machine learning & knowledge extraction

MDPI

Article

Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers

Mohammad H. Alshayeji

Computer Engineering Department, College of Engineering and Petroleum, Kuwait University, Safat, P.O. Box 5969, Kuwait City 13060, Kuwait; m.alshayeji@ku.edu.kw

Abstract: Thyroid disease is among the most prevalent endocrinopathies worldwide. As the thyroid gland controls human metabolism, thyroid illness is a matter of concern for human health. To save time and reduce error rates, an automatic, reliable, and accurate thyroid identification machine-learning (ML) system is essential. The proposed model aims to address existing work limitations such as the lack of detailed feature analysis, visualization, improvement in prediction accuracy, and reliability. Here, a public thyroid illness dataset containing 29 clinical features from the University of California, Irvine ML repository was used. The clinical features helped us to build an ML model that can predict thyroid illness by analyzing early symptoms and replacing the manual analysis of these attributes. Feature analysis and visualization facilitate an understanding of the role of features in thyroid prediction tasks. In addition, the overfitting problem was eliminated by 5-fold cross-validation and data balancing using the synthetic minority oversampling technique (SMOTE). Ensemble learning ensures prediction model reliability owing to the involvement of multiple classifiers in the prediction decisions. The proposed model achieved 99.5% accuracy, 99.39% sensitivity, and 99.59% specificity with the boosting method which is applicable to real-time computer-aided diagnosis (CAD) systems to ease diagnosis and promote early treatment.

Keywords: machine learning; thyroid; data mining; ensemble model; feature engineering

Citation: Alshayeji, M.H. Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1195–1213. https://doi.org/10.3390/make5030061

Academic Editor: Nada Lavrač

Received: 26 July 2023
Revised: 2 September 2023
Accepted: 13 September 2023
Published: 18 September 2023

1. Introduction

Thyroid hormones are primarily responsible for the regulation of human metabolism, neuronal growth, and large increases in reproductive activity. When the thyroid gland is unable to produce regular hormone levels, normal body functionality is disrupted. This is known as thyroid disease. Thyroid disorders in medical science can result in thyroiditis and thyroid cancer. The primary thyroid conditions are hyperthyroidism and hypothyroidism [1]. The proportion of people with thyroid dysfunction is rising globally, accounting for between 30 and 40% of patients seen in endocrine clinics [2]. An estimated 20 million Americans have thyroid illness, of which approximately 60% are unaware [3].

It might be challenging to distinguish thyroid disease from other illnesses owing to the conditions it produces. Diagnosis of thyroid illness is a difficult and time-consuming process. The traditional procedure for diagnosing thyroid illness involves clinical examination, as well as many blood tests. However, the main issue is being able to make a precise diagnosis of the condition in its early stages [4]. Blood tests are recommended if doctors suspect thyroid dysfunction because they can offer crucial details regarding various thyroid hormones, including thyroid-stimulating hormone (TSH), triiodothyronine (T3), thyroxin (T4), and thyroid-stimulating immunoglobulin (TSI) [5].

Thyroid disease should never be underestimated, as it can lead to severe complications such as thyroid storm (extreme hyperthyroidism) and myxedema (life-threatening end-stage untreated hypothyroidism) [6]. Therefore, managing the progression of an illness and even avoiding death greatly depend on early disease identification, diagnosis, and therapy. Despite numerous tests being conducted, medical diagnosis is sometimes regarded

as a difficult task [7]. Data mining uses a semi-automated process to identify patterns and relationships in large databases [8]. The classification of thyroid dysfunction can be resolved using data-mining methods. The use of machine-learning (ML) and DM techniques has become more widespread among academics and researchers.

It has become increasingly clear in recent years that the complexity of life and changes in food preferences have contributed to the sharp increase in medical issues. In addition, the cost of medical therapy is thought to be higher, particularly for compliance that requires surgical intervention. Data science, technology, and intelligent systems can be used to enhance medical diagnosis [9]. The strategic prediction of thyroid disease is crucial for providing appropriate care, reducing medical expenses, and preventing avoidable fatalities. By understanding each patient's complete medical history over time and analyzing the patterns of biological indicators and other parameters, it is possible to predict individual responses to therapy and modify treatments accordingly. Thus, it is essential to enhance medical procedures and leverage cutting-edge technologies for the early identification and prevention of thyroid problems. With advancements in data processing and computation technologies, ML approaches have been increasingly employed for accurate thyroid diagnosis.

A general outline of the workflow that combines data mining and ML is shown in Figure 1. The dataset chosen for the model training consisted of 29 clinical features. Because the data are gathered in real time, data preparation must be performed to obtain appropriate ML classifier inputs. Data preprocessing involves data encoding, cleaning, resampling, and normalization. To address any imbalances in the dataset and mitigate bias, a synthetic minority oversampling technique (SMOTE) algorithm can be employed to balance the data, which generates minority class synthetic samples by using K-nearest neighbors. In addition, to ensure that the models can be generalized to new data and are not unduly specialized in the training data, overfitting can be checked using cross-validation. Among the various thyroid diseases that include thyroiditis, goiter, thyroid cancer, etc., the present study focused on hypothyroid diagnosis. However, the same framework can also be employed for other diseases by training the model with the appropriate database. The primary contributions of this study are as follows.

Figure 1. Workflow outline.

1. The development of a simple, automatic, precise, and reliable ML thyroid prediction model can be incorporated into computer-aided diagnosis (CAD) systems.
2. Correlation analysis, heatmap generation, and other visualizations of various clinical features were implemented to understand their role in thyroid risk prediction.
3. Involvement of 29 features that can predict and classify the thyroid by analyzing early symptoms. In addition, it replaces the tedious manual analyses of these parameters. The SMOTE algorithm was implemented to achieve data balancing and to ensure that the results were not biased.
4. Ensemble learning ensures the reliability of the prediction model owing to the use of several classifiers, instead of a single ML algorithm.

A detailed survey of various existing studies that address thyroid prediction is presented in Section 2. Section 3 provides additional information on the materials and techniques used in this study. Section 4 provides a detailed methodology. The corresponding experiments and results are discussed in Sections 5 and 6 provides the conclusions of the work.

2. Related Works

A thorough analysis of the different prediction methodologies presented in recent studies for diagnosing thyroid illness is provided in this section by categorizing them as conventional ML- and DL-based approaches.

Traditional ML approaches: The authors created a thyroid prediction model [10] that uses a kernel-based classification approach with "multi-kernel SVM". To enhance the effectiveness of the classification process, optimal feature selection was performed using enhanced gray wolf optimization (GWO) [11], which is a population-based meta-heuristic technique that mimics the natural leadership structure and hunting strategy of gray wolves. With respect to GWO, the three fittest candidate solutions—alpha, beta, and delta—lead the population to favorable regions of the search space. By adaptively searching for the feature space, we determined the optimal feature combination. This approach achieved 97.49% accuracy, 99.05% sensitivity, and 94.5% specificity. However, this technique requires lengthy computations; hence, the authors suggest the need for new and powerful procedures that will improve the performance and allow for the diagnosis of thyroid disease.

In [12], the authors followed an empirical method by comparing the performance of random forest (RF), artificial neural networks (ANN), decision tree (DT), and K-nearest neighbor (KNN) on the dataset to enhance disease prediction based on the dataset's specified parameters. The dataset was also altered to allow for a precise classification prediction. The dataset was modified to enable an accurate classification prediction, and categorization was performed on the sampled and unsampled datasets to improve comparability. After modifying the dataset, the RF algorithm achieved an accuracy of 94.8% and a specificity of 91%.

Using ML techniques, such as gradient-boosting machine (GBM), deep neural network (DNN), logistic regression (LR), SVM, and RF, Garcia et al. [13] identified molecules that are highly likely to initiate thyroid hormone homeostasis. In the initial phases of thyroid illness, the early prediction of these compounds is helpful for additional testing. The ToxCast database provides information on molecular events. The best predictive performance was observed for the thyroid hormone receptor (TR) and thyroid peroxidase (TPO), with F1 values of 0.81 and 0.83.

Another study [14] investigated three feature selection algorithms in conjunction with ML algorithms: univariate feature selection (UFS), principal component analysis (PCA), and recursive feature extraction (RFE). PCA [15] transforms high-dimensional data into lower dimensions, whereas UFS [16] selects the strongest features and RFE [17] removes the weakest features until a specified number of features are reached. PCA is used when dealing with high-dimensional data, UFS is preferred for a simple selection approach, and RFE is performed by iteratively selecting the strongest features. The RFE and ML classifiers performed better than the other classifiers and achieved an accuracy of 99.35%. However, the sample size was small (n = 519). A sizable dataset was required to assess the efficacy of the method.

