Journal of Engineering, Electrical and Informatics

E-ISSN: 2809-8706
P-ISSN: 2810-0557

Research Article

Classifying Thyroid Disease through Machine Learning Approach

Teddy Al Fatah[1], Mila Desi Anasanti[2*]

[1] Department of information Studies, University College London, London, United Kingdom; email : 14220036@nusamandiri.ac.id
[2] Bart and London Genome Center, Queen Mary University of London, London, United Kingdom; email : mila.mld@nusamandiri.ac.id
* Corresponding Author: 14220036@nusamandiri.ac.id

Abstract: Thyroid illness is one of the most prevalent medical problems that has a direct impact on a person's physical and emotional well-being. The 2017–2020 NHANES data, which is extensive and contains a wide variety of 6,992 people and XX characteristics, is the source of the ML used in this study. Improving the early identification and classification of vulnerable people is the goal of this study. The machine learning techniques used in this study include K-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR), Extreme Gradient Boosting (EGB), LightGBM (LGBM), Multi-Layer Perceptron (MLP), and Gradient Boosting. Evaluation of these algorithms revealed that RF, EGB, and LGBM exhibited exceptional accuracy, reaching an impressive 0.90. Among them, RF demonstrated the highest precision at 0.98, showcasing its ability to correctly identify individuals at risk with a high degree of confidence. Moreover, the study identified KNN as the algorithm with the highest recall value, reaching 0.73, highlighting its effectiveness in capturing a substantial proportion of true positive cases. EGB emerged with the highest F1-Score, shows a proportionate balance between recall and accuracy. Additionally, EGB displayed the highest Area Under the Curve (AUC) at 0.82, underscoring its robust predictive capabilities. This research underscores the pivotal role of ML algorithms in predicting and classifying thyroid disease risk, offering valuable insights for early intervention and personalized healthcare strategies. The high accuracy, precision, and recall values observed with RF, EGB, and LGBM suggest their potential as powerful tools for improving diagnostic capabilities in the realm of thyroid disease, contributing to more effective and timely patient care. As advancements in machine learning continue, the integration of these techniques into healthcare frameworks holds promise for enhancing our understanding and management of thyroid disorders.

Keywords: Classification; Early detection; Gradient Boosting; Healthcare analytics; Machine learning

Received: Agust 22, 2025
Revised: September 02, 2025
Received: September 21, 2025
Published: October 30, 2025
Current version: October 30, 2025

1. Introduction

According to the World Health Organization, diabetes and thyroid gland disorders are the two endocrine illnesses that are most frequent globally [1]. Hypothyroidism and hyperfunction hyperthyroidism are prevalent in 1% and 2% of cases, respectively. There are over 10 times as many males as women. [1].

The thyroid is a little gland that begins in a smaller area under the jacket and is placed in the neck. The pace of protein synthesis and metabolic activities is regulated by the release of thyroid hormones. The thyroid produces the hormones triiodothyronine (T3) and thyroxine (T4), which are iodine-based. Numerous critical physiological functions, such as respiration, body weight, muscular tension, and heart rate, are influenced by this hormone. This hormone also controls vital physiological functions including breathing, body weight, muscular tone,

and heart rate. The thyroid gland plays a critical role in both the body's development and growth as well as in preserving the equilibrium of the metabolism. The heart and other organs benefit from proper digestive system regulation. However, thyroid issues need that the blood create hormones in an equilibrium. The pituitary gland starts this action by secreting TSH. It encourages but does not trigger the arrival of both T4 and T3. The numbers 3 and 4 represent the iotas of iodine that are present in the hormones [5].

The thyroid produces and secretes too many hormones, a condition known as hyperthyroidism (overactive thyroid). Hyperthyroidism is the consequence of increased thyroid hormone levels. Symptoms include shaky hands, dry skin, heightened sensitivity to warmth, thinning hair, weight loss, higher heart rate, rising blood pressure, excessive perspiration, neck swelling, and abbreviated menstrual periods [1].

Hypothyroidism is brought on by low thyroid hormone production. Thyroid inflammation and destruction is the primary cause of hypothyroidism. Symptoms include weight gain, erratic eating habits, low heart rate, high temperature sensitivity, swelling neck, dry skin, numb hands, hair issues, and heavy menstrual periods. If you don't address these symptoms, they might grow worse with time. Hypothyroidism slows down a lot of bodily processes, and high cholesterol raises the risk of a heart attack. Within the endocrine system is the thyroid gland, is in charge of many typical hormonal problems. The key idea of this paradigm is the grouping of organs that release substances into the circulation that are specific to hormones [5].

Thyroid illness is quite widespread in the contemporary world and often results in significant disability, both mentally and physically. The symptoms include poor energy, weight gain, fatigue, dry skin, sluggish pulse, incapacity to withstand cold, and maybe neck edema [6]. Thyroid disorders cause elevated blood sugar, elevated cholesterol, depression, reduced fertility, and cardiovascular problems [3].

