Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 15 No. 3
jtec.utem.edu.my

Prediction of Thyroid Disease using Machine Learning Approaches and Featurewiz Selection

Shiuh Tong Lim[1], Khai Wah Khaw[1], XinYing Chew[2] and Wai Chung Yeong[3]
[1]*School of Management, Universiti Sains Malaysia, 11800 Penang, Malaysia.*
[2]*School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia.*
[3]*School of Mathematical Sciences, Sunway University, Petaling Jaya, Malaysia.*

khaiwah@usm.my

Article Info

Article history:
Received Mar 10th, 2023
Revised Apr 27th, 2023
Accepted June 8th, 2023

Index Terms:
Thyroid Disease
Decision Tree
K-Nearest Neighbor
Logistic Regression
Naïve Bayes
Support Vector Classifier
Random Forest
Extreme Gradient Boost
Featurewiz

Abstract

Thyroid disease is one of the most disturbing hormonal disorders faced by the global population. To help the healthcare industry to diagnose the disorder rapidly and accurately, supervised machine learning algorithms and feature selection were introduced to play an essential role in predicting whether a patient has developed thyroid disease from his/her various characteristics. Therefore, in this work, a new feature selection library was introduced, which was the Featurewiz in the Python library. The goals were to present the performance of the Featurewiz library and to decide on a remarkable model for thyroid disease prediction among several machine learning models, such as Decision Tree, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Classifier, and ensembled machine learning algorithms (Random Forest and Extreme Gradient Boost). A data set consisting of records of thyroid patients in Australia was used to develop the machine-learning models. After the data set was cleaned, exploratory data analysis was carried out. The models were then built in two ways: without feature selection and with feature selection. The feature selection process was conducted by using a new Python library called Featurewiz. The performances of the models from the two operations were evaluated using three performance metrics, including accuracy, F1-score, and AUC (Area Under Curve) value from ROC (Receiver Operating Characteristics Curve). From the two operations, the results are similar in the way that tree-based models, especially those formed by the ensemble method, outperform the statistical models. Initially, in the process without feature selection, the champion model is XGBoost with 99.23% accuracy, while Random Forest ranks second with 98.79% accuracy. However, after the feature selection, the result reveals that the champion model is Random Forest. This model achieves an improvement of 0.66% in accuracy (99.45%), making it the best model from both operations. The model also scores 0.99 and 0.97 in F1-score and AUC values, respectively. The valuable insights gained from this study can serve as a comprehensive framework for machine learning applications in predicting thyroid illness. Additionally, the study highlights the advantageous utilization of the Python feature selection library, Featurewiz. With the combination of Featurewiz and machine learning applications, medical authorities can save time and reduce the risk of misdiagnosis when identifying patients with thyroid disease.

I. INTRODUCTION

The thyroid is a crucial gland that resembles a butterfly in shape. In addition to having a pivotal role in basal metabolic rate (BMR), it plays a vital role in the control of calcium metabolism and stimulates physical and psychological growth [1]. The hormones released by the thyroid are decisive in protein management as well as energy transport and transmission in numerous regions of the body. They are released in response to body temperature [2]. When thyroid hormone is needed, the pituitary gland releases Thyrotropin-Stimulating Hormone (TSH), which travels through the circulation to reach the thyroid gland. TSH then induces the synthesis of T4 and T3 hormones by the thyroid glands [3]. Those functions are stimulated by the main hormones, which are released by the thyroid gland into the bloodstream: thyroxine (T4) and triiodothyronine (T3).

Thyroid disorders and diseases are common hormonal issues that impact most of the world's population. The first indirect references to the thyroid gland, which is associated primarily with disorders, could well be found in Egyptian, Chinese, and later Greek medical books dating back to 2700 BC [4]. Iodine is considered as the fundamental building block of the thyroid glands for the two thyroid hormones, T3 and T4, and is prostrated in a few particular diseases, some of which are extremely common. The two most common manifestations of thyroid disease are hyperthyroidism and hypothyroidism. Hyperthyroidism is caused by an overreaction of the thyroid gland, releasing too much thyroid hormone into the bloodstream, while hypothyroidism is caused by the opposite [5]. Early and accurate diagnose of these diseases requires measuring the T4, T3, and TSH hormone levels [6].

However, diagnosing thyroid disease is a complex and time-consuming procedure that needs much expertise and

information. The typical method of diagnosing thyroid illness involves a clinical examination and a series of blood testing. Fortunately, computational biology is advanced enough for the healthcare industry to help collect stored patient data for medical disease prediction. In addition, with the help of data mining applications, the extensive data collected from healthcare organizations can be transformed into meaningful information and knowledge with high organizational value. Through various data mining methods, hidden relationships and trends in medical data can be efficiently discovered at a lower cost, which subsequently increase profit, and maintain patient's high-quality healthcare [3]. These patterns are useful in predictive modeling.

