Personalized Patient Care Through AI-Driven Segmentation Predictive Modelling Thyroid Disease Detection Using Machine Learning

Dasari Girish

Ford Senior Data Scientist.

Keywords

Machine learning · Classification algorithm · Thyroid disease · Hypothyroidism Hyperthyroidism · Explainable artificial intelligence (XAI)

Abstract

The thyroid gland is the crucial organ in the human body, secreting two hormones that help to regulate the human body's metabolism. Thyroid disease is a severe medical complaint that could be developed by high Thyroid Stimulating Hormone (TSH) levels or an infection in the thyroid tissues. Hypothyroidism and hyperthyroidism are two critical conditions caused by insufficient thyroid hormone production and excessive thyroid hormone production, respectively. Deep learning modelscan be used to precisely process the data generated from different medical sectors and to build a model to predict several dis-eases. In this paper, we use different machine-learning algorithms to predict hypothyroidism and hyperthyroidism. Moreover, we identified the most significant features, which can be used to detect thyroid diseases more precisely. After completing the pre-processing and feature selection steps, we applied our modified and original data to several classification models topredict thyroids. We found Random Forest (RF) is giving the maximum evaluation score in all sectors in our dataset, and Naive Bayes is performing very poorly. Moreover, selecting the feature by using the feature importance method RF provides the best accuracy of 91.42%, precision of 92%, recall of 92% and F1-score of 92%. Further, by analyzing the characteristics and behavior of the dataset, we identified the most important features (TSH, T3, TT4, and FTI) of the dataset. In terms of accuracy and other performance evaluation criteria, this study could advocate the use of effective classifiers and features backed by machine learning algorithms to detect and diagnose thyroid disease. Finally, we did some explain ability analysisof our best classifier to understand the internal black-box of our machine learning model and datasets. This study could further pave the way for the researcher as well as healthcare professionals to analyze thyroid disease in real time applications.

1.Introduction

Thyroid disease is the most common endocrine condition worldwide, second only to diabetes, according to the World Health Organization (https://www.who.int/). (1). The most common thyroid gland illnesses, hyperthyroidism and hypothyroidism, have been identified in over 110 countries worldwide, endangering 1.6 billion individuals; most of these cases occur in Asia, Africa, and Latin America [2]. At the moment, more than 25,000 urgent care centers worldwide gather patient data in different ways. Unfortunately, studies are carried out utilizing the time-consuming and expensive traditional technique [3], which involves traditional examination and quantifiable tests. According to medical professionals, early illness diagnosis, treatment, and discovery are essential to halting the progression of an illness or even death. Clinical diagnosis is often considered an arduous endeavor, even after countless tries [4]. The thyroid is a small gland near the base of the neck that resembles a butterfly and is located directly below Adam's apple [5]. Numerous bodily functions are organized by the intricate network of glands that makes up the endocrine system. The thyroid gland secretes hormones that control metabolism in humans. While iodine deficiency is the most frequent cause, there are other potential causes as well [6]. The three hormones that the thyroid gland produces are

T3, T4, and calcitonin, with T3 and T4 being the most literal forms [7]. Both of these hormones are created with the help of iodine. Our bodies are unable to synthesize this trace element; thus we must obtain it through diet. Food in our intestines absorbs iodine into our circulation, where it eventually produces thyroid hormones. The condition known as hypothyroidism, or underactive thyroid, occurs when the thyroid gland does not generate enough of a particular hormone [8]. Several signs had been observed in the early stages of hypothyroidism. Absent

Giving hypothyroidism a lot of attention, fat may result from this. In addition, a number of additional issues such as heart disease, joint pain, and occasionally infertility may be observed in the patients [9]. A malfunction known as hyperthyroidism occurs when the thyroid gland creates an excessive amount of thyroid hormones that are released into the bloodstream. Among the signs of hyperthyroidism include jitters, irritability, and heightened appetite [10]. Machine learning, a branch of computer science that has gained enormous prominence recently and is probably going to continue doing so, could be used to predict thyroid problems early on. Several benefits of a machine learning algorithm include high parallelism, speed, self-learning, and fault tolerance to noise [11]. With the use of machine learning, people can now make sense of enormous volumes of data that would otherwise be too complicated or impossible to process. It is possible to forecast hypothyroidism and hyperthyroidism using patient symptoms and a machine learning model, which is a time- and money-efficient method. The input data used to train the machine learning model comes from multiple databases. Once trained, it can be used to provide predictions for further input data. The literature has a number of supervised machine learning methods [12–16]. In our study, we used a variety of classifiers to predict thyroid disease, including Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, K-Nearest Neighbor, Logistic Regression, and Support Machine Vector. By comparing the algorithms' performances, we were able to determine the best approach for more accurately predicting thyroid disease. Ultimately, the output and results generated by the black box machine learning algorithms are comprehended and trusted through the application of a post hoc method called explainable artificial intelligence.

The main objective of this study is

- To identify a machine learning classification technique that is trustworthy and uses the fewest features feasible to predict thyroid illness.

- Determining the key characteristic for thyroid illness detection

- Applying explainable AI to validate the experimental outcomes.

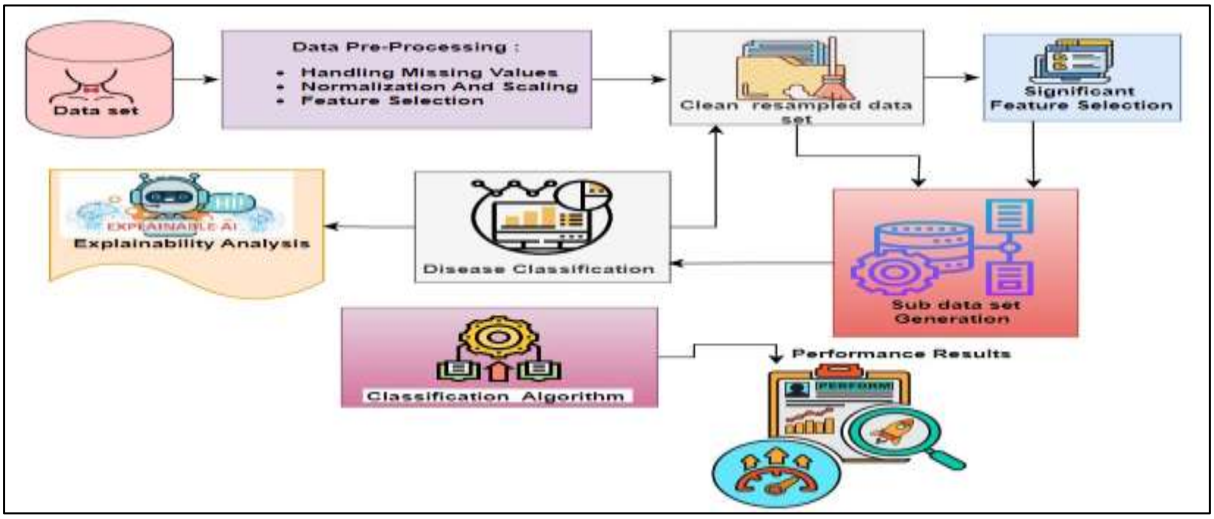


Figure -1 Proposed work flow block Diagram

The identified the key elements in our datasets that are most crucial for predicting thyroid disease. Lastly, we think that our study may have a big influence on the scientific community's comprehension of and use of machine learning in the medical profession, particularly in the prediction of thyroid disease. Fig. 1 shows the overall procedure for this investigation. First, we gather the data that has been cleaned, preprocessed, and resampled. After it, two subsets of the original dataset were created utilizing it to predict thyroid in total, we built the thyroid prediction model using seven phases. Finally, we used explainable AI to validate our findings.

2. Literature Review

Numerous studies have been conducted in this pertinent field thus far. Based on the thyroid dataset obtained from the UCI repository (https://archive.ics.uci.edu/ml/ datasets/thyroid + ease), the authors of [17] used a variety of classification algorithms, including Decision Tree, Support Vector Machine, Artificial Neural Network, and the K-Nearest- Neighbor algorithm. Classification and prediction were carried out, and accuracy was assessed based on the output supplied. These two approaches were compared in [18] using Logistics Regression and SVM machine learning algorithms to assess the Thyroid Dataset and RMS error, Precision, Recall, F1 measure, and ROC. They stated that logistic regression was discovered to be the most effective classifier. Awasthi and Anil Antony [19] talked about classifying and identifying thyroid ailment using machine learning methods, support vector machines (SVM), and KNN. In order to approximate missing values in user input for thyroid diagnosis, they used the K-nearest neighbor algorithm. In [20], a categorization scheme for the two types of thyroid disease—hyperthyroidism and hyperthyroidism was put forward. The mean value of the matching column is used to fill in any gaps left by missing values that are not a numerical limitation that are found during the preprocessing stage. The method of differential evolution is employed to convert parent records into child subsets. SVM was employed by the authors in [21] as a classifier to differentiate thyroid illness. Two datasets form the basis of this analysis. In [22], the authors used support vector machines and Naive Bayes for classification.