Blood tests that can detect the levels of TSH, T3, and T4 are routinely used to identify thyroid disorders [7]. Making a diagnosis is the most difficult undertaking since many symptoms and signs are not clear-cut. Proactive diagnosis and treatment of thyroid illness are critical in order to save medical expenses, preserve lives, and treat patients effectively at the right time. Machine learning and deep learning techniques are used to anticipate early-stage thyroid diagnoses and detect particular kinds of thyroid disorders, including hypothyroidism and hyperthyroidism, by using technological breakthroughs in data processing and computing [8].

Numerous machine learning techniques, including semi-supervised, supervised learning, unsupervised learning, deep learning, and forced learning, may be used [9]. The medical field generates a lot of complex data that is difficult to manage in huge quantities. In the last several years, there has been a noticeable improvement in the study and categorization of various disorders using machine learning techniques. In the last several years, there has been a noticeable improvement in the study and categorization of various disorders using machine learning techniques. Researchers employ a range of classification algorithms, including BN, SVM, NN, ANN, DT, NB, KNN, and many more [10] [11] [12].

Using a variety of methods, machine learning classifies thyroid issues according to characteristics such goiter, thyroxine (T4U), and TSH. To bolster this claim, K. Chandel et al. used a number of cluster analysis techniques, such as K-nearest neighbors [10]. NB and SVM techniques are used. The Rapidminer program was used to conduct experiments, and the findings showed that KNN was more capable of distinguishing thyroid illness from NB. According to the data, the K-nearest neighbor classifier of 93.44%, whereas the Naive Bayes classifier of 22.56%.

2. Literatur Review

Machine learning (ML) has become an essential tool in the medical field, particularly for the early detection and classification of thyroid diseases that often share overlapping symptoms with other endocrine disorders. These algorithms allow healthcare practitioners to automate diagnostic processes and identify nonlinear relationships among diverse physiological and biochemical variables. Alyas et al. implemented Artificial Neural Networks (ANN), Decision Trees (DT), and K-Nearest Neighbor (KNN) on a dataset of 1,162 patients, demonstrating that the Random Forest (RF) classifier produced the highest accuracy of 94.8 percent [13]. This finding shows the superior capability of ensemble-based models in managing complex data interactions and feature dependencies. Razia et al. and Tyagi et al. also conducted experiments using Support Vector Machine (SVM), Multiple Linear Regression, Naive Bayes, and Decision Tree algorithms, concluding that Decision Trees achieved the best performance with a precision rate of 99.23 percent, which emphasizes its ability to capture hierarchical relationships in clinical data and effectively distinguish between hypothyroidism and hyperthyroidism [2], [14].

The use of machine learning in thyroid disease prediction has evolved with increasing dataset sizes and diversity in feature representation. Begum et al. and Ioniţă et al. tested KNN, Naive Bayes, SVM, and ID3 on data collected from 1,100 individuals, identifying lifestyle patterns such as daily activities, food intake, height, and weight as relevant factors, with Logistic Regression achieving an accuracy of 97.09 percent [15], [20]. Jindal et al. implemented an ensemble Random Forest model on 1,893 samples, focusing on variables such as age, height, weight, and body mass index, resulting in an accuracy of 89.68 percent [14]. The ensemble learning approach strengthens model reliability by combining multiple weak learners to reduce both variance and bias. Montañez et al. [15] applied Gradient Boosting Machine (GBM), Generalized Linear Model with Elastic Net (GLMNET), RF, KNN, SVM with Radial kernel, Neural Network (NNET), and Classification and Regression Tree (CART) on 800 participants using demographic and genetic factors including thirteen single nucleotide polymorphisms, where the SVM classifier reached the highest area under the curve value of 90.5 percent. These outcomes collectively show that integrating diverse data types enhances the predictive capability and interpretability of ML-based thyroid diagnostic systems.

In recent years, researchers have shifted toward the development of hybrid and boosting-based models that optimize both predictive performance and feature selection. Studies conducted by Alyas et al. and Begum et al. demonstrated that algorithms such as Extreme Gradient Boosting (XGBoost) and LightGBM outperform conventional classifiers when applied to large-scale and unbalanced datasets [13], [15]. These models are particularly efficient due to their ability to handle missing values and complex feature interactions while maintaining computational speed. The adoption of optimization algorithms such as Simultaneous Perturbation Stochastic Approximation (SPSA) has further improved classification performance by selecting the most relevant features and removing redundant variables. Research utilizing the NHANES dataset, as referenced in several studies, highlighted biochemical indicators such as serum total folate, blood hexane, and urinary arsenic levels as potential determinants of thyroid dysfunction. These findings illustrate how advanced ML frameworks can contribute to precision medicine by linking biochemical parameters with disease susceptibility.