In another study [18], in-depth analyses of thyroid prediction were conducted using various ML classifiers, with and without feature selection techniques. The DHQ Teaching Hospital in Dera Ghazi Khan, Pakistan, provided the data. It also included the extra parameters of blood pressure, body mass index, and pulse rate, which stood out from previous research. The experiment was conducted with and without L1 and L2 norm feature selection [19]. L1 regularization aids feature selection by eliminating irrelevant features, whereas L2 reduces model overfitting by minimizing weight magnitudes. L1 is better when a subset of features is important, whereas L2 maintains a larger set of features. The final model with the naïve Bayes (NB) classifier and L2 selection achieved 100% accuracy. Another study [20] experimented with different separate distance functions of the KNN classifier for the same database. In addition, they utilized the L1 norm and chi-squared test to perform feature selection. The chi-squared test [21] identifies features strongly associated with the target variable. This is useful for dealing with categorical or discrete data. In addition to improving model performance, it prevents overfitting

and enhances interpretability. Chi-squared techniques with KNN, Euclidean, and cosine distance functions achieved 100% accuracy. However, the proposed model has not yet been validated by using a public thyroid database. Hence, it cannot be verified based on overfitting and real-time usage.

To predict hypothyroid illnesses, Mishra et al. [22] used ML techniques, such as sequential minimum optimization (SMO), RF, DT, and K-star classifiers. SVM training involves solving large quadratic programming optimization problems. SMO breaks them into smaller analytically solved problems, reducing the time-consuming optimization [23]. In this investigation, 3772 distinct records were considered. RF and DT performed better and achieved 99.44% and 98.97% accuracy, respectively. However, hyperthyroid prediction was not considered in the present study. The efficacy of supervised and unsupervised classifiers in predicting thyroid conditions has received considerable attention. Instead of following this pattern, another study [24] concentrated on the idea of feature relevance and its clinical application. They listed the top four features most likely to indicate the presence of thyroid disease, and demonstrated how easily and affordably practitioners could test these features. Moreover, they pointed out the drawbacks of widespread clinical practice in many nations with universal healthcare of not testing the whole thyroid panel. Finally, the results are stable and unlikely to vary depending on the classifier used or because of the fundamental characteristics of the dataset, such as imbalance.

DL-based approaches: Some studies make use of deep-learning-driven algorithms to provide good performance for the automatic identification of thyroid disorders, which provides doctors with support for diagnostic decision-making and reduces human false-positive diagnostic rates. One study [25] used two preoperative medical imaging modalities to classify thyroid diseases. It developed a diagnostic model for thyroid disease based on a convolutional neural network (CNN) architecture, and it achieved remarkable performances of 0.972 and 0.942 accuracy for computed tomography (CT) scans and ultrasound images, respectively, using Xception. They conducted experiments using DeseNet121 and InceptionV3 in addition to Xception.

To improve the prediction accuracy, the authors in [26] followed a series of procedures for input data before being input into the DL networks. After preprocessing the real data, they experimented with feature-reduction techniques, such as singular value decomposition (SVD) and PCA. SVD provides singular vectors with reduced dimensionality. The resulting data were then fed into KNN and neural network classifiers. In addition, the same data were applied to the DNN after the data augmentation. The experiment utilized 2 hidden layers (16 neurons), 1 input layer (23 inputs), and the ReLU activation function for prediction. They achieved an accuracy of 99.95% when using a DNN. Although it achieved a higher prediction accuracy, it could not identify the features based on which the network made the prediction decisions. Additionally, we could not determine whether model overfitting had occurred or not. Another study [27] proposed a collection of multiple multilayer perceptron (MMLP) neural networks with backpropagation error handling to improve the generalization and prevent overfitting during training. They also employed an adjustable learning rate strategy to address the convergence and local minima issues associated with backpropagation error. The MMLP model outperformed a single network, increasing the accuracy by 0.7%. Additionally, using an adjustable learning rate algorithm improved the accuracy of Internet of Medical Things (IoMT) systems by 4.6%, ultimately achieving 99% accuracy compared with classical backpropagation.

To identify five distinct thyroid conditions, hypothyroidism, hyperthyroidism, and thyroid cancer, a CNN-based modified ResNet architecture was used in another study [28]. In the proposed study, the training method was improved by employing dual optimizers to achieve greater accuracy and outcomes. It was found that the modified ResNet operational efficiency increased when using Adam and stochastic gradient descent (SGD) optimizers. The improved ResNet design offers 97% accuracy with SGD compared with 94% accuracy for the baseline ResNet architecture. Table 1 presents an overview of the studies reviewed.

Table 1. Related works.

Work	Methodology	Input Variable	Output	Performance Measures
[10]	Gray wolf optimization + multi-kernel SVM	29 clinicopathological characteristics	Hypothyroid, hyperthyroid, and normal	Accuracy 97.49%, sensitivity 99.05%, and specificity 94.5%
[12]	Customized Alexnet	29 clinicopathological characteristics	Thyroid, normal	Accuracy 94.8% and specificity 91%.
[13]	ML classifiers	Molecular descriptors	Thyroid peroxidase (TPO) active, inactive	F1-score 0.83
[14]	RFE + ML classifier	ID, age, sex, FT3, FT4, T3, T4, and TSH	Hypothyroid, non-hypothyroid	Accuracy 99.35%.
[18]	L2 selection + NB classifier	ID, gender, age, body mass index, pregnant, pulse rate, blood pressure, T3, T4, and TSH	Hypothyroid, hyperthyroid, and normal	Accuracy 100%
[20]	Chi-square test + KNN	ID, gender, age, body mass index, pregnant, pulse rate, blood pressure, T3, T4, and TSH	Hypothyroid, hyperthyroid, and normal	Accuracy 100%
[22]	SMO + RF classifier	29 clinicopathological characteristics	Hypothyroid, non-hypothyroid	Accuracy 99.44%
[24]	Work on the concept of thyroid prediction feature importance and its clinical implications			
[25]	CNN	CT image, ultrasound image	Normal, thyroiditis, cystic, multi-nodular goiter, adenoma, and cancer	Accuracy 97.2%
[26]	DNN	23 clinicopathological characteristics	Sick-euthyroid, negative	Accuracy 99.95%
[27]	MMLP	21 clinicopathological characteristics	Hypothyroid, hyperthyroid, and normal	Accuracy 99%
[28]	ResNet	X-ray images	Thyroid nodules, hypothyroid, hyperthyroid, thyroid cancer, and thyroiditis	Accuracy 97%

In the reviewed existing works based on conventional approaches, the prediction accuracy requires improvement because, while considering ML in the medical field, the prediction accuracy measure plays a significant role. Additionally, the models were built on a single ML classifier and there was no evidence to recheck their reliability. Only after ensuring reliability can the proposed models be incorporated into CAD systems. Although a few studies have achieved a high accuracy, they were developed on private datasets instead of public datasets and obtained permission to acquire data, privacy, secrecy concerns, etc. The remaining studies that are based on DL networks face the limitations of the requirement of huge training data, high computational power, large training time, and black-box nature. Additionally, the underlying decision-making features are unknown. To resolve these issues, This study proposed a model with sufficient reliability by implementing an ensemble of classifiers, in which more than one classifier is involved in making diagnosis decisions. The classic ML algorithms employed ensured model simplicity and speed. In addition, the features can be accessed and analyzed at any stage. Training the model with public data will help it handle diverse data, and thereby, it can be incorporated into real-time CAD systems.

3. Materials and Methods

The database and the existing techniques used in the present study will be presented here.

3.1. Database

In this study, we used the thyroid dataset sourced from the University of California, Irvine (UCI) ML repository [29]. It includes 6 databases from the Garavan Institute in

Sydney, Australia, with 2800 training and 972 test instances. This was used for binary classification. In the database, the predictive variable was given as categorical where "P" represented the presence of hypothyroidism and the opposite with "N". The dataset includes various attributes that represent the different clinicopathological characteristics of patients. In total, 29 attributes that are either categorical or real were included. The attributes are age, sex, thyroxine (whether currently taking thyroxine medication), query on thyroxine (whether a patient is currently under investigation), sick, thyroid_surgery, antithyroid medications, query_hyperthyroid, I131_treatment, pregnancy, query_hypothyroid, TSH_measured (whether the TSH level was measured for patients), TSH (actual value of TSH), T3, referral_source, TT4_measured, T4U, T3_measured, TBG, TT4, FTI_measured, TBG_measured, T4U_measured, FTI, lithium, tumor, goiter, psych, and hypopituitory were considered from their corresponding levels in the blood. Hypothyroidism was used as the target variable.

3.2. SMOTE

Real-world datasets face data-imbalance problems. The data-imbalance issue with the thyroid database was solved using the synthetic minority oversampling technique (SMOTE) algorithm [30,31]. To address the imbalanced dataset problem, one common strategy is to oversample the minority class. However, simply replicating existing examples from a minority class may not provide new insights into the model. Instead, SMOTE generates new synthetic examples by combining the information from existing examples. The SMOTE technique selects examples that are close to each other in the feature space. It constructs a line linking the chosen example and its nearest neighbors (typically five neighbors) and then generates a new sample along this line. This synthetic example fills the gap between two instances and their randomly selected neighbors, thereby effectively increasing the representation of the minority class. By utilizing SMOTE, we mitigated the problem of overfitting that can arise from random oversampling. This technique ensures that synthetic examples are generated in a controlled manner based on the distribution of existing minority class instances, thus improving the generalization capabilities of the model.