One of the most critical applications in data mining is classification algorithms, which help make decisions and predictions in a wide range of real-world problems, including diagnosing diseases in the healthcare industry. Moreover, supervised machine learning (ML) algorithms have shown a promising performance, surpassed traditional illness diagnostic methods, and assisted medical professionals in the early detection of high-risk diseases. Nonetheless, to make a precise disease prediction model, the features selected from the different datasets to be applied as a classification of a healthy patient should always be prioritized. Otherwise, misclassification could lead to unneeded treatment for healthy patients. In cases like thyroid disease prediction models, they often involve many potential predictors, such as demographic information, medical history, and laboratory test results. However, not all these predictors are equally important, and the inclusion of redundant or irrelevant predictors can lead to overfitting, decreased model performance, and increased computational complexity. Feature selection methods, by contrast, can help to identify and focus on the most important predictors, eliminating those that are less important. This can lead to a more parsimonious and accurate model, with improved performance and better generalizability to new data. Additionally, this method can help to identify potential causal factors or biomarkers for thyroid disease, which can be useful for further research and clinical practice. In short, the primary key to predicting any disease in association with thyroid disease is paramount [2].

An increased number of papers have been published to introduce machine learning algorithms in predicting thyroid disease in recent years. However, the results of different papers can vary widely without any significant pattern. Therefore, in this work, we selected the ML algorithms that have achieved outstanding performance across various papers in the past few years. By comparing these algorithms, we aim to gain a deeper understanding of which ML method excels in this specific application. Besides that, the ensembled ML algorithms (bagging and boosting) were also included. It is important to note that some of the papers do not include feature selection when building the prediction model. This can lead to errors in clinical decision-making, which can cause health hazards and exorbitant medical costs. This paper seeks to determine the influences of feature selection and obtain significant features by introducing a new Python library, named Featurewiz. This library was introduced by Soham Das in 2019. Featurewiz is believed to be able to select vital features from abundant variables by considering the feature importance and permutation importance. Additionally, it can distinguish whether the problem is regression or classification. More notably, Featurewiz makes the feature selection process easier to comprehend and time-saving [7]. Meanwhile, this paper also proposes the optimal machine-learning model for thyroid disease prediction. This framework can also be applied in other disease prediction, such as cardiovascular disease, cancer, Alzheimer's, and infectious diseases.

II. RELATED WORK

In recent years, papers have been published to introduce machine learning algorithms in the prediction of thyroid disease, such as Naïve Bayes, Decision Tree (DT), Multilayer Perceptron and RBF Network in 2016 [8]. According to Ioniță *et al*., the best classification model was the DT model, with an accuracy of 97.35% after removing some insignificant model attributes. In 2018, a paper prepared by R. Pal *et al.* revealed that most of the studies applied DT, Artificial Neural Networks (ANN) and other techniques for thyroid disease prediction. Considering less work implemented Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN), the study adopted these three models and they found that KNN was superior after dimension reduction with 96.90% accuracy [6]. Nonetheless, in the same year, another paper showed that among ANN, KNN, SVM, and DT, SVM outperformed KNN as the optimal model with 99.63% accuracy, while DT was, surprisingly, the worst with only 75.76% accuracy. However, the operations did not include feature selection [2].

In 2021, P. Duggal and S. Shukla presented a paper showcasing the efficiency of the Recursive Feature Elimination (RFE) method for feature selection. They found that using this technique, the SVM outperformed both Random Forest and Naïve Bayes, achieving the highest accuracy of 92.92%. However, a subsequent study in 2022 showed that the Naive Bayes classifier could achieve a perfect 100% accurate rate in predicting thyroid disease, regardless of whether feature selection was applied. This later study also revealed that other classifiers saw enhanced performance when utilizing L_1-based feature selection [9].

Other than the previously mentioned classifier such as DT, KNN and SVM, the year 2022 saw the emergence of Random Forest as a formidable model for thyroid disease prediction. A study by Islam S. *et al.* introduced Random Forest as a champion mode, achieving an accuracy rate of 98.93% [10]. This finding was further supported by a research conducted by Alyas T. *et al.*, where the Random Forest model emerged as the top performer, registering a 94.8% accuracy rate, outdoing both DT and KNN [11]. Recent development have also brought Extreme Gradient Boost (XGBoost) into the limelight. A comparative study involving DT, Logistic Regression, and KNN showed that XGBoost outperformed all other models in thyroid disease prediction, with an accuracy rate of 98.59% [12].

