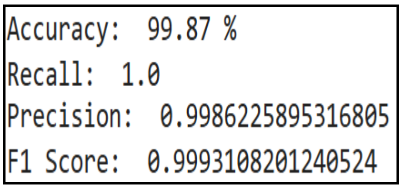
It is a probabilistic classifier that relies on Bayes' theorem and assumes that features are conditionally independent and have a (normal) Gaussian distribution. Given the input features, it determines each class's probability and chooses the one with the highest likelihood. Due to its dependence on probability distribution assumptions, the NB model works well in situations when feature independence is maintained, but it may not work well with linked features that are frequently found in intricate medical data. One of the most important features of this classifier is its computational efficiency, which makes it very suitable in cases of lack of computational resources for treatment or the need for a rapid initial diagnosis.

3.6 ADABOOST

AdaBoost is a boosting-type ensemble learning algorithm that combines multiple weak classifiers to form a strong one. It operates iteratively, assigning higher weights to misclassified samples to focus subsequent classifiers on harder cases, refining overall accuracy. AdaBoost classifier was used to predict thyroid illness using 50 estimators. AdaBoost's iterative method is its main strength. By concentrating on situations that are challenging to categorize, AdaBoost enhances its predictions, which makes it very useful with challenging datasets.

4. RESULTS

Bagging Classifier achieved exceptional performance in predicting thyroid disease, with an accuracy of 99.87%, a recall of 100%, a precision of 99.86% and a f1 score of 99.93%, as shown in Figure 5.



```
Accuracy:  99.87 %
Recall:  1.0
Precision:  0.9986225895316805
F1 Score:  0.9993108201240524
```

FIGURE 5. - **Performance metrics of the Bagging Classifier, demonstrating high accuracy, precision, recall, and F1-score**

Figure 6 presents the accuracy scores of the four ML models used for thyroid disease prediction. The Bagging model achieved the highest accuracy at 99.87%, followed closely by AdaBoost with 98.87%. This visual comparison underscores the superior accuracy of ensemble techniques for predictive tasks in complex datasets.

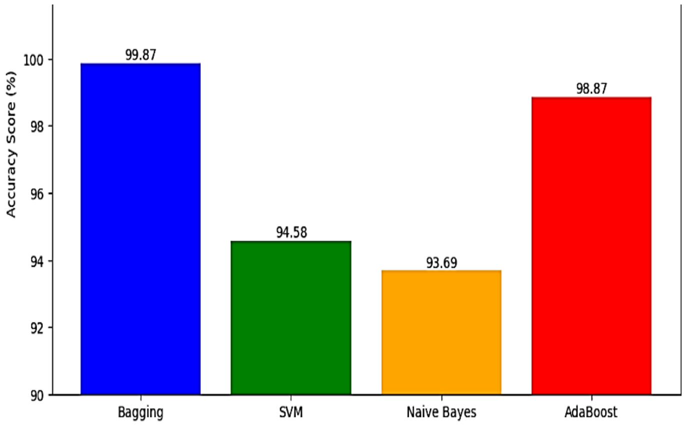


FIGURE 6. - **Accuracy of the four machine learning models**

Figure 7 shows a comparison of precision scores among the four models for thyroid disease prediction. The Bagging classifier achieved the highest precision at 99.86%, followed closely by AdaBoost at 99.18%.

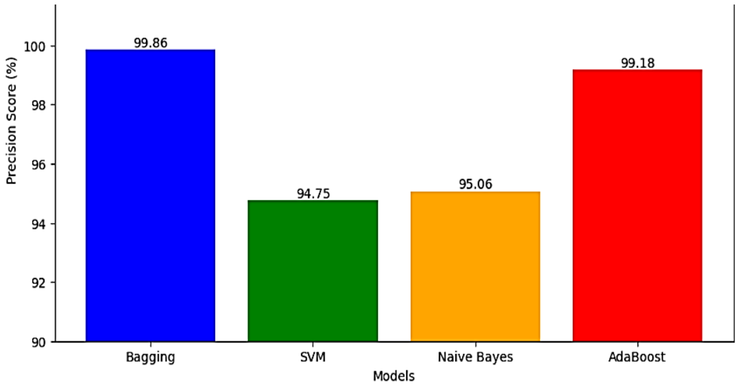


FIGURE 7. - Precision of the four machine learning models

Figure 8 presents a bar chart comparing the recall scores. The models' high recall, especially the 100% recall of the Bagging Classifier, has important ramifications for medical diagnosis. Reducing false negative results is essential when it comes to thyroid illness identification because untreated cases can cause serious health issues. The healthcare priority of early and accurate disease identification is aligned with a high recall, which guarantees that almost all patients with the ailment are correctly identified. This promotes prompt medical intervention and lowers the chance of treatment delays. This performance demonstrates the model's potential usefulness in therapeutic contexts where sensitivity is crucial.