This theory is supported by a number of grouping algorithms, such as the K-nearest neighbor. The study was conducted using the Rapid Miner gadget, and the results show that K-nearest neighbor is a more accurate thyroid issue diagnosis method than Naive Bayes. With an accuracy of 93.44%, the K-nearest neighbor classifier was the most dependable, while the Naive Bayes classifier only achieved 22.56%. SVM outperformed Bayesian and K-Nearest Neighbor with an accuracy of 84.62 percent in [23]. KNN autonomously identified the nearest neighborhood. The authors of [24] have put out a number of data mining-based Thyroid prediction algorithms. They looked into the relationship between hyperthyroidism and hypothyroidism, as well as T3, T4, and TSH. Furthermore, many machine learning modification techniques have been employed by other writers recently to predict thyroid [25, 26]. Furthermore, for the thyroid classification in [27], authors employed a variety of feature engineering techniques, including forward-backward and bidirectional feature deletion. Pawar et al. [28] used the XAI approach in the field of healthcare to model integrity, openness in feature selection, result monitoring, and model sophistication. In 2021, the same author used explainable AI to provide a platform for healthcare professionals to understand machine learning models. To the best of our knowledge, Arjaria et al. in 2022 [29] employed XAI to predict the precision of decision tree algorithms with an explanation of important aspects, increasing the precision of the models and increasing their accountability by having the models justify each decision. We found that there is still a gap in explainable machine learning's capacity to forecast thyroid disease, despite the broad use of AI and machine learning in the diagnostics and medical domains. In order to categories thyroid illnesses, we first predict the optimal ML model and then use XAI to examine the "black box" of the optimal ML model.

3. Proposed Methods

3.1 Description of Dataset

Gathering data is a logical initial step in the creation of a machine learning model. The UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Thyro id+Disease) provided the data [30]. Our final dataset, which has 3221 entries, is the result of our merging three datasets from the UCI machine learning repository: ill, hyperthyroid, and hypothyroid.

There are thirty features in all; Table 1 illustrates that six of the features are real number qualities and the remaining features are category traits. To enhance the quality of the dataset acquired for additional analysis, pre-processing is carried out. After removing the two attributes (TBG measured and TBG) due to a significant missing value, the histogram of all the attributes is shown in Fig. 2.

3.2 Pre-processing of Data

Real-world raw data is sometimes unreliable, lacking in specific behaviors or trends, and incomplete. They probably contain a lot of errors as well [31]. Therefore, before they are collected, they are pre-processed into a format that the machine learning algorithm can use for the model. To get the highest model quality, considerable attention should be paid to the data pre-processing stage. It comprises a number of actions used in the procedure to increase the relevance of the data. The following procedures are what we used to preprocess the data for this study.

Many ambiguous values initially had no real significance. Therefore, we eliminated those ambiguous values in order to reduce the dataset's properties and obtain better results from this approach. Since there were so many missing entries in our dataset, we then replaced the missing value. To deal with missing numbers, we used a variety of techniques. For instance, we filled in the gaps with the median and mode. Additionally, the integer format of the categorical data was encoded so that data

Table 1 Description of thyroid

Sl. No.	Attribute	Value type	Sl. No.	Attribute	Value type
01	Age	Continuous	16	Psych	f, t
02	Sex	M, F	17	TSH measured	f, t
03	On thyroxine	f, t	18	TSH	Continuous
04	Query on thyroxine	f, t	19	T3 measured	f, t
05	On antithyroid medication	f, t	20	T3	Continuous
06	Sick	f, t	21	TT4 measured	f, t
07	Pregnant	f, t	22	TT4	Continuous
08	Thyroid surgery	f, t	23	T4U measured	f, t
09	I131 treatment	f, t	24	T4U	Continuous
10	Query hypothyroid	f, t	25	FTI measured	f, t
11	Query hyperthyroid	f, t	26	FTI	Continuous
12	Lithium	f, t	27	TBG measured	f, t
13	Goitre	f, t	28	TBG	Continuous
14	Tumor	f, t	29	referral source	WEST, STMW, SVHC, SVI,

					SVHD, other
15	Hypopituitary	f, t	30	category Negative, hypo-thyroid, sick, hyperthyroid	

M male, F female, t true, f false, TSH thyroid stimulating hormone, T3 triiodothyronine hormone, TT4 thy-roxine hormone, T4U thyroxine utilization rate, FTI free thyroxine index with transformed category values may be fed into models toimprove prediction accuracy. Furthermore, we handled the imbalanced data from our datasets in which the target classhad an unequal distribution of observations. For balancing our dataset, we employed the resampling technique. Finally,we spilled the datasets into training and test sets. The training dataset was utilized to fit the model, and test sets were used to make predictions and compare them to the predictedvalues. In this study, 70 out of 100 data was used for training, and 30 out of 100 was used for testing.

3.3 Feature Selection Methods

Feature selection is a strategy for limiting the input variable to the model by removing insignificant data and only using valuable data [32]. The purpose of feature selection in machine learning is to determine the best set of characteristics for building effective models of the phenomena being studied. In this study, for selecting the most important feature, we used the univariate feature selection approach and the feature importance method [33].

3.4. Selection of the Classification Algorithms

Before selecting an algorithm, there are a few things to remember, the size of the training data, the output's accuracy and/or interpretability, time spent on training or speed, linearity, and the number of features [34]. In this investigation, we took seven popular machine learning classification algorithms for solving this dataset because we are trying to figure out which algorithm performs better on our dataset. In order to predict the thyroid, we use Decision Tree Classifier, Random Forest Classifier, Gra- dient Boosting Classifier, Naive Bayes Classifier, Logistic Regression, K-Nearest Neighbor, and Support Vector Machine (SVM) algorithms. Supervised Machine Learning algorithms like Decision simplistic, properly trained and optimized models frequently outperform them. Computational efficiency, ease of regularization, and sim- plicity in implementation are some benefits of Logistic Regression. However, its inability to tackle non-linear problems, susceptibility to overfitting, and poor performance until all independent variables are recognized may sometimes causes problem of using this algorithm. Unsupervised K Means Clustering is a widespread choice for clustering problems. When variables are large, it is computationally more effective than hierarchical clustering. The algorithm's order of complexity, making it computationally efficient. However, K value prediction is challenging and the performance of globular clusters is compromised. We can use SVM for both classification and regression problem. The decision boundary, a hyperplane is required to divide a collection of objects into their many classes. It can manage structured and semi-structured data. Moreover, it can manage complex functions if the right kernel function can be determined. SVM has a lower likelihood of over fitting. With a huge data collection, though, its performance suffers because of the longer training times.

3.5. Evaluation of the Model

In machine learning, performance metrics refer to how well an algorithm per- forms depending on various criteria such as precision, accuracy, recall, and F1 score [35–37]. The following sections go through several performance metrics.

3.5.1. Accuracy

The percentage of correct test data predictions referred to as accuracy. It is easy to calculate by dividing the number of forecasts by the number of correct guesses. The formula for calculating the accuracy is given below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad (1)$$

Trees are typically used to tackle classification and regression issues by splitting the data based on specific criteria. While the data is divided among the nodes, the final decision is provided by the leaves. The problems with this method are over-fitting, although Random Forest offers a solution that is based on an

3.5.2 Precision

The precision score is used to assess the model's correctly counting genuine positives among all positive predictions. Thefollowing is the formula for calculating precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad (2)$$

3.5.3 Recall (Sensitivity)

The recalls score used to assess the model's performance in terms of accurately counting true positives among all actual positive values. Below is the formula for determining the recall.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (3)$$

3.5.4 F1 Score

The F1-score is the harmonic mean of precision and recall score, and utilized as a metric in situations when choosing either precision or recall score can result in a model with excessive false positives or false negatives. The F1 score measured as follows.

$$\text{F1 Score} = \frac{2\,(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \qquad (4)$$

After combining three datasets, our final thyroid dataset had 3221 number of instances of 3221 patients. Along with the target value, we had 30 attributes. There were no missing values in our data. When we looked back at the original dataset, there were missing values in several columns. 'nan' is used to replace these values. Then we convert it into the numerical format. Because the missing values, except sex, are from numeric attributes, they are replaced with the median value of the respective columns. However, sex is categorical attribute, and the missing value of it is replaced with a mode value of the respective attribute. Initially, wedropped two attributes, TBG and TBG measured, as the majority of values of these attributes are missing. Because the majority of data of these columns are missing. Our categorical attribute is mapped to numeric values, done manually with programming. For converting those values intonumeric values, we use a label encoder. Our other attributes are in the form of objects. As a result, we convert them to integer format to ft them into our model. Our dataset is imbalanced because the target class has an uneven distribution of observations. There are 2753 observations under the negative class label, 220 observations under the hypothyroid class label, 171 observations under the sick class label, and77 observations under the hyper- thyroid class label. When dealing with unbalanced datasets, typical machine learning methods may create biased, erroneous, and unsatisfactoryclassifers. Standard classifier methods favour classes with many instances, such as Decision Tree and Logistic Regression. Typically, they can only anticipate data from the vast majority of classes. The minority class's traits are frequently dismissed as noise and ignored. As a result, the minority class has a higher chance of being classified than the majority class. Because Machine Learning Algorithms are typically design to improve accuracy by reducing error, this occurs. Therefore, we convert the dataset into a balanced one to obtain the desired result. We use a resampling technique to ensure that the minority and majority classes are equal. Finally, the distribution of observations in our dataset is even across our entire class.