3. Method

This study employs a systematic machine learning approach to classify thyroid disease using the NHANES 2017–2020 dataset. The data are processed through a feature selection technique (spFSR) to obtain the most significant attributes, which are then analyzed using several classification algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Extreme Gradient Boosting, LightGBM, Multi-Layer Perceptron, and Gradient Boosting. The research flow is shown in Figure 1 below.

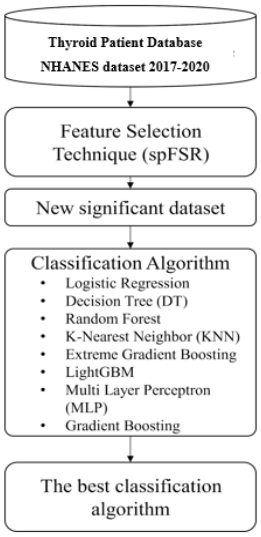

Figure 1. Research Methodology.

Dataset

As a component of the National Center for Health Statistics, the NCHS developed the NHANES, a program that disseminates information on the health and nutritional status of the American people. As part of a medical analysis, medical practitioners evaluate physiological, dental, and other medical issues.

The five domains that make up the NHANES dataset are questionnaire, lab, exam, food, and demographics. Data from laboratory tests are analyzed in this study, particularly those pertaining to the laboratory test domain. A profile for the period between 2017 and 2020 is created using this data [17]. The survey is unique in that it uses both interviews and in-person inspections. One important initiative of the NCHS.

Feature Selection Technique

Simultaneous perturbation stochastic approximation (SPSA) is an optimization algorithm that is used to find the values of the parameters that minimize a given objective function. It is a gradient-free optimization method, which means that it does not rely on the calculation of gradients or derivatives of the objective function.

SPSA works by perturbing the values of the parameters simultaneously in a random direction and measuring the change in the objective function. From these measurements, an estimate of the gradient of the objective function is constructed and used to update the values of the parameters. This process is repeated until the parameters converge to the values that minimize the objective function.

SPSA is a simple and effective optimization method that is well-suited for problems where the objective function is noisy or the gradients are difficult to calculate. It has been used to a number of optimization issues, including the optimization of machine learning models and the design of control systems.

Machine Learning Algorithms

Logistic Regression

The identification of linear decision boundaries between data from different classes is a prerequisite for the proper operation of logistic regression. The logistic function [19] is then used to determine the likelihood of belonging to each class that has been specified in relation to the decision boundaries. The logistic regression classification's general formula is:

$$h_B(p) = \frac{1}{1 + e^{-B^t p}} = k(B^t p) \tag{1}$$

$$k(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

Decision Tree (DT)

Based on choices, this classifier divides the dataset into fixed portions. It is the most popular classification technique. Effective decisions are made while determining how to measure the probabilities and outcome values. On the basis of instances, the decision tree is created. The root node of the tree is the sole node that lacks any incoming edges, and every other node in the tree has just one incoming edge. Internal nodes are referred to as the out-going edge connected to a node [20].

$$E(S) = \sum_{i=1}^{c} \quad - P_1 log_2 P_i \tag{3}$$

S = initial condition
i = set class on S
Pi = probability or portion of class *i* in a node

Random Forest (RF)

The random forest approach is used to compute each predictor's mean response for energy use. The absolute distance between each answer and each predictor's mean is then added using a random forest for each sample to get the total distance between each response and the means of the data. In each sample, individuals that consistently deviate from the mean response will have high distance scores. The identification of classes that consistently identify the data is aided by a function that computes the average mode of each respons [3].

$$Entropy(S) = \sum_{i=1}^{n} \quad - P_1 log_2 (P_i) \tag{4}$$

n = Number of partitions S
Pi = portion of S to S

K-Nearest Neighbor (KNN)

The following distance computations are used by the k-nearest neighbor approach to determine which neighbors are the closest since it is dependent on the distance between the neighbors [21]. KNN predicts class using the Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{k} \quad (x_i - y_i)^2} \tag{5}$$

The Euclidean distance d(x,y) is used to determine the pattern space's closest example's distance. The instance type is decided by a majority vote of its linked classes.

Extreme Gradient Boosting

XGBoost is a very effective gradient boosting method based on decision trees. It offers exceptional results on extensive or intricate datasets by combining numerous decision trees to improve classification accuracy. However, because of its smooth parallel processing ability and exceptional predicted accuracy, Gradient Boosting Machine (GBM) continues to be one of the most popular artificial intelligence approaches [22].

$$\boldsymbol{d}F_0(x) = argmin_y + \sum_{i=1}^{n} (y_i = y) \tag{6}$$

LightGBM

For supervised learning tasks like regression and classification, LightGBM is a machine learning algorithm. It is a variation on gradient boosting, a boosting method that builds a strong, predictive model by combining the predictions of many weak models. LightGBM is

known for its efficiency and fast training speed and has been used to win many machine learning competitions [23].