3.3. Bayesian Optimization

To search for a global optimization issue efficiently and effectively, Bayesian optimization [32] offers a systematic method based on Bayes' theorem. Contrary to random or grid search, Bayesian techniques retain notes of previous assessment outcomes. This methodology involves creating a surrogate function $((score \mid hyperparameters)\ or\ P(y \mid x))$ that serves as a probabilistic model of the objective function. The surrogate function is then effectively explored using an acquisition function to select candidate samples to evaluate the actual objective function. In the context of applied machine learning, Bayesian optimization is commonly used for the hyperparameter tuning of well-performing models on a validation dataset.

Compared to the objective function, the surrogate function is easier to optimize, and Bayesian approaches operate by selecting hyperparameters that perform best on the surrogate function to determine the new hyperparameter set to assess the actual objective function. Following each objective function evaluation, these methods update the surrogate probability model, which is consistent with the Bayesian reasoning goal, i.e., to become "less wrong" with more data. With fewer iterations, Bayesian approaches can uncover better model parameters than random searches.

3.4. Ensemble Learning

While dealing with class imbalance, different classifiers would produce different results, whereas other classifiers might not be able to perform better in terms of classification. Consequently, an ensemble-learning technique [33] was used to enhance the ML results. To solve this problem, many models have been developed and combined. The core assertion is

that more accurate and resilient models can be produced by appropriately merging the weak models. Compared to a single model, this strategy delivers a better prediction performance.

The two methods of ensemble learning implemented were boosting and bagging. Bagging (bootstrap aggregating) takes homogeneous weak learners, trains them independently in parallel, and combines them using a deterministic averaging process. It uses Breiman's random forest algorithm with decision tree learners. Numerous boosting algorithms have been proposed. The first versions, proposed by Robert Schapire [34] and Yoav Freund [35], lacked adaptability and were unable to fully capitalize on weaker learners. Schapire and Freund [36] later created AdaBoost, an adaptive boosting algorithm that earned the coveted Gödel Prize. The first truly successful boosting algorithm created for binary classification was called AdaBoost. Boosting frequently considers homogeneous weak learners, trains them sequentially in a highly adaptive manner, and then combines them in accordance with a deterministic strategy. It follows AdaBoost with decision tree learners. The bias was reduced by boosting.

4. Proposed Methodology

A detailed workflow diagram of the present study is shown in Figure 2. Early identification, diagnosis, and treatment are critical to stop the progression of thyroid illnesses and lower the mortality rate. The accurate prediction of disease outcomes and understanding of the interdependencies of clinical features play a crucial role in medical diagnosis and therapy. By developing a fully automatic ML model that can be incorporated into a CAD system, a higher prediction time and error rate can be eliminated.

The public thyroid database from the UCI ML Repository [29] was used in this study. With 29 clinical variables that will help in early diagnosis, this is the only public thyroid dataset with such a large sample size. By using a public dataset for model development, the proposed model will be capable of handling real-world data limitations.

Preprocessing the dataset is a crucial step that both data mining and ML rely on. Real-world data have inherent irregularities and noise, and there is a chance that some of the data will be missing, duplicated, or irrelevant. This can result in false information being learned and a decline in algorithm performance. The dataset is encoded, resampled, normalized, and thereby put into the proper format using preprocessing to make it ready for processing. In the database, data values are also written in words as well as by numbers. The training data are frequently labeled in words to make them human-readable and intelligible. However, machines require numeric representations, and hence, label encoding transforms actual data into numeric. Machine-learning algorithms can then analyze and determine the optimal functioning of these labels. Converting categorical data into numeric values is essential in supervised learning for structured datasets, ensuring accurate analysis.

To verify that the data given into the ML model are not redundant, duplicate rows were then eliminated as part of the data-cleaning process. A given class is typically underrepresented in relation to other classes in datasets that accurately reflect the real world. When learning a concept from a class with few examples, it might be difficult due to the "class-imbalance" problem. One among the main issues associated with data mining and pattern recognition is the data-imbalance problem. Unbalanced datasets significantly impede learning because the bulk of currently used ML techniques presuppose a balanced class distribution or an equal penalty for misclassification. To solve this problem associated with our database, data resampling was performed with the SMOTE technique, which uses oversampling. Patients without hypothyroidism made up the minority group in this case. Because SMOTE was applied ahead of data splitting, the minority class had sufficient samples and the data imbalance was mitigated.

SMOTE has several benefits over conventional oversampling. Duplicating existing data may bias the model in conventional oversampling. SMOTE interpolates between the minority class's existing samples to construct synthetic samples. By adding diversity to the oversampled dataset, the likelihood of overfitting is decreased. SMOTE can assist in enhancing the model's capacity to discriminate between classes in the critical regions by

producing samples along the decision boundary. Additionally, it can assist in decreasing the effect of noisy data points.

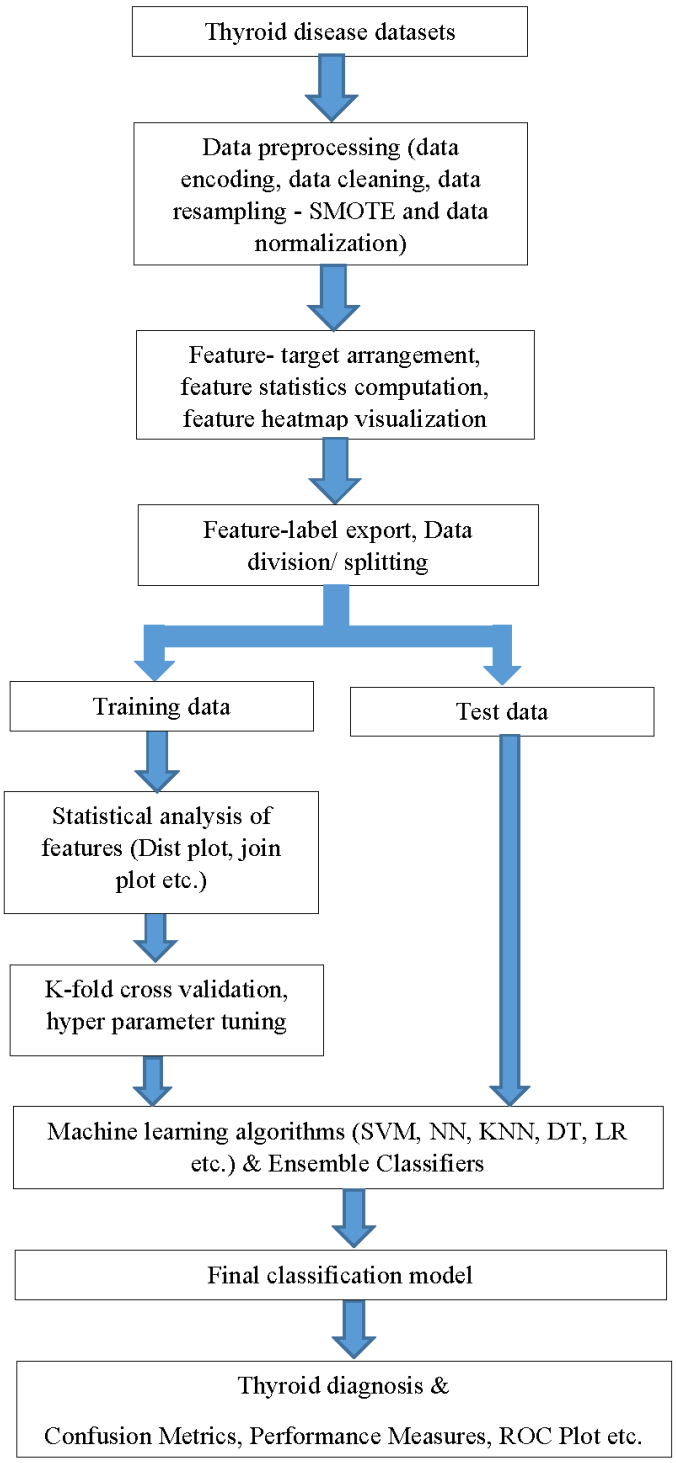

Figure 2. Complete flow diagram.

This study utilized SMOTE class implementation from the imbalanced-learn Python library. This class is like a scikit-learn data transform object and needs to be defined and configured before it can be applied to a dataset. By using SMOTE, it initially oversampled the minority class, ensuring it contained 10% majority class instances. Required ratios were passed as arguments to the SMOTE and RandomUnderSampler classes, which were

combined into a pipeline. This pipeline allowed us to apply the transformations step by step, resulting in a final dataset with the desired number of transformations.