3.6. Artificial Intelligence Explained

Explainable artificial intelligence (XAI) is a set of methodologies and strategies that eventually allow human users to understand and trust the output and results produced by black-box machine learning technologies. A post-hoc XAI technique was employed to explain the model in this work. The post-hoc methodologies analyse the model after it has been trained, but without limiting the model's complexity. As a result, the model's explain ability has no effect on its performance. The use of intrinsic techniques limits the complexity of the machine learning model. Again, depending on the scope, there could be two options. There are two sorts of explain ability: global and local. A machine learning (ML) model's global explanation specifies which the features are critical to the overall model's success. by contrast A local method explains only a single data point. In this case, we employed Shapley additive explanations (SHAP) [38] in this investigation. and locally interpretable model-independent explanations (LIME). [39] for both global and local explanations. For N number of explanatory variables, in terms of local accuracy each prediction made by the SHAP method is approximated by f(x) with g(x′), and a quantity ϕj∈R. Which can be defined as follows [40]:

$$F(x) = g(x') = \emptyset_0 + \sum_{j=1}^{N} \emptyset_j \; x_j^i \tag{5}$$

Three major properties of SHAP Local accuracy, messiness, and consistency can only be satisfied by one explanatory model defined by as follows:

$$\emptyset_j \; (f, x) = \sum_{zj}^{N} \frac{\lfloor z' \rfloor! (N - \lfloor z' \rfloor - 1)!}{N!} (f_x \lfloor z' \rfloor) - f_x \lfloor z' \rfloor / J \tag{6}$$

where z′∈{0,1} N: binary variable’s linear function, z′\j representing setting at zi ′=0, and non-zero entities is denoted by |z′|.

Whereas LIME attempt to ft a local model with sample data points that resemble the observation being addressed. Thus, each observation x of LIME can be obtained by as follows [40]:

$$\lambda(x) = \mathrm{argmin}_{\phi \epsilon Q} L(f, q, \pi_x) + \psi_q \tag{7}$$

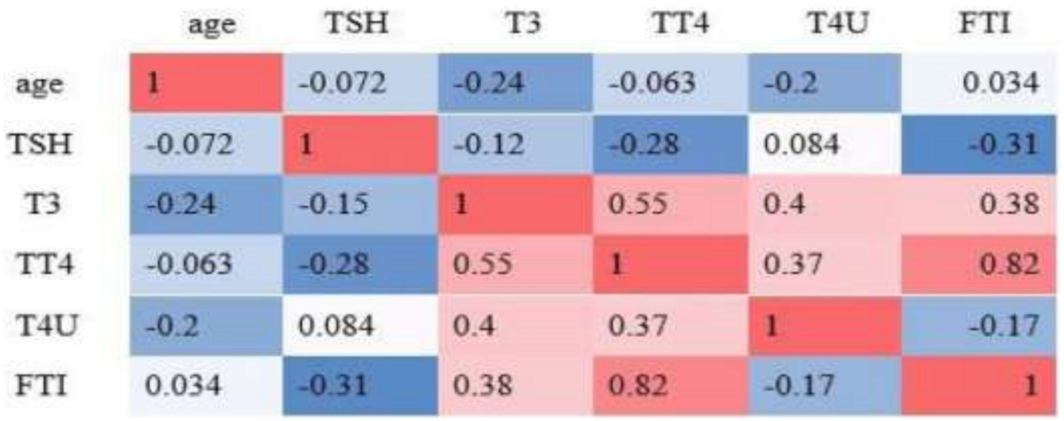
where locality aware loss L, potentially interpretable models is denoted by Q, πx(z): distance between an instance z and x, and ψ(q): A metric for the explanation's complexity q∈Q.

4 Result and Discussion

A comparison of seven different machine-learning algorithms was conducted in this study. Decision Tree Classifier, Random Forest Classifier, Naive Bayes Classifer, Gradient Boosting Classifer, Logistic Regression Classifer, K- Nearest Neighbor, and Support Vector Machine were utilized for thyroid disease prediction. Firstly, we collect and pre-processed the data and then fed the data to train the model. By comparing the scores, various performance criteria, including accuracy, precision, recall, and F1-score, are utilized to establish whether an algorithm is superior to others. We divide our dataset into three formats: the frst set considering all attributes, the second set with 14 feature selection process attributes, and the third with 14 univariant feature selection process attributes. We narrowed down attributes based on their correlation with the target, which we calculated with the feature selection process and univariant feature selection methods. Overall, the results of various algorithms are explained in the next part of this result analysis.

4.1 Descriptive Statistics of the Dataset

Exploratory data analysis (EDA) is a sort of data analysis that employs data visualization to evaluate and investigate data sets and describe their key properties [41, 42]. EDA is mainly used to examine what data might reveal outside formal modelling or hypothesis testing tasks and to understand variables and their interactions better. It can also help us to figure out if the statistical methods we are contemplating for data analysis are appropriate. Our dataset has 28 attributes, with only six of them being numeric. Therefore, we give a short descriptive statistic of our dataset in Table 2. We can see that all of the attributes have 3221 values in this table. So, before we train the model, we use various techniques to fall in the missing values. We can also see that the average age of the patients is 52.4, implying that the most patients were elderly. The youngest person was one-year-old, and the oldest person was 94 years old. The age distribution of the data is skewed, indicating that the population with a low age is absent. The standard deviation is 19.1, indicating the sparseness of the age group, which ranges from 57 to 73 years old. TSH mean was 6.322 mIU/L, indicating that most patients' TSH levels were not expected. TSH levels should be between 0.5 and 5.0 mIU/L to be considered normal. TSH had a minimum value of 0.005 mIU/L and a maximum value of 478.0 mIU/L. The mean T3 value was 1.95 nmol/L, with a minimum of 0.05 nmol/L and a maximum of 10.6 nmol/L. The mean value of TT4 is 107.55. The maximum value of TT4 is 430 and the minimum of TT4 is 2. In the case of T4U, the mean value is 0.988 mIU/mL. The maximum value of T4U is 2.12 mIU/mL and the minimum value of T4U is 0.31 mIU/mL. Moreover, the mean value of FTI is 110.26. The correlation between all the numeric data is depicted in Figure 3 shows that TT4 and FTI have a strong relationship. We can get a better understanding of this correlation table if we look at the heat map. Fig. 4 depicts a heat map of all attribute correlations. Form the heat map and numerical correlation of the above figure, we can draw interpretation about the correlation among the variable. It is clear from the heat map that T4U measured and FTI measured has very strong correlation. Moreover, some other parameter's also visualized very strong relationship like TT4 with T3, T4U with FTI, TT4 with FTI and TTI with T4U.



	age	TSH	T3	TT4	T4U	FTI
age	1	-0.072	-0.24	-0.063	-0.2	0.034
TSH	-0.072	1	-0.12	-0.28	0.084	-0.31
T3	-0.24	-0.15	1	0.55	0.4	0.38
TT4	-0.063	-0.28	0.55	1	0.37	0.82
T4U	-0.2	0.084	0.4	0.37	1	-0.17
FTI	0.034	-0.31	0.38	0.82	-0.17	1

Fig. 3 Correlations among the numeric value attributes of the dataset

Table 2. Descriptive statistics of numeric value of our dataset

Characteristics	**Age**	**TSH**	**T3**	**TT4**	**T4U**	**FTI**
Count	3221	3221	3221	3221	3221	3221
Unique	94	264	65	218	139	210
Unit	Years	mIU/L	nmol/L	–	mIU/mL	–
Freq	91	247	589	142	276	274
Mean	52.4	6.322	1.95	107.55	0.988	110.26
Std	19.1	26.54	0.8399	38.09	0.186	35.967

Min	1.0	0.005	0.05	2.0	0.31	2.0
25%	37.0	0.58	1.6	86.0	0.88	93.0
50%	55.0	1.5	1.9	102.0	0.97	106.0
75%	68.0	3.0	2.2	123.0	1.07	123.0
Max	94.0	478.0	10.6	430.0	2.12	395.0

4.2. Performance Analysis of Different Algorithm

Our original dataset, which included all features, was first utilized to evaluate several machine learning measures. After that, we used our balanced dataset to test multiple machine-learning models. This study selected the dataset's importantfeatures using feature importance methods and univariate feature selection techniques. In our experiments, those vitalfeatures are then used to identify the model's precision, accuracy, recall, and F1 score. The data we use is typically divided into two categories: training data and test data. In this study, 70% of the data was utilized for training and 30% for testing. So, out of our11,012 dataset instances, 7708 were used for the training set.3304 of the 11,012 dataset instances were used in the testingset. Using the testing, we can determine the accuracy of ourmodel and how well it can predict thyroid disease. We usedthe Sklearn library to split our data set as a train and test set.Sklearn, model selection train, and test split library component, split the dataset randomly with specified portion, and we get the random train and test part from the entire data- set. After training the model with all algorithms, the testingdataset was used to test the methods. The F1-score, recall, precision, and accuracy were used to evaluate the models. Performance. The entire study aimed to see which algorithmcould best classify diseases. This section highlights the study's outcomes and introduces the top performer based on several performance criteria. At first, performance was measured using our raw data-set. Secondly, performance was measured using a dataset containing 14 attributes derived from the feature importance method. Third, performance was determined by considering 14 attributes from the univariate feature selection. Finally, we compare various performance metrics of various algorithms and feature categories