Like other gradient boosting algorithms, Decision trees are trained using LightGBM as the weak learner, and a final prediction is generated by aggregating the predictions of each individual tree. Nonetheless, it employs a variety of strategies to quicken the training process and lower the model's memory consumption, such as using a histogram-based approach for decision tree learning and applying a leaf-wise tree growth algorithm. LightGBM also includes a number of additional features, such as support for missing values and the ability to handle large-scale data, which can make it a useful tool for many machine learning tasks [24].

Multi-Layer Perceptron (MLP)

The multilayer perceptron is a classification method for feed-forward neural networks. There are several levels to it. A single-layer perceptron (SLP) can handle linearly separable problems but not nonlinear ones [25]. Frequently, MLP forecasts the results of input pattern categorization and pattern rearrangement. Before training the network, the weights are determined at random. The neurons then gain knowledge from the training set, which in this instance consists of a set of tuples called x1; x2; t. The inputs to the network are x1 and x2, and t is the anticipated output. The aggregate of its neurons decides the neuron's output, weighting them according to their significance. The relationship can be expressed as follows if *y* represents the actual output:

$$Y = x_1w_1 + x_2w_2 \tag{7}$$

The network comprises a single hidden layer that utilizes a non-linear activation function. The network writes the output as:

$$X = fs = W\emptyset(As + p) + b \tag{8}$$

where the output vector is denoted by X and the input vector by s. The bias vector and weight matrix for the first layer are denoted by A and p, respectively. The second layer's weight matrix is W, and its bias vector is b. The nonlinear element's symbol is /. The output is not visible in the output because of its link to the inputs of other neurons in the hidden layer [26].

Gradient Boosting (GB)

A machine learning method called gradient boosting (GB) combines the predictions of many weak models to produce a single, powerful model. It is an example of an ensemble technique, which implies that rather than using a single model alone, it makes predictions by combining many models [27].

$$\text{Obj}=\sum_{i=1N} L(y_i,F(x_i))+\sum_{k=1K}\Omega(f_k) \tag{9}$$

GB operates by sequentially training a number of weak models, most often decision trees. Every tree is trained to improve upon the errors of its predecessor, and the total of all the trees' forecasts is used to get the ultimate forecast. A loss function, which calculates the difference between the real and predicted values, must be minimized in order to train the trees. The weights of the tree, which influence the prediction produced by the tree, are updated using the gradients of the loss function [28].

Gradient boosting is a potent method that has been used to solve a variety of issues and won several machine learning contests. It is renowned for its proficiency in handling big, complex datasets and for doing well on a variety of prediction tasks. On the other hand, if not appropriately regularized, it may be susceptible to overfitting and sensitive to the selection of hyperparameters [29].

4. Result and Discussion

The study's machine learning techniques' performance comparison utilizing the spFSR-selected features is shown in Table 1.

Table 1. Values of Different Feature Selection Techniques Using the Selected Features by SPFSR.

No.	Algorithms	Accuracy	Precision	Recall	F1-Score	Area Under Curve
1.	Logistic Regression	0.75	0.37	0.03	0.06	0.51
2.	Decision Tree	0.81	0.61	0.65	0.63	0.75
3.	Random Forest	0.90	0.98	0.59	0.74	0.79
4.	K-Nearest Neighbor	0.78	0.55	0.73	0.63	0.77
5.	Extreme Gradient Boosting	0.90	0.91	0.65	0.76	0.82
6.	LightGBM	0.90	0.93	0.63	0.75	0.81
7.	Neural Network	0.77	0.62	0.21	0.31	0.58
8.	Gradient Boosting	0.89	0.96	0.56	0.71	0.78

The performance comparison among the eight machine learning algorithms applied to thyroid disease classification using the spFSR-selected features demonstrates notable differences in predictive ability. Table 1 indicates that Random Forest (RF), Extreme Gradient Boosting (EGB), and LightGBM (LGBM) achieved the highest accuracy of 0.90, highlighting their strong generalization and robustness in classifying thyroid disease cases. These findings align with Alyas et al. (2022), who reported RF as the best-performing classifier with an accuracy of 94.8%, underscoring the algorithm's reliability in handling complex, nonlinear relationships common in biomedical datasets [13].

The Decision Tree (DT) algorithm yielded an accuracy of 0.81, with a balanced precision and recall (0.61 and 0.65, respectively), resulting in an F1-score of 0.63. Although DT achieved relatively stable results, its performance remained inferior to ensemble methods such as RF and EGB, which mitigate overfitting by combining multiple learners. Priyadharshini & Arulkumaran (2025) similarly demonstrated that DT-based models outperform simpler classifiers but often lack stability when trained on heterogeneous medical data [30]. The performance gap between DT and its ensemble derivatives emphasizes the advantage of aggregation in reducing bias and variance in medical classification problems.