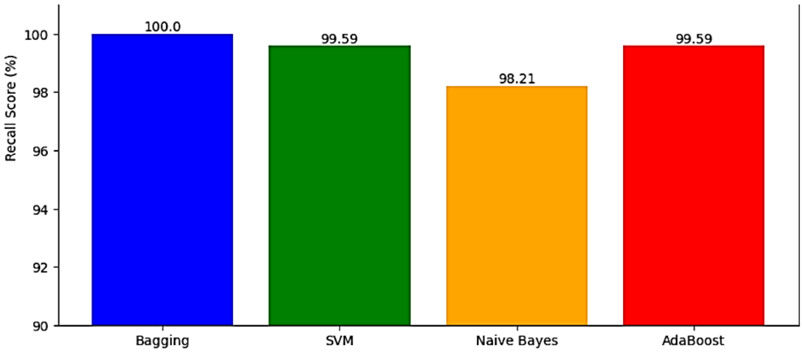


FIGURE 8. - Recall of the four machine learning models

Bagging achieved the highest F1 score at 99.93%, indicating its superior balance of precision and recall, as shown in Figure 9. AdaBoost followed closely with an F1 score of 99.38%, also showing strong performance. These results demonstrate that ensemble methods more reliable choices for accurate classification in this medical context.

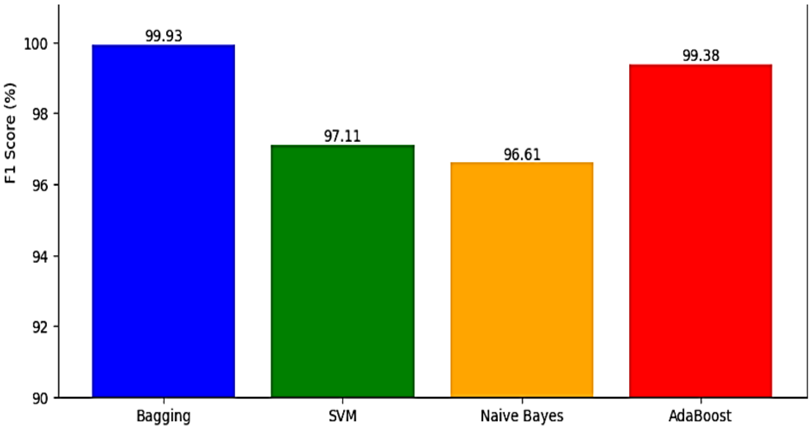


FIGURE 9. - F1 score of the four machine learning models

5. DISCUSSION

Table 3 compares the results of the proposed model with the results of two studies [10] and [12] that used the same dataset. We find that the proposed model outperforms significantly. Bagging outperforms Naive Bayes by combining predictions from multiple decision trees, reducing variance and improving the robustness of overfitting, while Naive Bayes relies on the assumption of feature independence, which may not hold up in complex medical datasets. The packing also exceeded the LSTMBAM model, which suffers from high computational complexity and sensitivity to hyperparameter tuning.

Table 3. - Comparing with related work

Ref.	Used Models	Best Model	Accuracy
T. Faiz (2022) [10]	KNN ANN Random Forest Naive Bayes	Naive Bayes	98%
D. Priyadharsini and S. Sasikala (2022) [12]	LSTMBAM	LSTMBAM	98.94%
Proposed Model	Bagging Classifier SVM AdaBoost Gaussian Naive Bayes	Bagging Classifier	99.87 %

The suggested bagging classifier's excellent performance suggests that it has good generalizability to other datasets, particularly in medical diagnosis. Because it can handle unbalanced data well, it can be used in situations where some groups are underrepresented. The model's adaptability can be demonstrated by applying it to related diagnostic tasks, including identifying diabetes or cardiovascular disorders, with little adjustment. Preprocessing steps, such as handling missing data, one-hot encoding for categorical variables, and scaling for numerical features, contributed significantly to the model's success. The comprehensive feature engineering ensured that the input data was properly prepared for the classifier, maximizing its predictive power. Additionally, the choice to drop columns with constant values and rows with missing values further improved model accuracy and efficiency.

The high computational cost of training numerous decision trees, especially when dealing with huge datasets or when quick predictions are required, is one of the drawbacks of the Bagging Classifier's excellent performance. The quality of the dataset has a significant influence on its efficacy; problems such as missing values or incorrectly scaled features might have an adverse effect on the outcomes. Although thorough preprocessing is crucial, mistakes made at this stage could jeopardize the accuracy of the model. Furthermore, the lower interpretability of ensemble strategies when compared to less complex models may limit their usefulness in situations requiring high transparency. Efficient data management, feature optimization, and striking a balance between computing economy and accuracy are all necessary to meet these challenges.