As the final part of preprocessing, data normalization is applied to convert the numerical values in the dataset's columns to a comparable scale while keeping their ranges. This method is noteworthy for maintaining the consistency of such value ranges. By using a linear data transformation called min–max normalization, the range of the dataset is condensed to a single range, set to 0 to 1. Here, it was applied to the TT4, T4U, FTI, and age data columns. Then, the statistical summary, such as the mean, count, standard deviation, percentiles, and minimum–maximum value ranges of the preprocessed data, was obtained.

To analyze the data, feature distributions and seaborn distribution charts, such as distplot and jointplot, were utilized. First, we selected one specific column of the dataset since the distplot frequently visualizes univariant sets of observations using histogram, that is, one observation. Then, a plot of two variables with bivariate and univariate graphs is created using a jointplot. In essence, it mixes two distinct narratives. Finally, the correlation between the data columns is evaluated, and a corresponding heatmap is generated. Data preprocessing was completed, and the data features and targets were separated from the preprocessed data. Feature-target values were exported into Excel files for later ML stages.

In the ML classification stage, DT, neural network (NN), SVM, LR, and KNN were initially used. Then, ensemble classifiers were employed since they would be more reliable than individual classifiers. A 5-fold cross-validation which prevents overfitting was also performed. Moreover, hyperparameters were tuned by employing Bayesian optimization. The ensemble-learning approaches of bagging and boosting were employed. Using bagging, several models were trained on a portion of the real dataset before combining the model results to produce the final prediction conclusion. Another meta-model was also trained combining the output predictions of many models to produce a concluding forecast. Boosting-based ensemble techniques contain models that have been trained several times, relying on previously trained performance errors as well as underperforming models. Afterward, a weighted average of the predictions was created using the accuracy of various models. Real-world scenarios benefit most from these ML models that have been trained to handle a class imbalance that characterizes all real-world data.

Some disease diagnosis may be affected by intricate relationships between features that are not immediately apparent. The model can capture these complicated relationships when all clinical features are used. To evaluate feature relevance, feature selection sometimes necessitates time-consuming data-preprocessing activities. The entire data-preparation procedure can be streamlined by omitting this step. Models using all features might provide more transparency in their decision-making process, as clinicians and researchers can observe the influence of various features on the predictions. Models without feature selection are better able to adapt to data changes, making them advantageous for real-time applications where new features become relevant in the future.

When the proposed approach is implemented into medical systems, thyroid disease screening can be carried out with minimal or no assistance from physicians. As a result, the standard thyroid screening approach could be changed, allowing for faster screening and treatment. Patients from remote locations and the elderly who do not have easy access to hospitals might benefit from the integration of this ML model into mobile applications. In principle, by retraining the suggested framework, it can be used for many diseases such as thyroid dysfunction, diabetes mellitus, pituitary tumors, acromegaly, etc. It, however, requires similar biomedical databases with clinical features that can aid in the prediction of a specific disease.

5. Experimental Results and Discussion

The results of each experiment conducted during the model development will be presented and discussed here. Several preprocessing approaches were used to improve the model performance that include locating and handling missing values and encoding categorical data. Furthermore, various ML classifier algorithms were tried along with

cross-validation and hyperparameter tuning. The accuracy, sensitivity, specificity, etc., were assessed to gauge the efficacy of the prepared model.

As a first step, a thyroid dataset from the UCI ML repository was used since we needed a real-time database with enough data for model training. A model developed from such a database would be useful as a CAD system in day-to-day hospital activities. The amount of data are essential for improved accuracy. When conducting research using private healthcare data, researchers must also address additional challenges such as obtaining permission to acquire the data, privacy, and secrecy concerns, etc. These are some of the reasons why many researchers use this public thyroid disease database. Since the dataset consisted of data-imbalance problems, the SMOTE algorithm was implemented in the preprocessing stage, and bar charts of both cases are given in Figure 3.

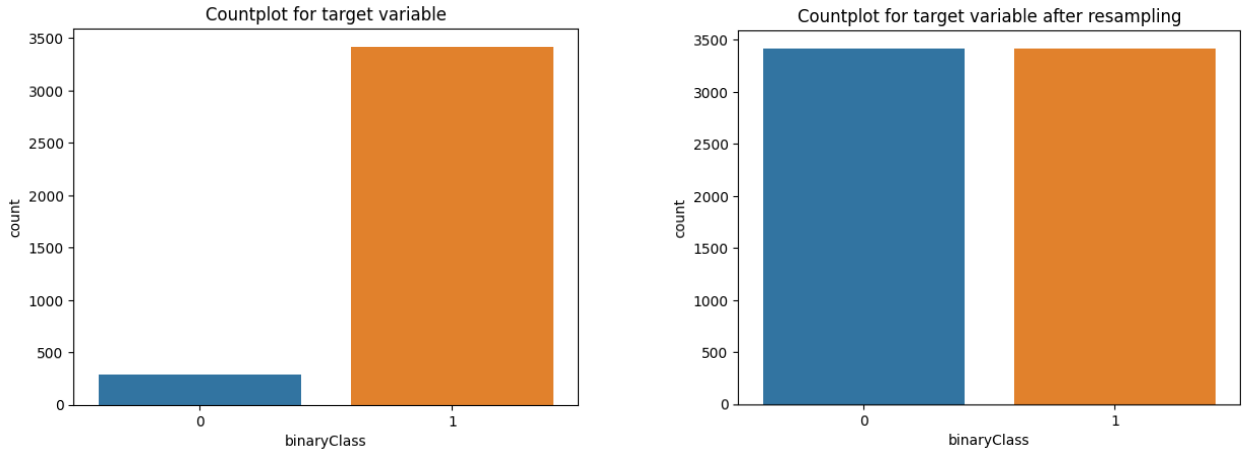

Figure 3. Count plot of binary classes of actual and resampled data.

Data preprocessing is a data-mining technique used to clean up and make the raw data that have been collected in the initial stage more workable. The Pandas python library was employed to perform the ML preprocessing. The actual data included attributes that have values such as letters or numbers. However, it must be encoded to make it readable by machines. Hence, data were converted into data frames, and encoding was carried out using the Sklearn label encoder. Then, all 29 attributes were converted into the "int64" data type. Later, data cleaning was performed by dropping duplicated rows from the encoded data, which ensured that the ML classifier input will not be redundant. Next, the statistical data summary was evaluated using count (nonempty values), mean, standard deviation, etc., as given in Table 2. How many values fall below the specified percentile is what the term "percentile" means.

Table 2. Statistical data summary details.

Index	Age	Sex	TSH	T3	TT4	T4U	FTI
count	3711	3711	3711	3711	3711	3711	3711
mean	46.493	1.266	123.688	30.4934	119.133	64.854	108.716
std	20.863	0.525	81.308	20.240	98.238	31.3301	97.0323
min	0	0	0	0	0	0	0
25%	28	1	63.5	18	21	46	17
50%	50	1	113	23	79	57	56
75%	63	2	167	32	226	71	221
max	93	2	287	69	241	146	234

In the field of medical diagnosis and treatment, the accurate prediction of disease outcomes and understanding the interdependencies among clinical variables or attributes are essential. Hence, before going further, the analysis of the data features was performed

using seaborn distribution plots, such as distplots and jointplots, where the former considers a single variable, and the latter handles two variables. Distplots showing the different attribute distributions are shown in Figure 4a–g.