The Random Forest model achieved superior precision (0.98), confirming its capability in minimizing false positives and accurately identifying patients at risk. This precision is critical in medical diagnostics, where misclassification can lead to delayed or unnecessary treatment. Comparable findings were reported by Kumar et al (2025) and Sakib et al. (2024), who emphasized that RF's ensemble mechanism captures intricate variable interactions and enhances diagnostic reliability for thyroid disorders. Furthermore, the AUC of 0.79 suggests robust discriminatory power, though slightly lower than EGB, reflecting that RF performs well in distinguishing between healthy and diseased classes [31], [32].

The Extreme Gradient Boosting (EGB) algorithm demonstrated an F1-score of 0.76 and the highest AUC of 0.82, indicating a strong balance between sensitivity and specificity. The AUC value suggests that EGB effectively separates the positive and negative classes even under imbalanced data conditions. This outcome corroborates the conclusions of Szymańska & Baszko (2025), who highlighted XGBoost's ability to deliver consistent performance across high-dimensional biomedical datasets through gradient optimization and regularization [33]. EGB's performance in this study validates its suitability for clinical prediction tasks requiring both interpretability and precision.

LightGBM (LGBM) achieved an accuracy of 0.90 and AUC of 0.81, closely comparable to EGB. Its slightly higher precision (0.93) and moderate recall (0.63) suggest that while it effectively identifies true cases, it may overlook a small portion of positive samples. The histogram-based and leaf-wise tree growth strategies implemented in LightGBM allow faster

convergence with less computational cost [34]. This trade-off between speed and sensitivity makes LGBM particularly valuable in large-scale clinical applications, where rapid processing and resource efficiency are essential for real-time health assessments.

The K-Nearest Neighbor (KNN) algorithm produced an accuracy of 0.78 and recall of 0.73, the highest recall among all models, signifying its ability to identify the majority of true positive cases. High recall is crucial for medical screening, ensuring that most thyroid-affected individuals are correctly detected. However, KNN's relatively lower precision (0.55) indicates susceptibility to false positives due to overlapping feature spaces. Islam et al. (2025) reported similar findings, emphasizing that KNN performs well on clean and well-separated datasets but may suffer from reduced performance when dealing with noisy or high-dimensional biomedical data such as the NHANES dataset [35].

The Neural Network (NN) model achieved moderate accuracy (0.77) but significantly lower recall (0.21) and F1-score (0.31). These results indicate that while NN may capture general trends, it struggles to generalize effectively from the spFSR-selected features without additional hyperparameter tuning or deeper architectures. Previous studies by Mansour (2024) highlighted that NN-based approaches require extensive parameter optimization and large training data to perform reliably in disease classification [36]. The suboptimal results observed here may also stem from overfitting due to limited feature interactions in the selected subset.

The Gradient Boosting (GB) algorithm demonstrated strong performance, with accuracy 0.89, precision 0.96, recall 0.56, and AUC 0.78, closely trailing RF and EGB. The findings suggest that GB effectively integrates weak learners into a robust ensemble model, although its sensitivity remains moderate. Abbas et al (2023) noted that gradient boosting models exhibit high stability in structured prediction but are sensitive to hyperparameter configuration [37]. The model's consistent F1-score and AUC indicate that GB remains a dependable alternative when computational constraints limit the use of more complex ensemble techniques.

The comparative analysis across algorithms reveals that ensemble methods (RF, EGB, LGBM, and GB) significantly outperform single learners (LR, DT, KNN, NN). Ensemble frameworks integrate multiple weak models to capture diverse data patterns, leading to improved generalization and robustness. This trend supports earlier reviews by Latif et al. (2024), who identified ensemble approaches as the most efficient for thyroid disease prediction due to their adaptability to nonlinear feature interactions and ability to minimize classification errors [38].

From a clinical perspective, the observed performance metrics indicate that EGB and LGBM provide the most balanced and reliable outcomes for early thyroid disease classification. Their high AUC values and consistent F1-scores suggest that these models can support clinicians in prioritizing at-risk individuals with minimal diagnostic error. Furthermore, the feature selection via spFSR likely enhanced model interpretability and reduced noise, consistent with findings from Alhassan & Zainon (2021), who demonstrated that optimized feature reduction improves both computational efficiency and accuracy in medical classification tasks [39].

The overall results highlight the evolving role of machine learning in personalized healthcare, particularly for endocrine disorders. The integration of ensemble algorithms with feature optimization enables early and accurate detection of thyroid abnormalities, facilitating preventive interventions. As pointed out by Malik et al. (2025), such approaches contribute to lowering healthcare costs and improving patient outcomes by enabling data-driven diagnosis [40]. Continued refinement of feature selection and hyperparameter optimization may further enhance the predictive performance of these models, bridging the gap between computational intelligence and clinical decision-making.