Drawn from the studies mentioned above, it can be concluded that the utilization of feature selecton processes tends to enhance the accuracy of prediction models. Among the various supervised machine learning algorithms, DT, KNN, SVM, Random Forest, and XGBoost have consistently shown high accuracy in thyroid disease prediction. However, the specific result may vary significantly when these models are applied to different datasets, particularly those involving additional features.

Table 1
Summary of the Results of Some Previous Works from 2016 to 2022

Author's	Feature Selection	ML Algorithms	Accuracy (%)
Ionită *et al*. (2016)	KNIME	- DT	- 97.35
		- Multilayer Perceptron	- 94.71
		- RBF Network	- 94.27
		- Naïve Bayes	- 89.96
R. Pal *et al* (2018*)*	Dimensional Reduction	- KNN	- 96.90
		- Naïve Bayes	- 94.78
		- SVM	- 92.78
A. Tyagi *et al*. (2018)		- SVM	- 99.63
		- KNN	- 98.62
		- ANN	- 97.50
		- DT	- 75.76
P. Duggal *et al*. (2021)	L1- and L2-Based Feature Selection	- Naïve Bayes	- 100
		- Logistics Regression	- 100
		- KNN	- 97.84
		- SVM	- 86.02
		- DT	- 76.92
Islam S. *et al* (2022)	Attribute Subset Selection Module	- XGBoost	- 95.33
		- Random Forest	- 94.79
		- Decision Tree	- 94.32
		- SVC	- 89.59
		- KNN	- 86.21
Sankar S. *et al*. (2022)		- XGBoost	- 98.59
		- KNN	- 96.88
		- DT	- 87.5
		- Logistics Regression	- 81.25

III. DATA EXPLORATION

Since the topic of this project is thyroid disease, a data set containing different thyroid disease patient records was used. The data was collected and supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia. The data set consisted of various features, such as demographic data, the characteristics, the mediations and treatments, different types of hormone indexes, referral sources, as well as the classes of each thyroid patient.

Table 2
Features Description

Variables	Description	Value Type
Class	Target Variable: Whether the patient is healthy (free from thyroid) or is sick	Healthy, Sick
Gender	Demographic Varaible: Male or Female	M, F
Age	Demographic Variable: In years	Continuous
Goitre	Categorical Variable: Does the patient have a goitre	No, Yes
Hypopituitary	Categorical Variable: Does the patient have hypopituitary	No, Yes
Tumour	Categorical Variable: Does the patient have a tumour	No, Yes
Pregnant	Categorical Variable: Is the patient pregnant	No, Yes
Psych	Categorical Variable: Does the patient have a mental illness	No, Yes
Sick	Categorical Variable: Is the patient ill	No, Yes
Query Hyperthyroid	Categorical Variable: Does the patient have any inquiry on hyperthyroid	No, Yes
Query Hypothyroid	Categorical Variable: Does the patient have any inquiries on hypothyroid	No, Yes
Query on Thyroxine	Categorical Variable: Does the patient have any inquiries on thyroxine	No, Yes
I131 Treatment	Categorical Variable: Is the patient undergoing I131 treatment	No, Yes
Lithium	Categorical Variable: Is the patient having lithium medication	No, Yes
On Antithyroid Medication	Categorical Variable: Is the patient undergoing antithyroid treatment	No, Yes
On Thyroxine	Categorical Variable: Is the patient having thyroxine medication	No, Yes
FTI	Continuous Variable: Free Thyroxine Index	Continuous
T3	Continuous Variable: Triiodothyronine	Continuous
T4U	Continuous Variable: Thyroid Utilization Hormone	Continuous
TSH	Continuous Variable: Thyroid-Stimulating Hormone	Continuous
TT4	Continuous Variable: Total Thyroxin	Continuous
Referral Source	Categorical Variable: The sources that are referred to	WEST, STMW, SVHC, SVI, SVHD, Other

IV. METHODOLOGY

The thyroid disease data set used in this work was retrieved from the Kaggle website. The framework of this paper was as shown in Figure 1.