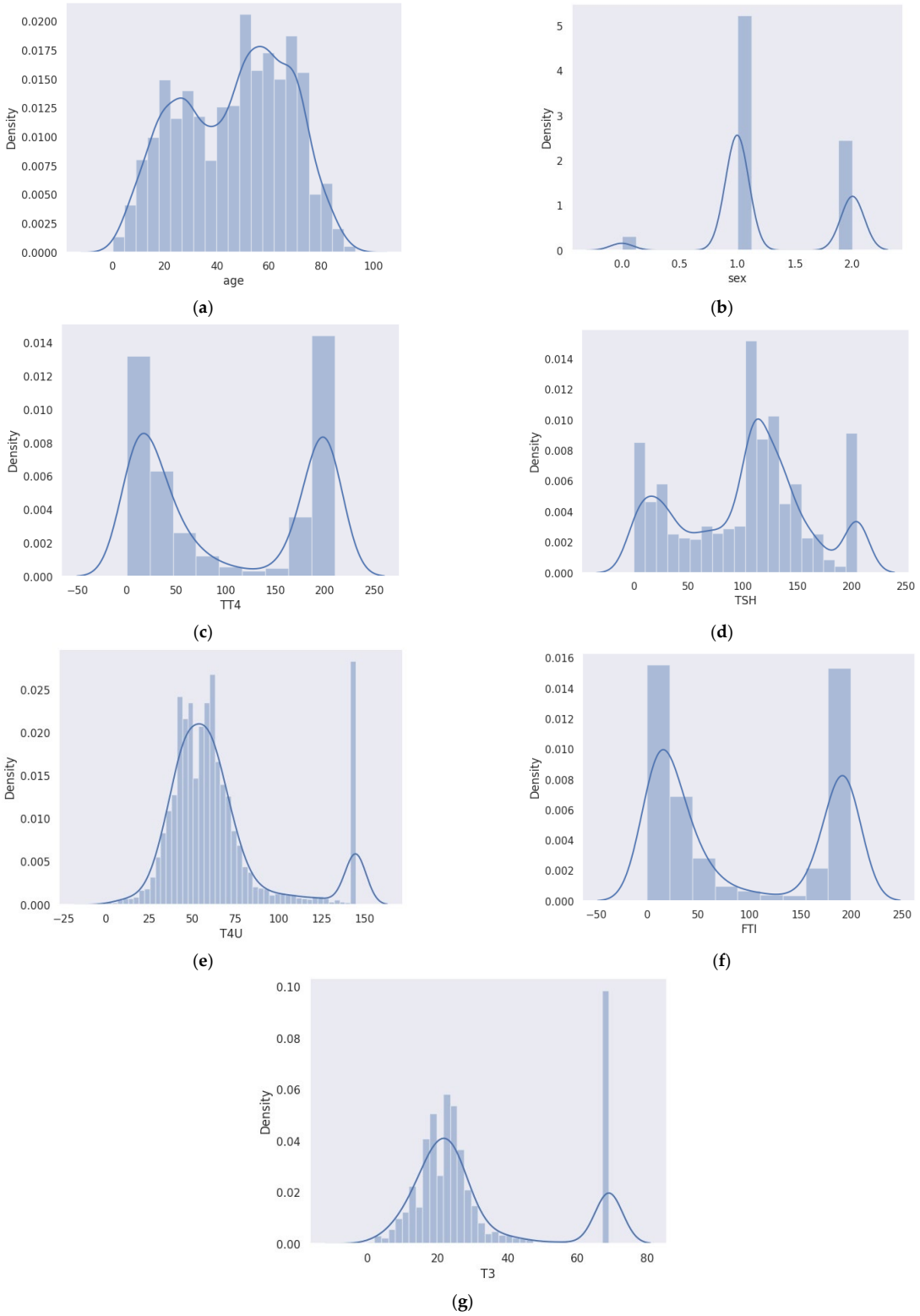

Figure 4. (**a**) Distplot of age. (**b**) Distplot of gender. (**c**) Distplot of total thyroxine (TT4). (**d**) Distplot of thyroid-stimulating hormone (TSH). (**e**) Distplot of thyroxine uptake (T4U). (**f**) Distplot of free thyroxine index (FTI). (**g**) Distplot of triiodothyronine (T3).

Distplots combine the functionality of histograms, kde plots, and rug plots in a simple and unified manner. Here, the rug plot was kept false. It is clear from the plot (Figure 4a) that thyroid issues become more common as people age. Additionally, the count of occurrence is more in females than in males when considering genderwise analysis, as shown in Figure 4b.

Total thyroxine (TT4) (Figure 4c) and triiodothyronine (T3) (Figure 4g) are thyroid hormones. Thyrotropin-releasing hormone, which is released by the hypothalamus, stimulates the pituitary gland to generate thyroid-stimulating hormone (TSH). The thyroid gland then releases T4 and T3 with the aid of TSH. TSH is essential to the system's operation. Therefore, the pituitary gland releases more TSH if the T3 and T4 levels are too low, as shown in the above distplot in Figure 4d. The gland will release less TSH if it is too high, but this give-and-take system only works if everything is in working order.

When the thyroid gland does not produce sufficient thyroid hormone to meet the body's needs, it causes hypothyroidism, also known as underactive thyroid. Primary hypothyroidism, caused by thyroid gland disease, is indicated by high FTI levels and elevated TSH levels (Figure 4f). Low levels of TSH and FTI suggest hypothyroidism caused by a malfunctioning pituitary gland. In the case of hyperthyroidism, TSH levels are low, while FTI levels are high. Figure 4e shows an increasing trend in T4U, whereas TT4 levels appear to be declining.

Both the main plot and the marginal plots make up the jointplot, as given in Figure 5. The combination of univariate and bivariate plots in a single figure is highly beneficial. This is so that the bivariate analysis can explore the link between two variables and explain the strength of that association, whereas the univariate analysis concentrates on one variable and describes, summarizes, and displays any patterns in your data. The Seaborn library's jointplot() function by default generates a scatter plot with two kernel density estimate (KDE) plots at the top and right edges of the graph. By setting the "hue" option to column "binary class" in this plot, the data points for thyroid (labeled as 1) and nonthyroid (labeled as 0) conditions are displayed in different hues and are clearly distinguishable. Regarding the marginal plots, density plots that separately display the data distribution for the two levels of the hue variable are plotted on both margins. If we observe the scatterplot of thyroid class, it can be noticed that the columns "age" and "TSH" appear to be positively correlated with one another as their values rise together. While considering the jointplot of "age" and "TT4", it is possible to observe the bimodal distribution of the density plots irrespective of age. However, for TSH and TT4, the nonthyroid class marginal plots are comparatively left skewed. From the analysis of these features, it can be concluded that each attribute plays a key role in predicting thyroid disease.

On the TT4, T4U, FTI, and age columns, min–max normalization was also performed. This is because, even with very rich data, if normalization is neglected, some traits may entirely outweigh others. Next, using the "Pearson" method (standard correlation coefficient), the pairwise correlation of all columns was determined. The row variable's correlation with the column variable determines the output cell value. Since the variable's correlation with itself is 1, each diagonal value is 1.00. Then, a heatmap was created that used various colors to depict the numerical values in the correlation data frame (Figure 6). Dark colors are used in this heatmap to show low values, and light colors are used to show high values. To the right of the figure is a color bar that illustrates how the colors and values relate to one another. Ticks are located at 0.75, 0.5, 0.25, etc., till 1.00 on the color bar. The minimum and maximum data values in the correlation data frame are used to compute the tick positions.

Following the completion of data preprocessing, the data features and target were separated from the preprocessed data frame. Excel files were used to export feature-target values to use in the ML stages. Using an 80:20 split, the dataset was divided into training and testing sets, using 80% data for training and 20% for testing. Here, we employed numerous ML models after partitioning the data with LR, NN, DT, KNN, and SVM algorithms with 5-fold cross-validation.

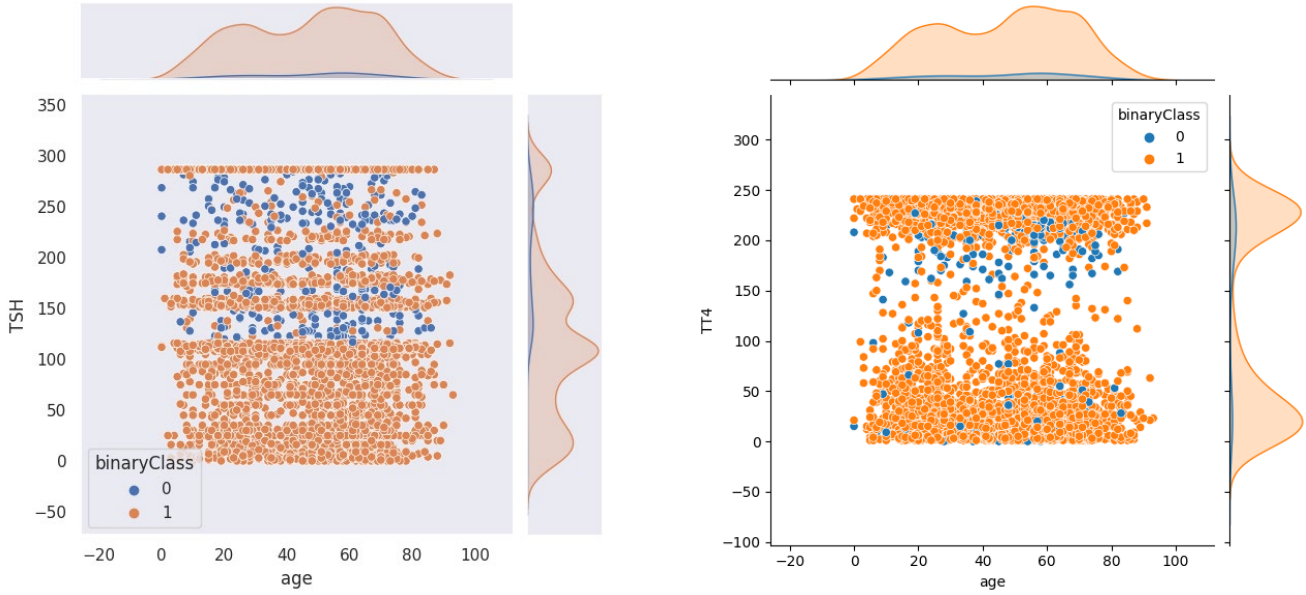

Figure 5. Jointplot of age with thyroid-stimulating hormone (TSH) and total thyroxine (TT4).