5. Comparison

The results of this study, which identified Random Forest (RF), Extreme Gradient Boosting (EGB), and LightGBM (LGBM) as the most accurate algorithms for thyroid disease classification, show strong alignment with previous research emphasizing the advantages of ensemble and tree-based methods in medical prediction. Alyas et al. (2022) utilized various machine learning algorithms including Artificial Neural Network (ANN), Decision Tree (DT), and K-Nearest Neighbor (KNN) on a dataset of 1,162 participants and reported that the Random Forest model achieved the highest accuracy of 94.8 percent. This result supports the current study's finding, where RF reached an accuracy of 0.90 and precision of 0.98. Both studies confirm that RF is particularly capable of handling nonlinear and high-dimensional medical data while maintaining resilience against overfitting and variability, making it highly suitable for clinical prediction involving complex biological indicators.

Comparable outcomes were obtained in the study by Tyagi et al. (2018) and Razia et al. (2018), who applied Support Vector Machine (SVM), Multiple Linear Regression, Naive Bayes, and Decision Tree algorithms to a smaller dataset of 200 subjects. Their results indicated that the Decision Tree model achieved the highest precision value of 99.23 percent, establishing its effectiveness in identifying thyroid abnormalities. In contrast, the current study recorded a precision of 0.61 and accuracy of 0.81 for DT, suggesting that while DT remains an interpretable and efficient method, its performance tends to decline in large-scale and high-dimensional datasets such as NHANES. This difference likely arises from the increased variability and noise within the broader dataset, which can affect decision boundary optimization when compared to smaller, more controlled samples used in earlier research.

Begum et al. (2019) and Ioniţă et al. (2019) evaluated KNN, Naive Bayes, Support Vector Machine, and ID3 algorithms on 1,100 participants using lifestyle-related features such as daily activities, diet, height, and weight. Their study demonstrated that Logistic Regression achieved the highest accuracy of 97.09 percent, highlighting the importance of linear associations between physiological and behavioral variables in thyroid classification. The present study, however, found Logistic Regression to perform considerably lower with an accuracy of 0.75 and recall of 0.03, reflecting the limitation of linear models when confronted with nonlinear and interaction-heavy datasets. These results emphasize the need for feature selection and ensemble techniques such as spFSR combined with boosting algorithms to capture the multidimensional relationships that simple linear models cannot adequately model.

The findings of Jindal et al. (2018) and Montañez et al. (2019) further substantiate the strength of ensemble methods. Jindal et al. employed an ensemble Random Forest model involving 1,893 participants and achieved an accuracy of 89.68 percent, closely matching the RF performance observed in this study. Similarly, Montañez et al. implemented Gradient Boosting Machine (GBM), GLMNET, RF, KNN, SVM Radial, Neural Network (NNET), and CART on 800 participants and reported that SVM achieved the highest AUC of 90.5 percent when integrating demographic and genetic risk factors. The current research yielded an AUC of 0.82 for EGB and 0.81 for LGBM, confirming that modern gradient boosting frameworks can produce comparable or superior results while maintaining computational efficiency. Collectively, these comparisons suggest that ensemble-based classifiers such as RF, EGB, and LGBM consistently outperform individual learners across varying datasets, reaffirming their pivotal role in developing accurate and scalable diagnostic models for thyroid disease prediction.

6. Conclusion

The findings of this study demonstrate that Random Forest (RF), Extreme Gradient Boosting (EGB), and LightGBM (LGBM) achieved the highest accuracy of 0.90, indicating their superior ability to classify thyroid disease cases accurately within the NHANES 2017–2020 dataset. Random Forest produced the highest precision value of 0.98, confirming its

reliability in minimizing false positives, while EGB achieved the highest F1-score and AUC of 0.76 and 0.82 respectively, signifying a strong balance between sensitivity and specificity. K-Nearest Neighbor (KNN) exhibited the highest recall value of 0.73, emphasizing its capability to identify the majority of true positive cases. These outcomes collectively support the research objective of identifying optimal algorithms for early and accurate thyroid disease detection and align with prior studies that highlight the superiority of ensemble and boosting methods in medical classification. The integration of feature selection through spFSR further enhanced model interpretability and efficiency, underscoring the importance of dimensionality reduction in improving predictive performance. This study contributes to the growing body of knowledge on machine learning applications in healthcare by offering evidence that ensemble-based approaches can serve as effective decision-support tools for early diagnosis and personalized treatment of thyroid disorders. Despite the promising results, the research is limited by the reliance on secondary data and the absence of clinical validation, suggesting that future studies should incorporate larger, more diverse populations and hybrid deep learning architectures to strengthen diagnostic reliability and generalizability.