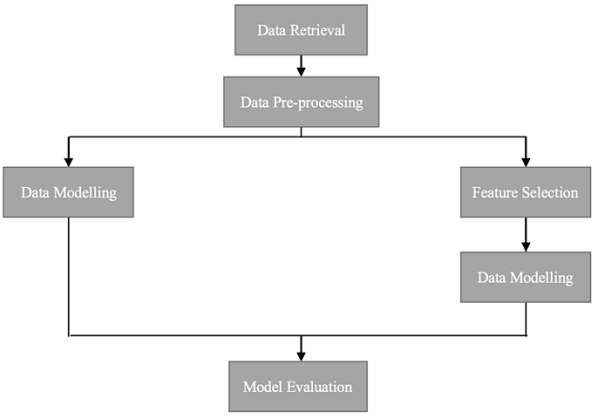

Figure 1. Framework

A. Data Pre-processing

During this stage, several operations were carried out, namely the removal of unnecessary columns, handling of missing values, and encoding of categorical variables. When encoding the categorical variables, to be read by the machine learning algorithms, all the text had to be converted into number form ('0' and '1'). Moreover, dummy variables were created to separate the categories in 'referral_source' into different features.

B. Data Modelling

Seven machine learning algorithms were run in data modelling in two ways: without feature selection and with feature selection. They are DT, KNN, Logistic Regression, Naïve Bayes, Random Forest, Support Vector Classifier (SVC), and XGBoost.

1) Decision Tree

The decision tree method uses a continuous data-splitting mechanism across predetermined parameters [13]. A decision tree is built by three nodes, that are internal nodes, leaf nodes,

and root nodes. The leaf node denotes the distribution of the class, the internal node denotes the test on an attribute, and the root node denotes the tree with the highest node [2]. The building of a decision tree involves a top-down approach by sorting from the root node to the leaf nodes to categorize the data. It can be applied to both continuous and discrete values; thus, it is well-known to be implemented in decision-making. However, when the data set is vast, the accuracy may be low as the model may experience overfitting [3].

2) K-Nearest Neighbor

The KNN model has always been famous for its 'laziness' as its technique does not require any presumption about the distribution of the data on which it is based. As there is no need for a training period for the KNN algorithm before making a prediction, any new data can be included seamlessly without any significant impact on the accuracy. KNN model is commonly evaluated by implementing the dataset; hence, working with datasets from the real world is advantageous. In addition, KNN also accomplishes well in predictive analysis and pattern recognition, especially in classification problems with discrete values [9]. KNN functions as a classifier using the distance function and the chosen K-value: when the classes are in an even number, K-value will be an odd number [13]. K- value, an indicator of how many neighbors is near, is influential for the prediction/classification [2]. Nonetheless, this process has a high demand for memory space and is time-consuming to get optimal K-value.

3) Logistic Regression

A linearly separable boundary is used in classification employing logistic regression. For logistic regression to function, linearly separable boundaries between samples from various classes must first be identified. The probability of belonging to each class described with respect to the decision boundaries is then gained utilizing the logistic function [14]. Logistic regression can examine risk factors as well as forecast the likelihood of getting a specific disease in several analyses by giving discrete predictions [9]. However, the model may face obstacles when there are many categorical features.

4) Naïve Bayes

Naïve Bayes classifier is a supervised machine learning algorithm that is probabilistic based on Bayes' theorem with the assumption that each pair of features is independent [3]. Because of its ability to be used as a sequential learner, its adaptivity to missing values, and its output's simplicity, it necessitates little storage space.

5) Random Forest

A random forest is an ensemble of decision tree machine-learning models. The model is ensembled by the bagging method. Ad hoc data samples are produced by each tree in the forest. The model determines the best prediction score according to the votes. Furthermore, it locates important components in a dataset and offers a clear indication of the feature's importance [13].

6) Support Vector Classifier

SVM model is built by dividing classes in target with the plotted features in n-dimensional space by producing a straight line known as a hyperplane. The points on the side of the hyperplane are considered as in the same class, while the others that fall in the opposite region are another class [9]. The primary purpose of the model is to decide the optimal line that can maximize the data point distance. In this work, the model used was emphasized as SVC instead of SVM as more options for selecting the penalty and loss functions were required, as well as the data set was extensive. LinearSVC and SVM differ in that SVC is defined in terms of libsvm, whereas LinearSVC is defined in terms of liblinear [13]. This model requires a longer training time.

7) Extreme Gradient Boost

XGBoost is a tree-based ensemble machine learning algorithm that obsesses relatively high predicting power. The initial model of XGBoost, which combines various decision trees in a boosting manner, is the gradient-boosting decision tree. The same gradient boosting principles are used by XGBoost, which also employs the maximum tree depth, learning rate, subsampling ratio, and the number of boosts to reduce overfitting and improve performance. More significantly, XGBoost optimizes the function's goal, the tree's size, and the weights' magnitude, all of which are governed by standard regularization parameters [15].