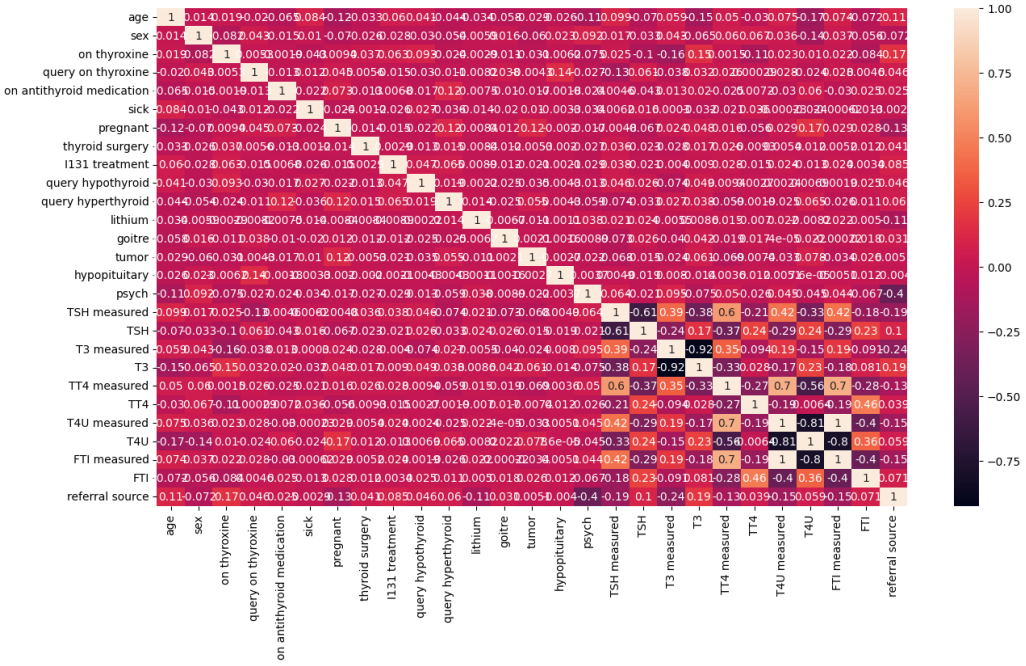

Figure 6. Heatmap showing the correlation between preprocessed data attributes.

The maximum classification accuracy of 95.62% was obtained from the DT classifier. However, the prediction accuracy required further improvement since the model is for medical diagnosis, where accuracy plays a key role. In addition, to incorporate the model into CAD real-time systems, reliability was another concern. To achieve these targets, the ensemble-learning algorithms of bagging and boosting were applied. They performed better than conventional classifiers, and the best accuracy achieved by boosting was 99.5%. Here, class 0 was used to denote nonthyroid classes, and 1 was used to denote thyroid classes. Also, we evaluated the model's negative predictive value (NPV), true-negative rate (TNR), positive predictive value (PPV), true-positive rate (TPR), and misclassification rate [37], as given in Table 3.

Table 3. Performance measures of machine-learning (ML) classifiers trained with preprocessed thyroid data.

Classifier	Class	PPV	NPV	TPR	TNR	Accuracy	Misclassification Rate
LR	0	0.8668	0.8893	0.8926	0.8628	87.77	0.1222
	1	0.8893	0.8668	0.8628	0.8926		
NN	0	0.9433	0.9541	0.9546	0.9426	94.86	0.0513
	1	0.9541	0.9433	0.9426	0.9546		
DT	0	0.9581	0.9626	0.9631	0.9494	95.62	0.0437
	1	0.9626	0.9581	0.9494	0.9631		
KNN	0	0.8919	0.9471	0.9505	0.8847	91.76	0.0823
	1	0.9471	0.8919	0.8847	0.9505		
SVM	0	0.8571	0.9062	0.9122	0.8479	88.01	0.1198
	1	0.9062	0.8571	0.8479	0.9122		
Bagging	0	0.9743	0.9890	0.9891	0.9739	98.15	0.0184
	1	0.9890	0.9743	0.9739	0.9891		
Boosting	0	0.9938	0.9958	0.9959	0.9938	99.5	0.005
	1	0.9958	0.9938	0.9938	0.9959		

Accuracy measures the count of data samples belonging to the test dataset that were properly categorized out of all the data samples (Equation (1)). The percentage of correct positive cases predicted out of all positive cases is known as sensitivity or TPR (Equation (2)). Specificity or TNR is the model's capacity to accurately categorize a data sample with a negative case among all the negative cases (Equation (4)). Precision or PPV assesses the model's performance by comparing the correct positive cases to those predicted by the model (Equation (3)). Negative predictive value (NPV) is the probability that a data sample with a negative screening test truly does not have the disease (Equation (5)). True positive (TP) and true-negative (TN) indicate positive and negative cases predicted correctly. False positive (FP) and false negative (FN) denote the false detection of negative cases as positive and positive cases as negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall\ or\ Sensitivity\ or\ True\ positive\ rate\ (TPR) = \frac{TP}{TP + FN} \tag{2}$$

$$Precision\ or\ Positive\ predictive\ value\ (PPV) = \frac{TP}{TP + FP} \tag{3}$$

$$Specificity\ or\ True\ negative\ rate\ (TNR) = \frac{TN}{TN + FP} \tag{4}$$

$$Negative\ predictive\ value\ (NPV) = \frac{TN}{TN + FN} \tag{5}$$

The models are fine-tuned by Bayesian optimization, and the classification error plot is shown in Figure 7. For the finalized boosting ensemble method, the training time required was 148.82 s with a prediction speed of 7800 obs/s. The hyperparameter range and optimized values are given in Table 4.

Table 4. Hyperparameters and values.

Hyperparameter	Range of Values	Optimized Value
Ensemble algorithms	GentleBoost, LogitBoost, AdaBoost, RUSBoost	GentleBoost
Number of learners	10–500	496
Learning rate	0.001–1	0.27021
Maximum number of splits	1–6839	5

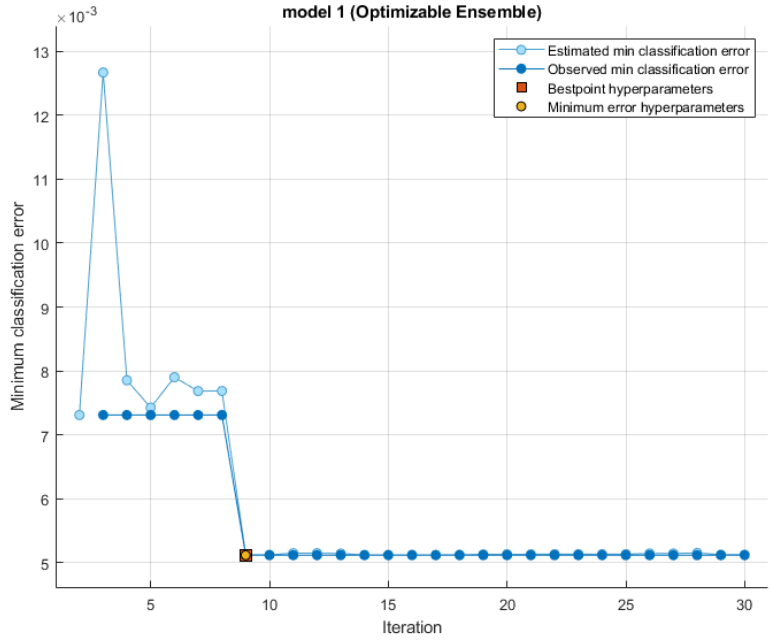

Figure 7. Classification error plot.

For the finalized thyroid prediction model using the boosting ensemble method, we exported the TPR and false-negative rate (FNR) plots (Figure 8) to see the classifier's performance for each class. Here, the false positives (FPs), true positives (TPs), false negatives (FNs), and true negatives (TNs) are 14, 3406, 21, and 3399, respectively. The TPR measures how many observations are accurately categorized for each true class. The FNR gives the ratio of mistakenly classified observations to correctly classified observations. The last two columns on the right of the plot display summaries for each true class. Since false positives are crucial to our classification problem, PPV and false-discovery rate (FDR) plots are obtained, as given in Figure 9. PPV represents the percentage of correctly classified data in each predicted class, while FDR measures the proportion of incorrectly classified observations for each predicted class.