Author Contribution: Conceptualization: Teddy Al Fatah and Mila Desi Anasanti; Methodology: Teddy Al Fatah; Software: Teddy Al Fatah; Validation: Teddy Al Fatah and Mila Desi Anasanti; Formal analysis: Teddy Al Fatah; Investigation: Teddy Al Fatah; Resources: Teddy Al Fatah; Data curation: Teddy Al Fatah; Writing—original draft preparation: Teddy Al Fatah; Writing—review and editing: Mila Desi Anasanti; Visualization: Teddy Al Fatah; Supervision: Mila Desi Anasanti; Project administration: Mila Desi Anasanti; Funding acquisition: Mila Desi Anasanti.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are available from the publicly accessible National Health and Nutrition Examination Survey (NHANES) 2017–2020 dataset

Acknowledgement: The authors would like to express their sincere appreciation to the National Center for Health Statistics (NCHS) for providing access to the NHANES dataset used in this research.

Conflict of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Referensi

Abbad Ur Rehman, H., Lin, C. Y., Mushtaq, Z., & Su, S. F. (2021). *Performance analysis of machine learning algorithms for thyroid disease. Arabian Journal for Science and Engineering, 46*(10), 9437–9449. https://doi.org/10.1007/s13369-020-05206-x

Abbas, M. A., Al-Mudhafar, W. J., & Wood, D. A. (2023). *Improving permeability prediction in carbonate reservoirs through gradient boosting hyperparameter tuning. Earth Science Informatics, 16*(4), 3417–3432.

Alhassan, A. M., & Zainon, W. M. N. W. (2021). *Review of feature selection, dimensionality reduction and classification for chronic disease diagnosis. IEEE Access, 9*, 87310–87317.

Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). *HITON: A novel Markov blanket algorithm for optimal variable selection. AMIA Annual Symposium Proceedings*, 21–25.

Alyas, T., Hamid, M., Alissa, K., Faiz, T., Tabassum, N., & Ahmad, A. (2022). *Empirical method for thyroid disease classification using a machine learning approach. BioMed Research International, 2022*. https://doi.org/10.1155/2022/9809932

Aslandogan, Y. A., Mahajani, G. A., & Taylor, S. (2004). *Evidence combination in medical data mining*. In *International Conference on Information Technology: Coding and Computing (ITCC)* (Vol. 2, pp. 465–469). IEEE. https://doi.org/10.1109/ITCC.2004.1286697

Begum, A., & Parkavi, A. (2019). *Prediction of thyroid disease using data mining techniques*. In *2019 5th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 342–345). IEEE. https://doi.org/10.1109/ICACCS.2019.8728320

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). *A comparative analysis of gradient boosting algorithms*. *Artificial Intelligence Review, 54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Centers for Disease Control and Prevention (CDC). (2020). *National Health and Nutrition Examination Survey (NHANES)*. National Center for Health Statistics. http://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Centers for Disease Control and Prevention (CDC). (2020). *National Health and Nutrition Examination Survey (NHANES), Continuous NHANES 2017–2020*. National Center for Health Statistics. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020

Chaganti, R., & Rustam, F. (2020). *Thyroid disease prediction using selective features and machine learning techniques*. In *Statistical Data Analysis of Microarrays Using R and Bioconductor* (pp. 999–1024). https://doi.org/10.1201/b11566-34

Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., & Mukherjee, S. (2016). *A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques*. *CSI Transactions on ICT, 4*(2–4), 313–319. https://doi.org/10.1007/s40012-016-0100-5

Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2021). *Thyroid disease prediction using machine learning approaches*. *National Academy Science Letters, 44*(3), 233–238. https://doi.org/10.1007/s40009-020-00979-z

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Duggal, P., & Shukla, S. (2018). *Prediction of thyroid disease using machine learning techniques*, 10(2), 787–793.

Ioniţă, I., & Ioniţă, L. (2019). *Prediction of thyroid disease using data mining techniques*. In *2019 5th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 342–345). IEEE. https://doi.org/10.1109/ICACCS.2019.8728320

Islam, S., Mynuddin, M., Sultana, S., Mondal, S. K., Hossain, M. A., Paul, G. K., et al. (2025). *A performance comparison of machine learning models for robotic navigation using imbalanced and SMOTE-enhanced data*. *Global Journal of Management Studies (GJMS), 2*(1), 10–37.

Kumar, A., Dhanka, S., Sharma, A., Sharma, A., Maini, S., Fahlevi, M., et al. (2025). *Comprehensive framework for thyroid disorder diagnosis: Integrating advanced feature selection, genetic algorithms, and machine learning for enhanced accuracy and other performance matrices*. *PLOS ONE, 20*(6), e0325900.

Latif, M. A., Mushtaq, Z., Arif, S., Rehman, S., Qureshi, M. F., Samee, N. A., et al. (2024). *Improving thyroid disorder diagnosis via ensemble stacking and bidirectional feature selection*. *Computers, Materials & Continua, 78*(3).