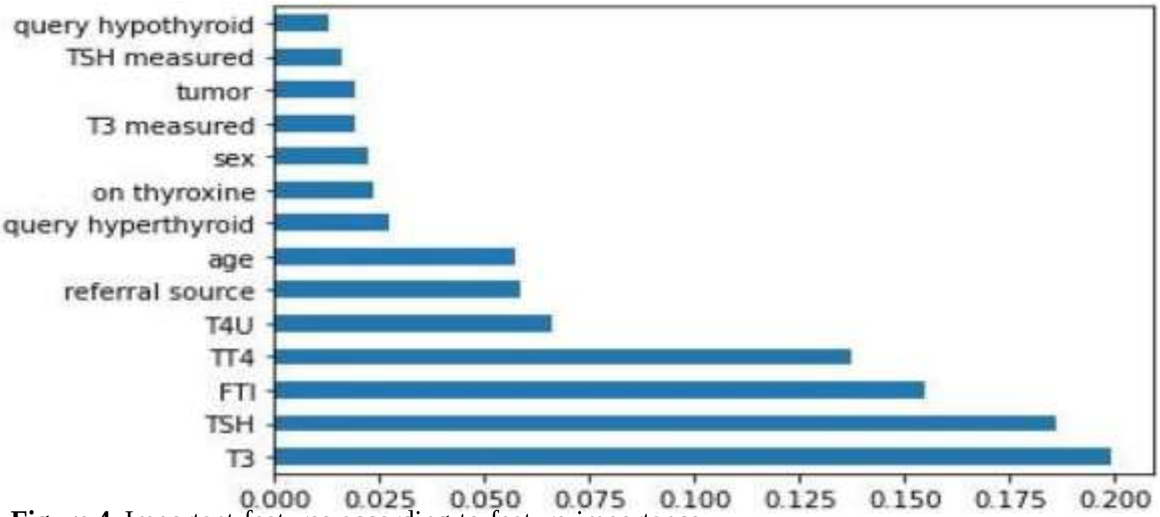


Figure 4. Important features according to feature importance

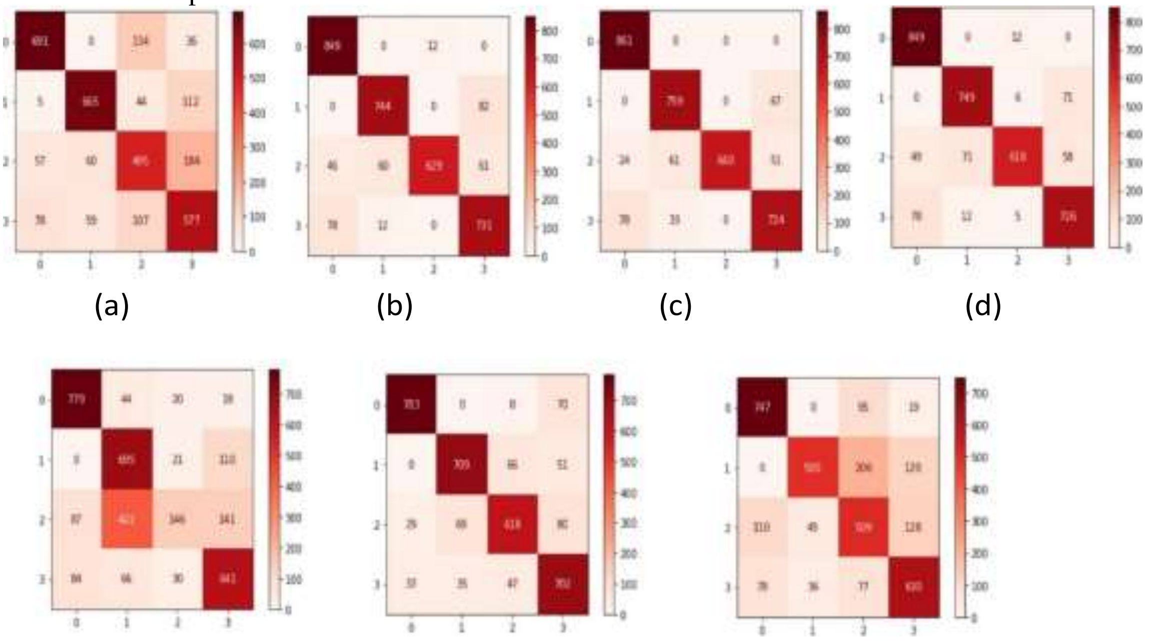
Table 3 Evaluation of algorithms with the features of feature importance

Algorithm names	Accuracy	Precision	Recall	F1-score
Decision tree classifier	90.43	91	90	90
Random forest classifier	91.42	92	92	92
Gradient boosting classifier	90.5	91	90	90
Naive Bayes classifier	67.86	68	67	64

K-nearest neighbor	86.22	86	86	86
Logistic regression	73.15	86	86	86
Support vector machine	73.7	74	74	74

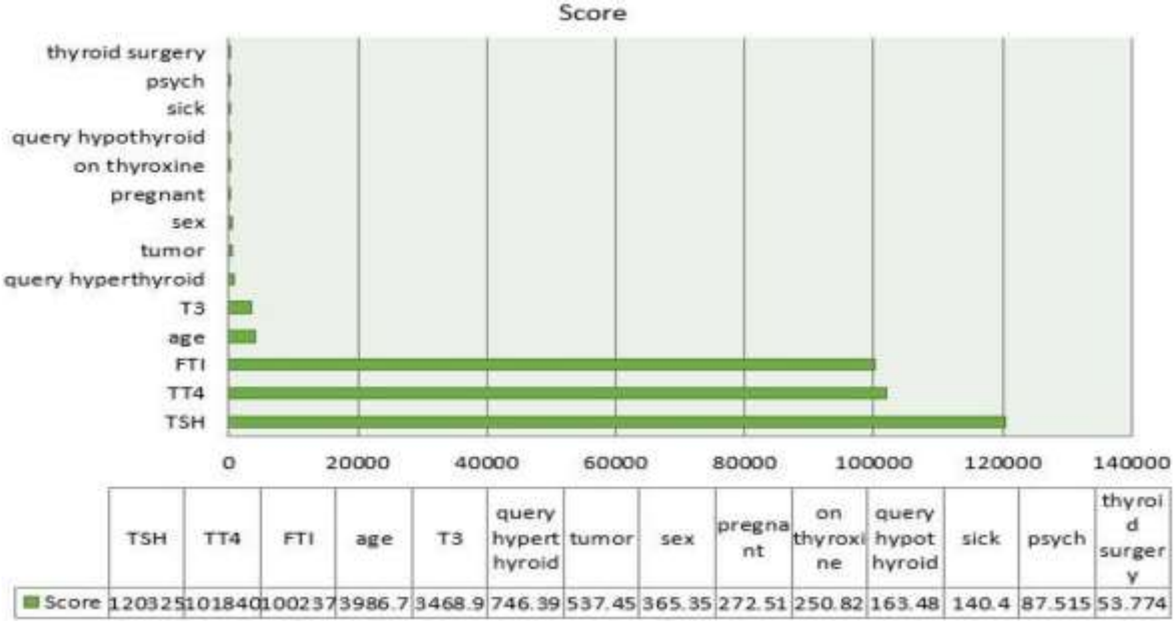
4.1.1 Results Using All Features

The applied selected algorithms to our dataset. Our datasethas 28 attributes; among them, the category is the target. Thealgorithms are then compared using various performance metrics. We can see from Table 4 that the Logistic Regression algorithm has the highest accuracy of any algorithm. After Logistic Regression, Support Vector Machine, Gradient Boosting Classifier, and Decision Tree Classifier have higher accuracy. Predictor accuracy refers to how well a predictor can forecast the value of a predicted characteristicfor fresh data.



(a) (b) (c) (d)

Fig. 5 Confusion matrix of different algorithms a SVM, b DT, c RF, d GB, e Naive Bayes, f KNN and



	TSH	TT4	FTI	age	T3	query hyperthyroid	tumor	sex	pregnant	on thyroxine	query hypothyroid	sick	psych	thyroid surgery
🟩 Score	120325	101840	100237	3986.7	3468.9	746.39	537.45	365.35	272.51	250.82	163.48	140.4	87.515	53.774

Figure 6. Top 14 features selectedusing univariate feature selection procedure based on their score

In contrast, classifier accuracy refers to a classifier's ability to predict the class label correctly. However,accuracy does not always provide good performance metrics to compare algorithms, so consider other metrics, for instance, recall, precision, and F1 score. We now assess our model's performance using various metrics such as recall, precision, and F1 score. Logistic Regression, as shown in Table 4, outperforms in terms of accuracy. However, this algorithm's precision, recall, and F1 score are all low. The accuracy of 84.48%, precision of 25%, recall of 24%, and F1 score of 25 from Logistic Regression, which outperforms the other six classification algorithms for this dataset. The Support Vector Machine, Gradient Boosting Classifier, and Decision Tree Classifier perform as well. However, precision, recall, and F1-score are all extremely low in each case. As a result, we can only measure them using accuracy. However, accuracy cannot always provide us with an accurate measure of performance. Random Forest has a 74.4% accuracy, but precision, recall, and F1 score are all low. The accuracy of the K-Nearest Neighbor is 72.18 percent. On the other hand, Naive Bayes gives us a low score for this experiment. This algorithm only has a16.44 percent accuracy, which is highly unsatisfactory. From the result, we can also say that Logistic Regression gives us the best prediction for our dataset. Naive Bayes gives us the poorest prediction in this case. As a result, we can conclude that for our dataset, Logistic Regression is the best classification algorithm, while Naive Bayes is the worst.