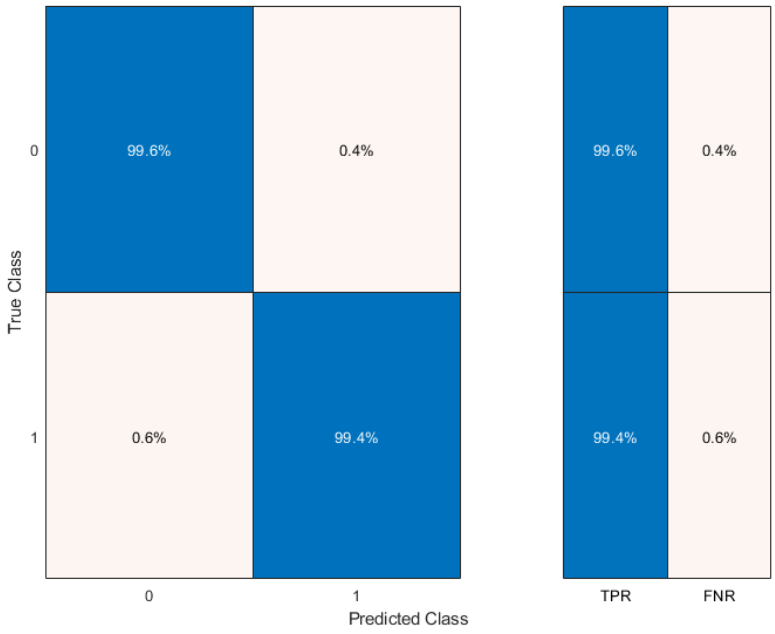

Figure 8. True-positive rate (TPR) and false-negative rate (FNR) plots.

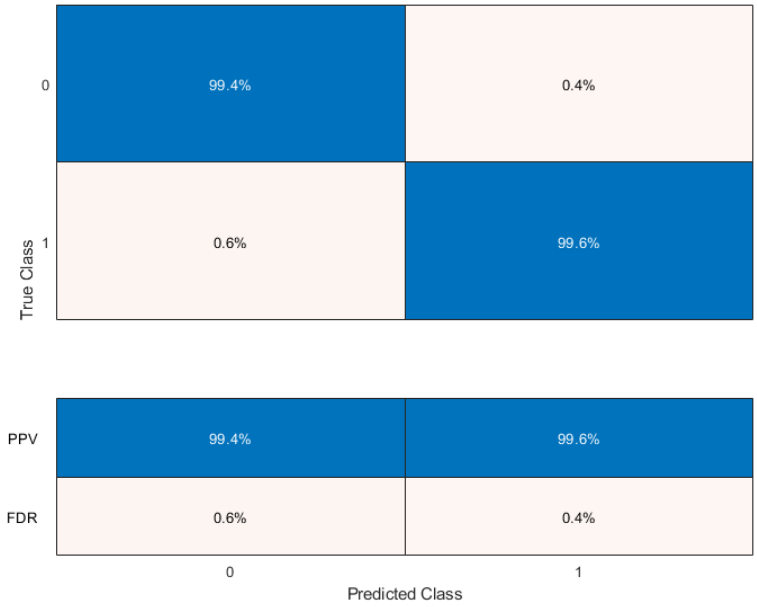

Figure 9. Positive predictive value (PPV) and false-discovery rate (FDR) plots.

The TPR and false-positive rate (FPR) for various thresholds of classification scores, calculated by the currently chosen classifier, are displayed on the receiver operating characteristic (ROC) curve (Figure 10). The integral of an ROC curve (TPR values) with respect to FPR from FPR = 0 to FPR = 1 is equivalent to the area under the ROC curve (AUC) value. The classifier's overall efficiency is gauged using the AUC value, and obtained a maximum value of 1, which indicates a higher classifier performance. An AUC of 1.000 indicates perfect separation between positive and negative classes in a model's predicted probabilities. However, this does not result in 100% performance in other metrics every time, as various matrices consider factors like classification threshold, class imbalance, and trade-offs. Each performance measure gives various performance aspects of a model's behavior.

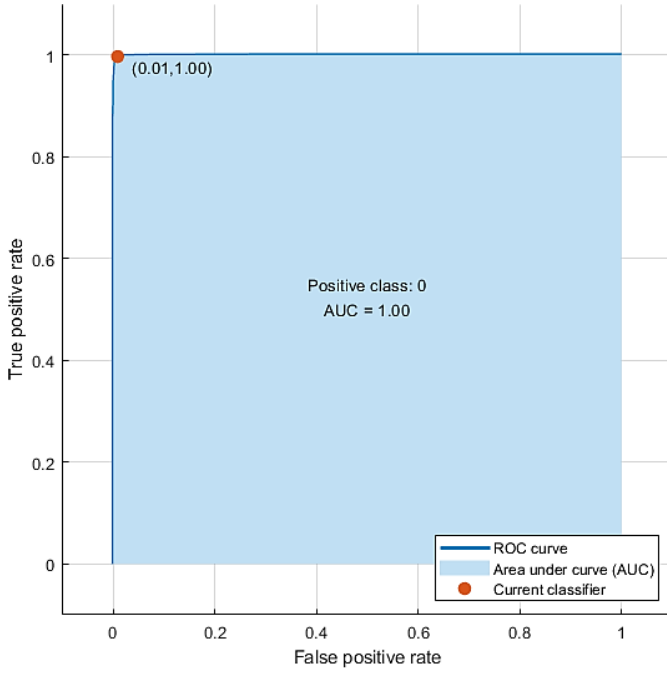

Figure 10. Receiver operating characteristic (ROC) curve.

A performance comparison of the proposed model with some similar works based on the same database is given in Table 5.

Table 5. Performance comparison table.

Reference	Methodology	Model Results
[38]	ANN classifier	Accuracy = 0.957, precision = 0.957, recall = 0.959, F1-score = 0.957
[12]	RF classifier	Accuracy = 94.8%, sensitivity = 94.8%, specificity = 91.2%
[27]	Multiple MLP	Accuracy = 99%
[39]	NN classifier	Accuracy = 98.4%
[40]	RF classifier	Accuracy = 99.14%
[41]	XGBoost classifier	Accuracy = 99%
Proposed work	Data mining + SMOTE + Bayesian optimization + Ensemble classifier (Boosting)	Accuracy = 99.5%, PPV = 99.59%, NPV = 99.39%, sensitivity = 99.39%, specificity = 99.59%, AUC = 1.00

Most of the individuals in the database are those who were referred to the thyroid clinic for evaluation. Since predictions relating to thyroid are the proposed model's main focus, this dataset is a useful resource for training and evaluating the model because it includes important clinical features for identifying thyroid disease. As a result, the model's applicability to this specific task could be ensured. SMOTE and cross-validation techniques are capable of mitigating bias and improving generalizability. Furthermore, this work could be extended for other related medical diagnosis applications by retraining the framework using an appropriate database.

The proposed model is valuable for real-time disease identification since it demonstrates several important qualities and characteristics such as high accuracy, robustness, balanced sensitivity and specificity, and adaptability, and it could be integrated into CAD systems as well. Incorporating our method into a software-based solution allows for patient data entry, and by utilizing the proposed ML model, it can predict the patient's thyroid state. Once a labeled public database is accessible in the future, a multiclass thyroid classification model will be tested. Additionally, deep networks will then be tried, coupled with the visualization of feature relevance in classification judgments. The proposed work details are available at Zenodo [42].

6. Conclusions

The identification of thyroid disorders has become increasingly important in the medical field due to their rising prevalence, early detection significance, treatment for preventing complications, and reducing mortality rates. In medical diagnosis and treatment, accurately predicting disease progression and understanding the interplay of clinical features are crucial. Hence, an effective real-time CAD system is required.

The majority of the studies that have already been conducted focus on a single model, which does not ensure model reliability or greater accuracy, which are the two main criteria for ML models used in medical systems. Even while certain DL techniques offer greater accuracy, in addition to the typical DL constraints, nobody is aware of the underlying criteria on which model decisions are made. Hence, in this work, all these limitations are eliminated by developing a conventional ML model with proper clinical feature analysis and an ensemble-learning approach. In medical diagnosis and therapy, predicting the course of diseases and the interdependence of clinical characteristics or features are fundamental. The model achieved an accuracy of 99.5% in addition to a PPV and specificity of 99.59%, NPV and sensitivity of 99.39%, and AUC of 1.00. Model reliability is ensured by the involvement of many models in making binary classification decisions. Additionally, model development with public databases helps the model to be stable with real-time system usage.

Funding: This research has received no external funding.

Institutional Review Board Statement: All methods were carried out in accordance with relevant guidelines and regulations.

Data Availability Statement: The data used in this paper is publicly available [29].

Conflicts of Interest: The author declares no conflict of interest.