6. CONCLUSION

This research is driven by the pressing need for early and precise thyroid illness identification. The emergence of machine learning presents a chance to improve diagnostic procedures and give physicians strong instruments to aid in medical judgment. The study used a dataset containing various attributes related to patient demographics, medical history, and thyroid-related test results. Preprocessing included removing rows with missing values and irrelevant columns to maintain data quality. Categorical variables were transformed using rapid single coding, and numeric variables were scaled using StandardScaler. 1850 samples were used to train four classifiers and 793 samples were used for testing. Bagging Classifier with 50 decision tree estimators achieved the highest accuracy of 99.87%, along with high precision (99.86%), recall (100%), and an F1-score of 99.93%. The results indicate that ensemble methods, especially bagging using decision trees, provided the most accurate and reliable results, closely followed by AdaBoost. Both models showed a high degree of overall accuracy, making them suitable for thyroid disease prediction tasks. This research highlights the effectiveness of machine learning in medical diagnosis and indicates that ensemble methods are promising tools for dealing with complex, high-dimensional medical datasets. Future research should focus on incorporating data from real-time patient monitoring to improve prediction accuracy. Continuous data collection from wearable technology allows medical professionals to provide prompt interventions in response to patients' evolving health. Working together with clinics to collect a variety of demographic information will increase the model's applicability to various demographics. Patient feedback can also be used to continually update and improve the model's performance. Finally, engaging in interdisciplinary research involving healthcare providers, data scientists, and ethicists will ensure the ethical application and integration of these technologies into clinical settings.

FUNDING

None

ACKNOWLEDGEMENT

Thanks to everyone who helped whole this research and furnished recommendation and help.

CONFLICTS OF INTEREST

The authors declare no conflict of interest

REFERENCES

[1] A R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," CA: a cancer journal for clinicians, vol. 68, no. 1, pp. 7–30, 2018.

[2] T. Faiz, "Empirical method for thyroid disease classification using a machine learning approach," BioMed Research International, vol. 2022, pp. 1–10, 2022.

[3] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid disease prediction using machine learning approaches," National Academy Science Letters, vol. 44, no. 3, pp. 233–238, 2021.

[4] L. Aversano, M. L. Bernardi, M. Cimitile, M. Iammarino, P. E. Macchia, I. C. Nettore, and C. Verdone, "Thyroid disease treatment prediction with machine learning approaches," Procedia Computer Science, vol. 192, pp. 1031–1040, 2021.

[5] E. Sonuç, "Thyroid disease classification using machine learning algorithms," in Journal of Physics: Conference Series, vol. 1963, no. 1, p. 012140, July 2021.

[6] M. Hu, C. Asami, H. Iwakura, Y. Nakajima, R. Sema, T. Kikuchi, T. Miyata, K. Sakamaki, T. Kudo, M. Yamada, T. Akamizu and Y. Sakakibara, "Development and preliminary validation of a machine learning system for thyroid dysfunction diagnosis based on routine laboratory tests," Communications Medicine, vol. 2, no. 1, p. 9, 2022.

[7] K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, and V. H. C. De Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification," The Journal of Supercomputing, vol. 76, pp. 1128–1143, 2020.

[8] Alshayeji, M. H. (2023). Early thyroid risk prediction by data mining and ensemble classifiers. Machine Learning and Knowledge Extraction, 5(3), 1195-1213.

[9] S. Sankar, A. Potti, G. N. Chandrika, and S. Ramasubbareddy, "Thyroid disease prediction using XGBoost algorithms," J. Mob. Multimed., vol. 18, no. 3, pp. 1–18, 2022.

[10] T. Faiz, "Empirical method for thyroid disease classification using a machine learning approach," BioMed Research International, vol. 2022, pp. 1–10, 2022.

[11] M. Hosseinzadeh, O. H. Ahmed, M. Y. Ghafour, F. Safara, H. K. Hama, S. Ali, ... and H. S. Chiang, "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, pp. 3616–3637, 2021.

[12] D. Priyadharsini and S. Sasikala, "Novel Hybrid LSTBAM-Bidirectional Associative Memory Deep Learning Based Thyroid Disease Prediction," Journal of Pharmaceutical Negative Results, vol. 13, no. 10, pp. 1179–1185, 2022.

[13] Y. Wang, W. Yue, X. Li, S. Liu, L. Guo, H. Xu, ... and G. Yang, "Comparison study of radiomics and deep learning-based methods for thyroid nodules classification using ultrasound images," IEEE Access, vol. 8, pp. 52010–52017, 2020.

[14] I. D. Mienye and Y. Sun, "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," IEEE Access, vol. 10, pp. 99129–99149, 2022.

[15] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble learning for disease prediction: A review," in Healthcare, vol. 11, no. 12, p. 1808, MDPI, June 2023.

[16] [16] Y. Hessein, "Thyroid Disease Data Set," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/yasserhessein/thyroid-disease-data-set/version/1?select=hypothyroid.csv. [Accessed: Dec. 8, 2024].