Table 4 Evaluation of algorithms with the features of univariate feature selection

Algorithm name	Accuracy	Precision	Recall	F1-score
Decision tree classifier	89.55	90	89	89
Random forest classifier	90.4	91	90	90
Gradient boosting classifier	89.35	90	89	89
Naive Bayes classifier	56.3	63	55	50
K-nearest neighbor	86.07	86	86	86
Logistic regression	71.82	86	86	86
Support vector machine	74.15	74	74	74

4.1.1 Results for Our Dataset Using Feature ImportanceMethod

We determine our 14 best-correlated features from our dataset using the feature importance technique. We apply the seven algorithms to the 14 features chosen using the method. The algorithms are then compared using various performance metrics. All the selected features are presentedin Fig. 7 with their importance value.

We apply Random Forest Classifier, Decision Tree Classsifier, Gradient Boosting Classifier, Naive Bayes Classifier,Logistic Regression Classifier, K-Nearest Neighbor, and Support Vector. We can see from the above bar chart that the Random Forest algorithm outperforms all others in termsof accuracy. After Random Forest, Decision Tree Classifier and Gradient Boosting Classifier have higher accuracy.As previously stated, accuracy is not always an appropriatemetric when comparing algorithms, so consider alternative metrics like precision, recall, and F1-score. The performancemetrics of all seven algorithms are listed in Table 5.

Random Forest beats all other performance criteria, suchas accuracy, precision, recall, and F1 score, as seen in the table above. We have the highest accuracy of 91.92 per- cent, the highest precision of 92 percent, the highest recall of 92 percent, and the highest F1 score of 92 percent. So, forour dataset with 14 feature importance attributes, Random Forest outperforms the other six classification algorithms. Following that, the Gradient Boosting and Decision Tree Classifier perform admirably. However, both the Decision Tree Classifier and the Gradient Boosting Classifier have the same precision, recall, and F1 score. Moreover, in the case of Gradient Boosting, accuracy is improved. So, in terms of accuracy, we can say that Gradient Boosting outperforms Decision Tree Classifier. K-Nearest Neighbor has an accuracy of 86.22 percent and an F1 score of 86 percent. With a73.7 percent F1 Score, SVM provides 73.7 per- cent accuracy. With an F1 score of 86 percent, Logistic Regression has a 73.15 percent accuracy. Finally, Naive Bayes gives a 64 percent F1 score and 67.86 percent accuracy, respectively. The confusion matrix tells us how accurate the classifier is at making predictions. The confusion matrix of all seven classification algorithms is shown in Fig. 8. From the confusion matrix, as shown in Fig. 8, we can also say that Random Forest gives us the best prediction, and Naive Bayes gives us the poorest prediction in this case. As aresult, we can conclude that for our chosen dataset, RandomForest is the best classification algorithm.

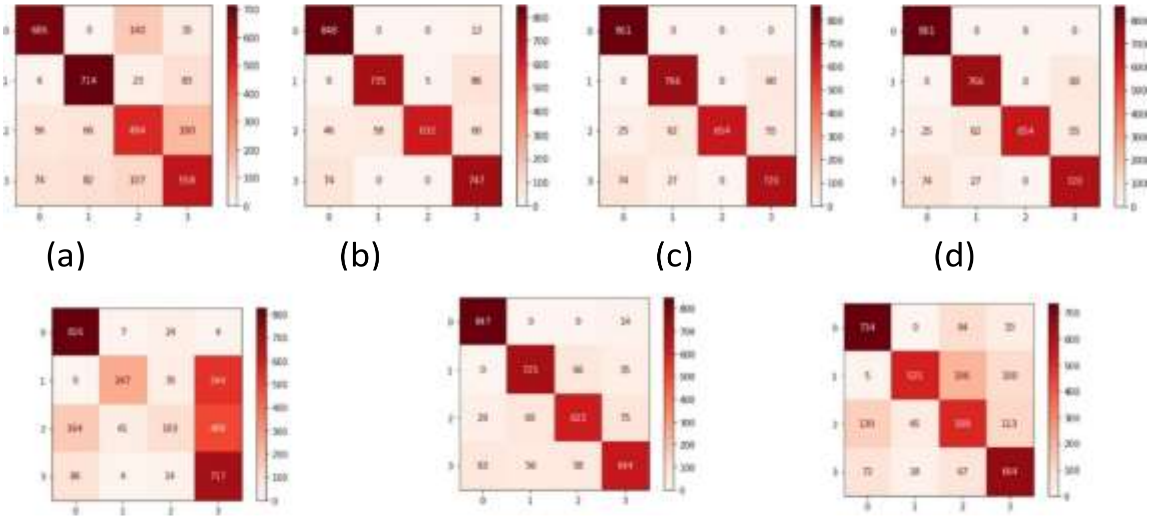
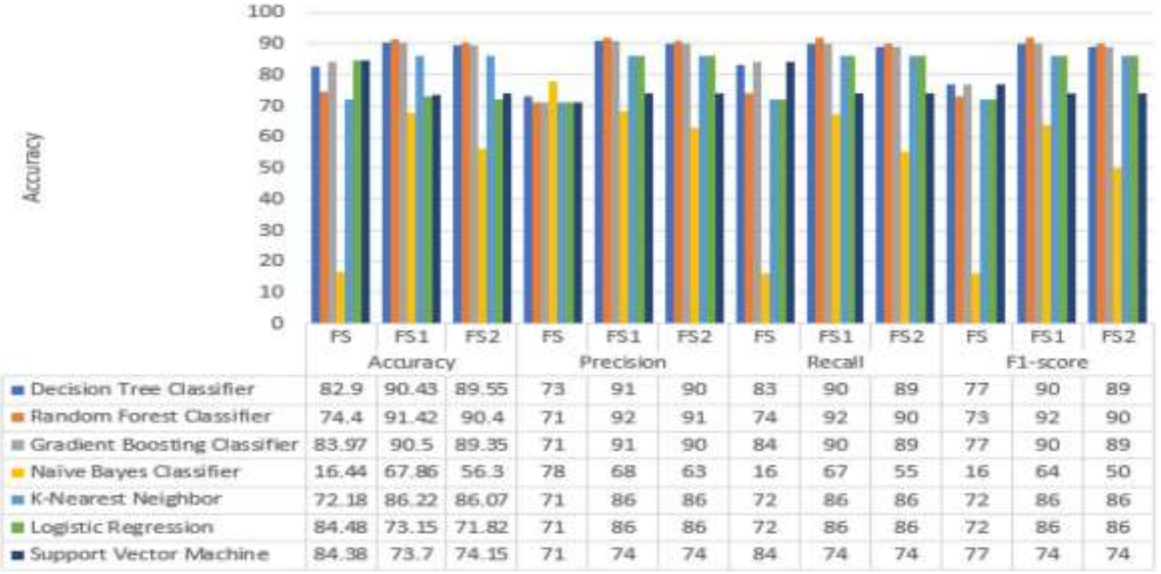
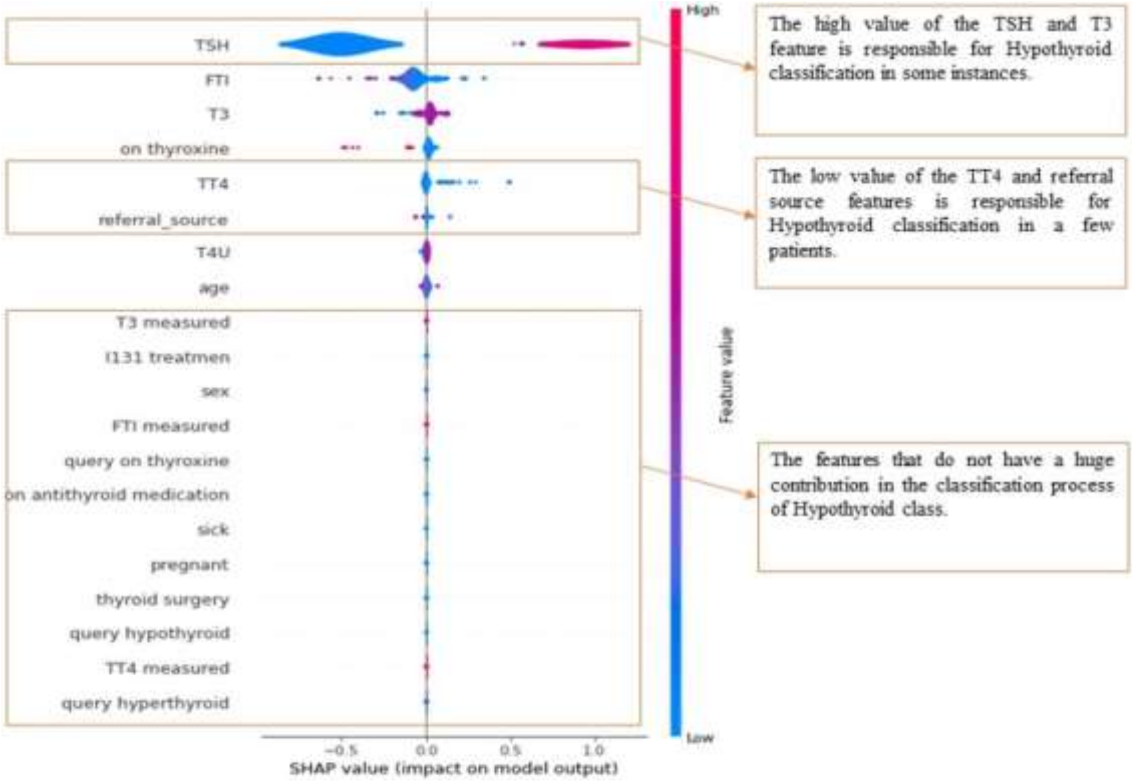