References

1. Thyroid Gland Overview. Available online: https://www.endocrineweb.com/endocrinology/overview-thyroid (accessed on 13 April 2023).
2. Rashad, N.M.; Samir, G.M. Prevalence, risks, and comorbidity of thyroid dysfunction: A cross-sectional epidemiological study. *Egypt. J. Intern. Med.* **2020**, *31*, 635–641.
3. American Thyroid Association. General Information/Press Room. Available online: https://www.thyroid.org/media-main/press-room/ (accessed on 20 May 2023).
4. Thyroid Disease: Causes, Symptoms, Risk Factors, Testing & Treatment. Available online: https://my.clevelandclinic.org/health/diseases/8541-thyroid-disease (accessed on 13 April 2023).
5. Thyroid Function Tests: Procedure, Side Effects, and Results. Available online: https://www.healthline.com/health/thyroid-function-tests (accessed on 13 April 2023).
6. Roser, S.M.; Bouloux, G.F. Medical Management and Preoperative Patient Assessment. In *Peterson's Principles of Oral and Maxillofacial Surgery*; Springer International Publishing: Cham, Switzerland, 2022; pp. 19–51. [CrossRef]
7. Mirbabaie, M.; Stieglitz, S.; Frick, N.R.J. Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction. *Health Technol.* **2021**, *11*, 693–731. [CrossRef]
8. Fernandes, E.; Moro, S.; Cortez, P. Data Science, Machine learning and big data in Digital Journalism: A survey of state-of-the-art, challenges and opportunities. *Expert Syst. Appl.* **2023**, *221*, 119795. [CrossRef]
9. Holzinger, A.; Keiblinger, K.; Holub, P.; Zatloukal, K.; Müller, H. AI for life: Trends in artificial intelligence for biotechnology. *New Biotechnol.* **2023**, *74*, 16–24. [CrossRef]
10. Shankar, K.; Lakshmanaprabu, S.K.; Gupta, D.; Maseleno, A.; de Albuquerque, V.H.C. Optimal feature-based multi-kernel SVM approach for thyroid disease classification. *J. Supercomput.* **2020**, *76*, 1128–1143. [CrossRef]
11. Nadimi-Shahraki, M.H.; Taghian, S.; Mirjalili, S. An improved grey wolf optimizer for solving engineering problems. *Expert Syst. Appl.* **2021**, *166*, 113917. [CrossRef]
12. Alyas, T.; Hamid, M.; Alissa, K.; Faiz, T.; Tabassum, N.; Ahmad, A. Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach. *Biomed. Res. Int.* **2022**, *2022*, 9809932. [CrossRef] [PubMed]
13. Garcia de Lomana, M.; Weber, A.G.; Birk, B.; Landsiedel, R.; Achenbach, J.; Schleifer, K.J.; Mathea, M.; Kirchmair, J. In Silico Models to Predict the Perturbation of Molecular Initiating Events Related to Thyroid Hormone Homeostasis. *Chem. Res. Toxicol.* **2021**, *34*, 396–411. [CrossRef]
14. Riajuliislam, M.; Rahim, K.Z.; Mahmud, A. Prediction of Thyroid Disease (Hypothyroid) in Early Stage Using Feature Selection and Classification Techniques. In Proceedings of the 2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021—Proceedings, Dhaka, Bangladesh, 27–28 February 2021; pp. 60–64. [CrossRef]
15. Omuya, E.O.; Okeyo, G.O.; Kimwele, M.W. Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Syst. Appl.* **2021**, *174*, 114765. [CrossRef]
16. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; pp. 1–600. [CrossRef]
17. Wahid, A.; Khan, D.M.; Hussain, I.; Khan, S.A.; Khan, Z. Unsupervised feature selection with robust data reconstruction (UFS-RDR) and outlier detection. *Expert Syst. Appl.* **2022**, *201*, 117008. [CrossRef]
18. Rehman, H.A.U.; Lin, C.Y.; Mushtaq, Z.; Su, S.F. Performance Analysis of Machine Learning Algorithms for Thyroid Disease. *Arab. J. Sci. Eng.* **2021**, *46*, 9437–9449. [CrossRef]
19. Demir-Kavuk, O.; Kamada, M.; Akutsu, T.; Knapp, E.W. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. *BMC Bioinform.* **2011**, *12*, 412. [CrossRef] [PubMed]
20. Rehman, H.A.U.; Lin, C.Y.; Mushtaq, Z. Effective K-Nearest Neighbor Algorithms Performance Analysis of Thyroid Disease. *J. Chin. Inst. Eng.* **2020**, *44*, 77–87. [CrossRef]
21. Zhai, Y.; Song, W.; Liu, X.; Liu, L.; Zhao, X. A Chi-Square Statistics Based Feature Selection Method in Text Classification. In Proceedings of the IEEE International Conference on Software Engineering and Service Sciences (ICSESS), Beijing, China, 23–25 November 2018; pp. 160–163. [CrossRef]
22. Mishra, S.; Tadesse, Y.; Dash, A.; Jena, L.; Ranjan, P. Thyroid disorder analysis using random forest classifier. *Smart Innov. Syst. Technol.* **2021**, *153*, 385–390. [CrossRef]
23. Platt, J.C. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. 21 April 1998. Available online: https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/ (accessed on 22 August 2023).

24. Balasubramanian, S.; Srinivasan, V.; Thomo, A. Identifying Important Features for Clinical Diagnosis of Thyroid Disorder. In Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey, 10–13 November 2022; pp. 363–369. [CrossRef]
25. Zhang, X.; Lee, V.C.; Rong, J.; Lee, J.C.; Liu, F. Deep convolutional neural networks in thyroid disease detection: A multi-classification comparison by ultrasonography and computed tomography. *Comput. Methods Programs Biomed.* **2022**, *220*, 106823. [CrossRef]
26. Jha, R.; Bhattacharjee, V.; Mustafi, A. Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society. *Wirel. Pers. Commun.* **2022**, *122*, 1921–1938. [CrossRef]
27. Hosseinzadeh, M.; Ahmed, O.H.; Ghafour, M.Y.; Safara, F.; Hama, H.K.; Ali, S.; Vo, B.; Chiang, H.S. A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. *J. Supercomput.* **2021**, *77*, 3616–3637. [CrossRef]
28. Prathibha, S.; Dahiya, D.; Robin, C.R.; Nishkala, C.V.; Swedha, S. A Novel Technique for Detecting Various Thyroid Diseases Using Deep Learning. *Intell. Autom. Soft Comput.* **2023**, *35*, 199–214. [CrossRef]
29. Ross, Q. Thyroid Disease. UCI Machine Learning Repository. 1987. Available online: https://archive.ics.uci.edu/dataset/102/thyroid+disease (accessed on 22 August 2023).
30. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
31. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Lect. Notes Comput. Sci.* **2005**, *3644*, 878–887. [CrossRef]
32. Bayesian Optimization Book. Available online: https://bayesoptbook.com/ (accessed on 26 April 2023).
33. Opitz, D.; Maclin, R. Popular Ensemble Methods: An Empirical Study. *J. Artif. Intell. Res.* **1999**, *11*, 169–198. [CrossRef]
34. Schapire, R.E. The strength of weak learnability. *Mach. Learn.* **1990**, *5*, 197–227. [CrossRef]
35. Freund, Y. Boosting a Weak Learning Algorithm by Majority. *Inf. Comput.* **1995**, *121*, 256–285. [CrossRef]
36. Experiments with a New Boosting Algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 3–6 July 1996. Available online: https://dl.acm.org/doi/10.5555/3091696.3091715 (accessed on 20 May 2023).
37. Alshayeji, M.H.; Sindhu, S.C.; Abed, S. CAD systems for COVID-19 diagnosis and disease stage classification by segmentation of infected regions from CT images. *BMC Bioinform.* **2022**, *23*, 264. [CrossRef] [PubMed]
38. Islam, S.S.; Haque, M.S.; Miah, M.S.U.; Sarwar, T.B.; Nugraha, R. Application of machine learning algorithms to predict the thyroid disease risk: An experimental comparative study. *PeerJ Comput. Sci.* **2022**, *8*, e898. [CrossRef] [PubMed]
39. Trivedi, N.K.; Tiwari, R.G.; Agarwal, A.K.; Gautam, V. A Detailed Investigation and Analysis of Using Machine Learning Techniques for Thyroid Diagnosis. In Proceedings of the 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 1–3 March 2023; pp. 1–5. [CrossRef]
40. Sengupta, D.; Mondal, S.; Raj, A.; Anand, A. Binary Classification of Thyroid Using Comprehensive Set of Machine Learning Algorithms. In *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*; Springer Nature: Singapore, 2023; pp. 265–276. [CrossRef]
41. Alnaggar, M.; Handosa, M.; Medhat, T.; Rashad, M.Z.; Author, C.; Alnaggar, M. Thyroid Disease Multi-class Classification based on Optimized Gradient Boosting Model. *Egypt. J. Artif. Intell.* **2023**, *2*, 1–14. [CrossRef]
42. Alshayeji, M.H. Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers. 2023. Available online: https://zenodo.org/record/8272107 (accessed on 22 August 2023).