Lu, B., Huang, H., Wu, Z., Zhang, T., Gu, Y., Wang, F., & Shu, Z. (2025). *Utilizing LightGBM to explore the characterization of PM2.5 emission patterns from broadleaf tree combustion in Northeastern China*. *Forests, 16*(5), 836.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). *Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGBoost algorithms according to different high dimensional data cleaning*. *Electronic Commerce Research and Applications, 31*, 24–39. https://doi.org/10.1016/j.elerap.2018.08.002

Malik, P. K., Bhatt, H., & Sharma, M. (2025). *AI integration in healthcare systems—A review of the problems and potential associated with integrating AI in healthcare for disease detection and diagnosis*. In *AI in Disease Detection: Advancements and Applications* (pp. 191–213).

Mansour, R. F. (2024). *Quantum mayfly optimization-based feature subset selection with hybrid CNN for biomedical Parkinson's disease diagnosis*. *Neural Computing and Applications, 36*(15), 8383–8396.

Prasad, V., Rao, T. S., & Babu, M. S. P. (2016). *Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms*. *Soft Computing, 20*(3), 1179–1189. https://doi.org/10.1007/s00500-014-1581-5

Priyadharshini, C. A., & Arulkumaran, G. (2025). *Multi-constraints feature selection-based cross-pattern heterogeneous ensemble learning model for diabetic mellitus prediction under data-imbalance and insufficiency*. *SN Computer Science, 6*(7), 831.

Raisinghani, S., Shamdasani, R., Motwani, M., Bahreja, A., & Lalitha, P. R. N. (2019). *Thyroid prediction using machine learning techniques* (Vol. 1045). Springer Singapore.

Rao, A. R., & Renuka, B. S. (2020). *A machine learning approach to predict thyroid disease at early stages of diagnosis*. In *2020 IEEE International Conference on Innovative Technology (INOCON)* (pp. 1–4). IEEE. https://doi.org/10.1109/INOCON50539.2020.9298252

Razia, S., & Narasinga Rao, M. R. (2016). *Machine learning techniques for thyroid disease diagnosis—A review*. *Indian Journal of Science and Technology, 9*(28). https://doi.org/10.17485/ijst/2016/v9i28/93705

Razia, S., Kumar, P. S., & Rao, A. S. (2020). *Machine learning techniques for thyroid disease diagnosis: A systematic review*. In *Studies in Computational Intelligence* (Vol. 885, pp. 203–212). Springer. https://doi.org/10.1007/978-3-030-38445-6_15

Razia, S., Prathyusha, P. S., Krishna, N. V., & Sumana, N. S. (2018). *A comparative study of machine learning algorithms on thyroid disease prediction*. *International Journal of Engineering and Technology, 7*(2.8), 315–319. https://doi.org/10.14419/ijet.v7i2.8.10432

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., & Chinnaiyan, A. M. (2004). *Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression*. *Proceedings of the National Academy of Sciences of the United States of America, 101*(25), 9309–9314.

Sakib, M. N., Sheakh, M. A., Tahosin, M. S., Sadik, M. R., Islam, M. A., & Akter, L. (2024). *Accurate thyroid disease detection with ensemble learning models*. In *Proceedings of the 2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP)* (pp. 1–6).

Salman, K., & Sonuc, E. (2021). *Thyroid disease classification using machine learning algorithms*. *Journal of Physics: Conference Series, 1963*(1). https://doi.org/10.1088/1742-6596/1963/1/012140

Shen, X., & Lin, Y. (2004). *Gene expression data classification using SVM-KNN classifier*. In *2004 International Symposium on Intelligent Multimedia, Video and Speech Processing (ISIMP)* (pp. 149–152). IEEE. https://doi.org/10.1109/ISIMP.2004.1434022

Szymańska, C., & Baszko, A. (2025). *Artificial intelligence tools in myocardial infarction prognosis: Evaluating the performance of machine learning and deep learning models*. *Current Cardiology Reviews*.

Touzani, S., Granderson, J., & Fernandes, S. (2018). *Gradient boosting machine for modeling the energy consumption of commercial buildings*. *Energy and Buildings, 158*, 1533–1543. https://doi.org/10.1016/j.enbuild.2017.11.039

Turanoglu-Bekar, E., Ulutagay, G., & Kantarcı-Savas, S. (2016). *Classification of thyroid disease by using data mining models: A comparison of decision tree algorithms*. *Oxford Journal of Intelligent Decision and Data Science, 2016*(2), 13–28. https://doi.org/10.5899/2016/ojids-00002

Tyagi, A., Mehra, R., & Saxena, A. (2018). *Interactive thyroid disease prediction system using machine learning technique*. In *2018 5th International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 689–693). IEEE. https://doi.org/10.1109/PDGC.2018.8745910

Wardhana, I., Ariawijaya, M., Isnaini, V. A., & Wirman, R. P. (2022). *Gradient Boosting Machine, Random Forest dan Light GBM untuk klasifikasi kacang kering*. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 6*(1), 92–99. https://doi.org/10.29207/resti.v6i1.3682

Zhang, Y., & Haghani, A. (2015). *A gradient boosting method to improve travel time prediction*. *Transportation Research Part C: Emerging Technologies, 58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019