Fig. 7 Confusion matrix of different algorithms **a** SVM, **b** DT, **c** RF, **d** GB, **e** Naive Bayes, **f** KNN and **g** LR using the univariate feature selection method



	Accuracy			Precision			Recall			F1-score		
	FS	FS1	FS2	FS	FS1	FS2	FS	FS1	FS2	FS	FS1	FS2
■ Decision Tree Classifier	82.9	90.43	89.55	73	91	90	83	90	89	77	90	89
■ Random Forest Classifier	74.4	91.42	90.4	71	92	91	74	92	90	73	92	90
■ Gradient Boosting Classifier	83.97	90.5	89.35	71	91	90	84	90	89	77	90	89
■ Naive Bayes Classifier	16.44	67.86	56.3	78	68	63	16	67	55	16	64	50
■ K-Nearest Neighbor	72.18	86.22	86.07	71	86	86	72	86	86	72	86	86
■ Logistic Regression	84.48	73.15	71.82	71	86	86	72	86	86	72	86	86
■ Support Vector Machine	84.38	73.7	74.15	71	74	74	84	74	74	77	74	74

Figure 8. Comparative analysis ofperformance measures of seven algorithms with three feature sets, where FS represented the data with all features, FS1 rep- resented the dataset generated using feature selection method and finally FS2 represented the dataset generated using univariate feature selection method.



The high value of the TSH and T3 feature is responsible for Hypothyroid classification in some instances.

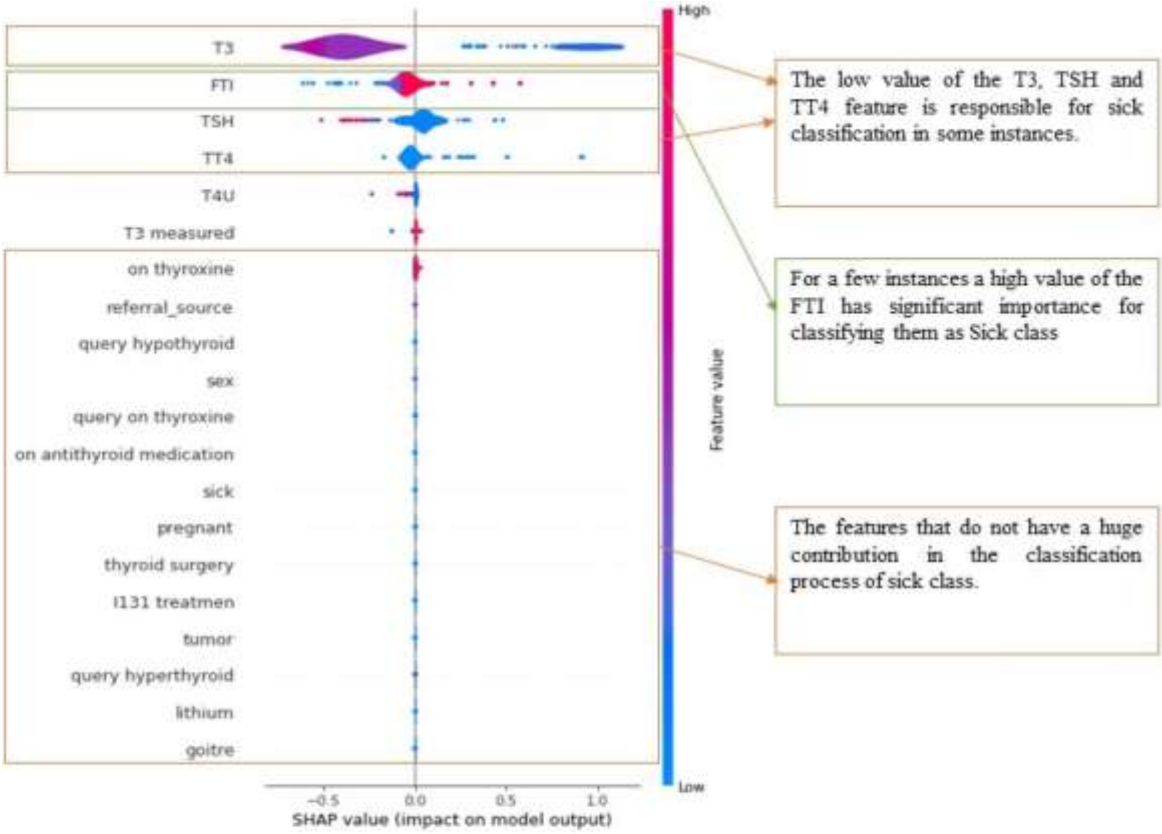
The low value of the TT4 and referral source features is responsible for Hypothyroid classification in a few patients.

The features that do not have a huge contribution in the classification process of Hypothyroid class.

Fig. 9 Violin summary plot using SHAP for the hypothyroid class

4.3.3 Results for Our Dataset Using Univariate Feature Selection Method

In this case, we use the univariate feature selection method to select our important features. The top 14 features with their correlated score with our target are given in Fig. 9. We apply the Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, Naive Bayes Classifier, Logistic Regression Classifier, K-Nearest Neighbor, and Support Vector to the selected features. We can observe from Table 5 that our results are slightly different from previous results. Random Forest provides the best accuracy of 90.4 percent this time as well. After Random Forest, Decision Tree Classifier and Gradient Boosting Classifier have higher accuracy. Decision Tree Classifier and Gradient Boosting Classifier both have an accuracy of 89.55 percent and 89.35 percent, respectively. K Neighbors has an accuracy rate of 86.07 percent. The accuracy of SVM in- creased to 74.5 per- cent, whereas that of Logistic Regression decreased to 71.82 percent. Besides, the accuracy of Naive Bayes fluctuates a lot for this dataset. As a result, we conclude that this method is ineffective compared to the feature importance technique. Other performance metrics of all seven algorithms on this dataset are also presented in Table 5 Table 5 shows that the performance metrics differ signifi- cantly from the previous test result. Logistic Regression, K Neighbors, and Support Vector Machine all have the same precision. The precision of the Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Naive Bayes Classifier, on the other hand, decreases. K-Neighbors, SVM, and Logistic Regression all have the same recall. On the other hand, the recall of the Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Naive Bayes Classifier falls. F1-Score provides a comprehensive view of precision and recall simultaneously, as shown by the fact that the F1-Score is the same for Logistic Regression, K Neighbors, and SVM. The F1 Score of Naive Bayes decreases. So, based on the table above, we can conclude that Random Forest is the bestperformer. After that, the Decision Tree Classifier performs admirably. Gradient Boosting Classifier and Decision Tree Classifier are nearly equal in this race, but Decision Tree Classifier outperforms Gradient Boosting Classifier by a small margin. However, Naive Bayes reduces performance across the board. The confusion matrix of all the seven clas sification algorithms is shown in Fig. 10.Hence confusion matrix that Random Forest provides the best prognosis. In this case, Nave Bayes gives us the worst prediction. Overall results with all classifiers and features in this investigation are depicted in Fig. 11.



The low value of the T3, TSH and TT4 feature is responsible for sick classification in some instances.

For a few instances a high value of the FTI has significant importance for classifying them as Sick class

The features that do not have a huge contribution in the classification process of sick class.

Fig. 10 Violin summary plot using SHAP for the sick class

4.5 Discussion

As machine learning is using almost all aspect of data analysis. So practical implementation of machine learning model for medical data analysis especially for thyroid dis- ease detection may save huge amount of expert physician requirements in this field. However medical data is very sensitive therefore perfect model is the basic requirements for medical data analysis.

Selecting some appropriate feature data points as well as effective machine learning algorithm may pave the way for healthcare to automatic thyroid detection. In this study, we did feature engineering method to identify the best machine learning classification algorithm depending on the data feature used for thyroid detection. Furthermore, we validated our identified best performing model as well as the features which influenced most for the classification by XAI. It is clear from the performance of different algorithms that each algorithm per- formed better depending on whether a subset of features or full features were used. Depending on the situation, each algorithm has the inherent ability to outperform others. RF, for example, outperforms all other algorithms in terms of accuracy of 91.42% and 90.4% respectively in our dataset for the case of FS1 and FS2.

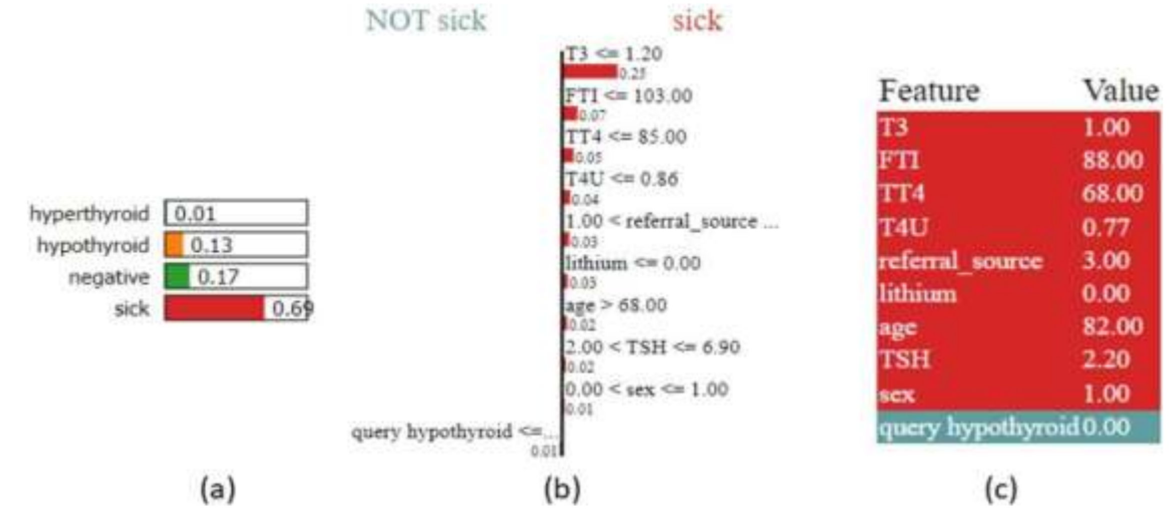


Fig. 11 LIME explanation for a patient **a** category, **b** explanation, **c** history, where actual class: sick, predicted class: sick

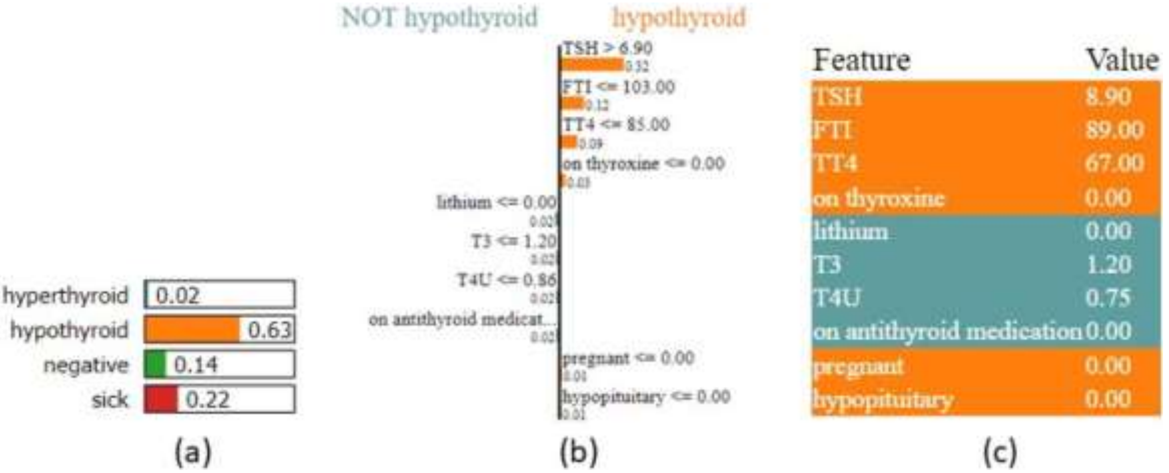


Fig. 12 LIME explanation for a patient **a** category, **b** explanation, **c** history, where actual class: hypothyroid

SVM performs better for small data sets, and ensemble- type classifiers like Random Forest perform better for large data sets. Missing values play a significant role in decision trees. Even after imputing, it cannot produce the same results as a perfect dataset. However, for our case, DT also performed very well of accuracy 90.43% in FS1. Another good classifier was Gaussian Naive Bayes. However, it did not perform well with our dataset. The presumption that all attributes were independent was the reason for this. Results and Analysis would have been less accurate if there was a dependency between the features in the dataset. The accuracy of the K-Nearest Neighbor increases as the number of K we choose increases. It ensures that the given point and the dataset are similar.

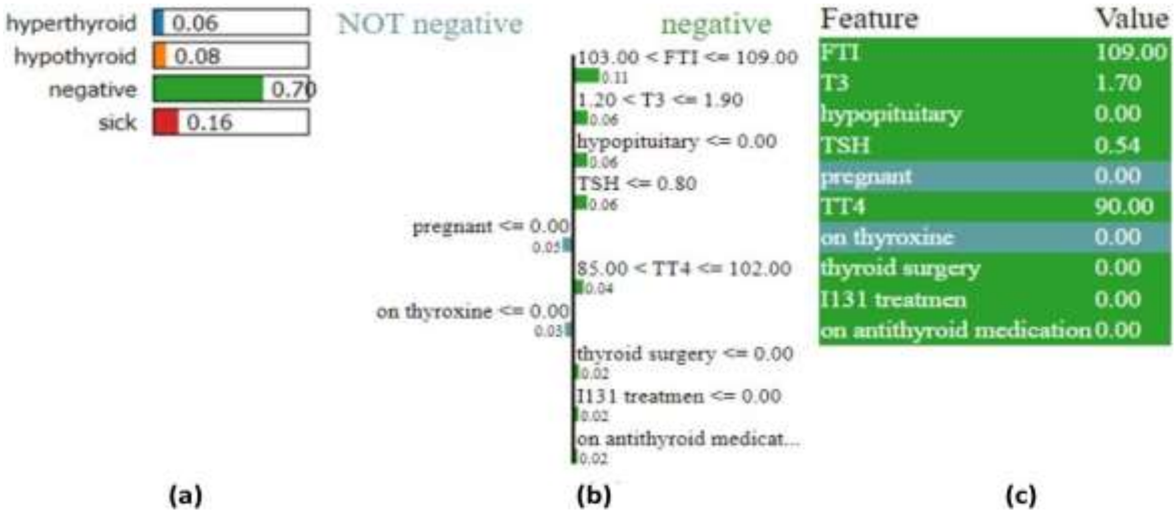


Fig. 13 LIME explanation for a patient a category, b explanation, c history, where actual class: negative, predicted class: negative

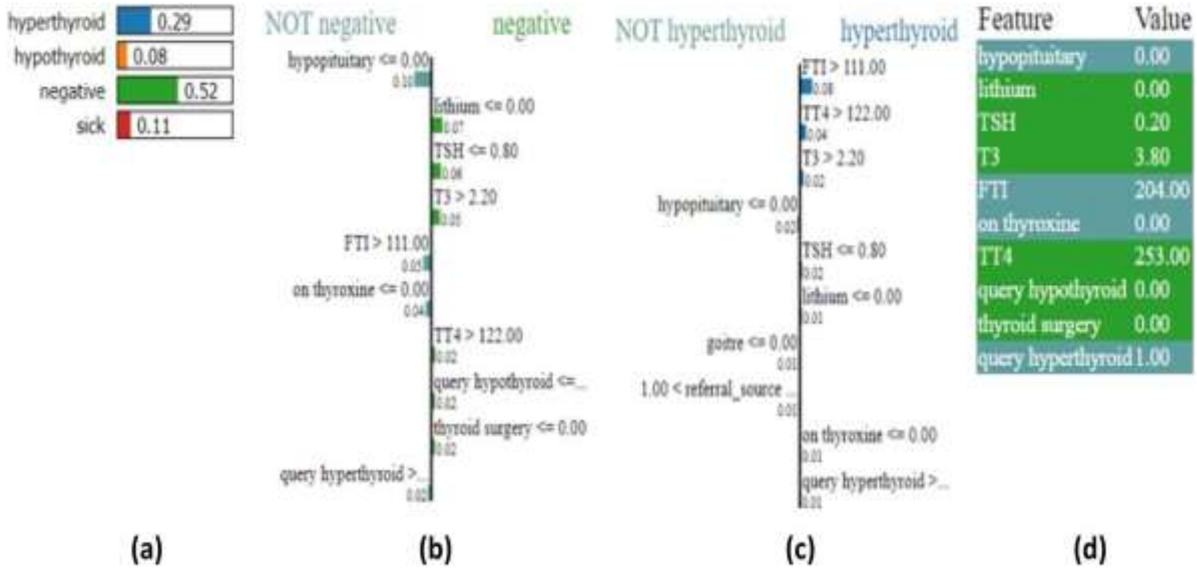


Fig. 14 LIME explanation for a patient **a** category, **b** explanation, **c** history, where actual class: hyperthyroid, predicted class: negative

The performance of algorithms that use all the dataset's features FS was poor relative to FS1 and FS2 for most of the algorithms. After reducing the attribute in the dataset, most of the cases machine learning algorithm performance improved. When there are many attributes, classifier algorithms become complicated, and prediction results vary. Because this is the standard process of evaluating algorithms, performance metrics after converting categorical values, balancing our dataset, and feature selection are used for dataset comparison. Therefore, by considering all of the situations and the performance of metrics used in this experiments we suggest that the Random Forest algorithm and FS1 features should be used to train the model to predict hypothyroidism and Hyperth roidism more correctly. Furthermore, from the XAI analysis we could also observed that the feature attribute TSH, T3, TT4, FTI and T3 contributed the most to classify hypothyroid and hyper- thyroid can be obtained by using the feature importance method FS1. As we use the limited data points in our study there is still a chance for the biasness of the result especially for the case of using all feature to classify the thyroid. Moreover, we balance the dataset before final classification which might not always produced perfect datapoints. Exact clinical data points might further validate our approach which we are looking for. With this UCI adjusted dataset, several studies [3, 28] have been conducted to determine the best suited machine learning model for thyroid categorization. Some other works[18, 21, 24] also noted for the prediction of thyroid disease using different dataset. As per our concern, in this paper, weapply XAI to verify the critical features from this dataset that led the best fitted model to predict specific classes usingexplainable artificial intelligence and our result is relativelycomparable to the existing work.

5. Conclusion

After reducing the features using the feature importance technique and univariate feature selection technique, we tested our collected dataset on various machine learning classifiers to see which classifier gave us the best accuracy. After analyzing the data, we discovered that Logistic Regression outperforms all other classification algorithms for our dataset. When all features are considered, Logistic Regression yields an accuracy score of 84.48 percent. When we use the feature importance method to narrow down the feature set, the Random Forest Classifier gives an accuracy score of 91.92 percent. The accuracy of the Decision Tree Classifier and the Gradient Boosting Classifier is 90.5 percent and 90.43 percent, respectively. When we use the univariate feature selection technique to narrow down the feature set, Random Forest also gives the highest accuracy score of90.4 percent. The second-best algorithm is the Decision Tree Classifier, which has an accuracy score of 89.55 percent; the third-best algorithm is the Gradient Boosting Classifier, which has an accuracy score of 89.35 percent. From explain- ability analysis, we can conclude that most instances have been classified as hypothyroid on the basis of the features TSH, T3 TT4. We can also have identified that the FTI and T3 test values are important for the hyperthyroid class. So, the feature importance technique is more accurate than the univariate feature selection technique in determining correlated features. Thus, after looking at all of the performance metrics, we decided that the Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier and feature importance technique might be a potential choice for predicting hypothyroidism and hyperthyroidism. Though we got relatively better result by feature engineering however there is still room for the search for more perfect model as well as dataset feature selection scheme to further improvement of the result. Moreover, we try our best to clear about biological term in this study however for any kind of our representational limitations we will improve in future.

References

[1] Biondi B, Kahaly GJ, Robertson RP. Thyroid dysfunction and diabetes mellitus: two closely associated disorders. Endocr Rev. 2019;40(3):789–824.

[2] Alam Khan V, Khan MA, Akhtar S. Thyroid disorders, etiology and prevalence. J Med Sci. 2002;2(2):89–94.

[3] Sonu CE, et al. Thyroid disease classification using machine learning algorithms. J Phys. 2021; 1963:12140.

[4] Yasir Iqbal Mirut SM. Thyroid disease prediction using two tier ensemble classifier. Int J Adv Sci Technol. 2020; 29:4460–71.

[5] Bhaladhare V, Chouragade NB, Balpande D, Bhande A, Ambad RS, Bankar N. Ayurvedic management of hypothyroidism. Nat Volat Essen Oil J. 2021;1440–7.

[6] Knudsen N, Laurberg P, Perrild H, Bulow I, Ovesen L, Jorgensen T. Risk factors for goiter and thyroid nodules. Thyroid. 2002;12(10):879–88.

[7] Garg MK, Mahalle N, Kumar K. Laboratory evaluation of thyroid functions: dilemmas and pitfalls. Princ Pract Thyroid Gland Dis- ord. 2017. https://doi.org/10.5005/jp/books/13094_4.

[8] Feller M, Snel M, Moutzouri E, Bauer DC, de Montmollin M, Aujesky D, Ford I, Gussekloo J, Kearney PM, Mooijaart S, et al. Association of thyroid hormone therapy with quality of life and thyroid-related symptoms in patients with subclinical hypothyroidism: a systematic review and meta-analysis. JAMA. 2018;320(13):1349–59.

[9] Unuane D, Velkeniers B. Impact of thyroid disease on fertility and assisted conception. Best Pract Res Clin Endocrinol Metab. 2020;34(4): 101378.

[10] Abbas S. To determine the frequency of undiagnosed hyperthy- roidism in patients presenting with generalized anxiety disorder. J Evol Med Dent Sci. 2013;2(8):930–8.

[11] Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255–60.
[12] Choudhary R, Gianey HK 2017 Comprehensive review on super- vised machine learning algorithms. In: 2017 International conference on machine learning and data science (MLDS), pp. 37–43. IEEE
[13] Crisci C, Ghattas B, Perera G. A review of supervised machine learning algorithms and their applications to ecological data. Ecol Model. 2012; 240:113–22.
[14] Osisanwo F, Akinsola J, Awodele O, Hinmikaiye J, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol (IJCTT). 2017;48(3):128–38.
[15] Praveena M, Jaiganesh V. A literature review on supervised machine learning algorithms and boosting process. Int J Comput Appl. 2017;169(8):32–5.
[16] Singh A, Thakur N, Sharma A 2016 A review of supervised machine learning algorithms. In: 2016 3rd international conference on computing for sustainable global development (INDI- ACom), pp. 1310–1315. IEEE
[17] Tyagi A, Mehra R, Saxena A 2018 Interactive thyroid disease prediction system using machine learning technique. In: 2018 fifth international conference on parallel, distributed and grid computing (PDGC), pp. 689–693. IEEE
[18] Godara S, Kumar S. Prediction of thyroid disease using machine learning techniques. Int J Electron Eng. 2018;10(2):787–93.
[19] Aswathi A, Antony A 2018 An intelligent system for thyroid dis- ease classification and diagnosis. In: 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1261–1264. IEEE
[20] Geetha K, Baboo SS. An empirical model for thyroid disease classification using evolutionary multivariate Bayesian prediction method. Global J Comput Sci Technol. 2016; 16:1–9.
[21] Kousarrizi, MRN, Seiti F, Teshnehlab M. An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification. IJECS. 2012; 12:13–9.
[22] Chandel K, Kunwar V, Sabitha S, Choudhury T, Mukherjee S. A comparative study on thyroid disease detection using k-nearest neighbor and naive bayes classification techniques. CSI Trans ICT. 2016;4(2):313–9.
[23] Singh N, Jindal A. A segmentation method and comparison of classification methods for thyroid ultrasound images. Int J Comput Appli. 2012;50(11):43–9.
[24] Begum A, Parkavi A (2019) Prediction of thyroid disease using data mining techniques. In: 2019 5th international conference on advanced computing & communication systems (ICACCS), pp. 342–345). IEEE
[25] Almahshi HM, Almasri EA, Alquran H, Mustafa WA, Alkhayyat A 2022: Hypothyroidism prediction and detection using machine learning. In: 2022 5th international conference on engineering technology and its applications (IICETA), pp. 159–163 (2022). IEEE
[26] Chaganti R, Rustam F, De La Torre Dıez I, Mazon JLV, Rodrıguez CL, Ashraf I. Thyroid disease prediction using selective features and machine learning techniques. Cancers. 2022;14(16):3914.
[27] 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. https://doi. org/10.1155/2022/9809932.
[28] Pawar U, O'Shea D, Rea S, O'Reilly R 2020 Incorporating explainable artificial intelligence (xai) to aid the understanding of machine learning in the healthcare domain. In: AICS, pp. 169–180
[29] Arjaria SK, Rathore AS, Chaubey G. Developing an explainable machine learning-based thyroid disease prediction model. Int J Bus Anal (IJBAN). 2022;9(3):1–18.
[30] Dua, D., Graff, C.: Uci machine learning repository [https://archi ve.ics.uci.edu/ml]. irvine, ca: University of California, school of information and computer science. IEEE transactions on pattern analysis and machine intelligence (2019)
[31] Kumar A, Tyagi AK, Tyagi SK. Data mining: various issues and challenges for future a short discussion on data mining issues for future work. Int J Emerg Technol Adv Eng. 2014;4(1):1.
[32] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3:1157–82.
[33] Jovi´CA, Brki´CK, Bogunovi´CN 2015 A review of feature selection methods with applications. In: 2015 38th international convention on information and communication technology, electron- ics and microelectronics (MIPRO), pp. 1200–1205. IEEE
[34] Cui S, Tseng H-H, Pakela J, Ten Haken RK, El Naqa I. Introduction to machine and deep learning for medical physicists. Med Phys. 2020;47(5):127–47.

[35] Juba B, Le HS 2019 Precision-recall versus accuracy and the role of large data sets. In: proceedings of the AAAI conference on artificial intelligence, vol. 33, pp. 4039–4048.

[36] Junker M, Hoch R, Dengel A 1999 On the evaluation of document analysis components by recall, precision, and accuracy. In: proceedings of the fifth international conference on document analy- sis and recognition. ICDAR’99 (Cat. No. PR00318), pp. 713–716. IEEE

[37] Powers DM 2020 Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061

[38] Janizek JD, Celik S, Lee S-I. Explainable machine learning pre- diction of synergistic drug combinations for precision cancer medicine. BioRxiv. 2018;331769.

[39] Khaled al Bashiti M, Naser M. Verifying domain knowledge and theories on fire-induced spalling of concrete through explainable artificial intelligence. Constr Building Mater. 2022; 348:128648.

[40] Gramegna A, Giudici P. Shap and lime: an evaluation of discriminative power in credit risk. Front Artificial intelligence. 2021; 4:752558. https://doi.org/10.3389/frai.2021.752558

[41] Javed U, Ijaz K, Jawad M, Ansari EA, Shabbir N, Kutt L, Husev O. Exploratory data analysis based short-term electrical load fore- casting: a comprehensive analysis. Energies. 2021;14(17):5510.

[42] Milo T, Somech A 2020 Automating exploratory data analysis via machine learning: An overview. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pp. 2617–2622