C. Feature Selection Using Featurewiz

When the data set contained high-dimensional data, not all features necessarily contribute to making accurate decisions or predictions. Some features may enhance the model's performance while others might be irrelevant. Identifying and eliminating these unhelpful features crucial in building an effective predictive model. In this work, after the data was split into independent variables (Xs), and target variables, the latest open-source Python library, named Featurewiz, was applied to run fast feature selection. Featurewiz is an effective and quick technique to identify significant variables from a dataset concerning the target variable. By employing this source, the correlation between the variables can be manipulated.

Furthermore, the Uncorrelated List of Variables (SULOV) method is involved in identifying the highly correlated pair of variables that exceeded the correlation threshold and determined their Mutual Information Score (MIS) to the target variable. Since MIS is a non-parametric scoring method, it applies to all kinds of variables. In this stage, the pairs of correlated variables that have lower MIS will be eliminated [7].

Subsequently, in this stage, only variables with less correlation and high MIS will be considered. The recursive XGBoost method is then used to iteratively select the best features from the remaining variables. First, the remaining variables in the dataset were split into training and validation sets. The validation set is then used to identify the top X features (for example, 10) on the training set for early termination to avoid overfitting. The identical procedure is carried out five more times with a different set of variables. Lastly, all the selected features were compiled and de-duplicated. By omitting the data in smaller datasets created from the entire dataset, it aids in finding the best characteristics in accordance with the target variable [16].

D. Model Evaluation

The last stage was model evaluation. Three performance metrics were carried out to quantify the performance of each classification model. The performance metrics were accuracy, F1-score, and Receiving Operating Characteristic (ROC) curve, which were based on the confusion matrix.

Table 3
Confusion Matrix

		Predicted Values	
		Healthy/Negative (0)	*Sick/Positive (1)*
Actual Values	**Healthy/Negative (0)**	True Negative (TN)	False Positive (FP)
	Sick/Positive (1)	False Negative (FN)	True Positive (TP)

1) Accuracy

How accurately a measurement reflects how close it is to the actual or acceptable value. Out of all predictions given, accuracy counts the number of instances that are correctly categorized, as shown in formula (1). A good model should have high accuracy and a low misclassification rate. However, if the classes in the dataset are imbalanced, accuracy may not be a reliable metric. This is because by only correctly predicting most of the classes, high accuracy and a low misclassification rate can be attained. Therefore, other performance metrics are required to select the champion model.

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

2) F1-Score

F1-score is defined as the harmonic mean of precision and recall. It combines precision (2) and recall (3) into a single number, as shown in formula (4). F1-score delivers the best performance when it reaches the perfect value of 1, whereas the worst is the value of 0. Since F1-score takes both FP and FN into account, it is more applicable than accuracy when it comes to uneven class distribution.

The fraction of correctly foreseen positive predictions among all positive predictions is known as precision. The percentage of accurately anticipated true positives out of all the observations in the actual class is known as recall, also known as sensitivity. Poor precision and recall are also indicators of a low F1 score, as shown in the equation. A high F1 score indicates that the classifier is doing an excellent job of correctly predicting the majority of positive observations, which is the underrepresented class for imbalanced classification issues.

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

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

3) Receiving Operating Characteristic Curve

The ROC curve acts as a probability curve that depicts two parameters at different threshold values: the false positive rate (FPR) on the horizontal axis and the true positive rate (TPR), or known as a recall on the vertical axis. Models that perform well possess ROC curves that sketch toward the top-left corner. From the ROC curve, the Area Under the Curve (AUC), a measurement of a classifier's capacity to distinguish between classes can be obtained. The AUC value is between 0 and 1, where AUC = 1 indicates that the model can accurately identify both the positive and the negative class points. On the contrary, when the AUC value equals 0, the model will give the wrong prediction: all positives will be negative, and vice versa. Meanwhile, the classifier is unable to distinguish between positive and negative class points when AUC is 0.5. Hence, it can be concluded that the higher the AUC value, the greater the prediction ability of the model.

$$FPR = \frac{FP}{TN + FP} \quad (5)$$

V. RESULTS

The results of each machine learning algorithm were revealed by showing accuracy, weighted average F1 scores, and AUC value from the ROC curve.

Table 4
Results of Performance Without Feature Selection Process

Classifier	Confusion Matrix	Accuracy (%)	F1-Score	AUC Value
DT	$\begin{bmatrix} 851 & 5 \\ 6 & 44 \end{bmatrix}$	98.79	0.99	0.94
KNN	$\begin{bmatrix} 845 & 1 \\ 29 & 31 \end{bmatrix}$	96.69	0.96	0.76
Logistic Regression	$\begin{bmatrix} 843 & 3 \\ 21 & 39 \end{bmatrix}$	97.35	0.97	0.82
Naïve Bayes	$\begin{bmatrix} 843 & 3 \\ 53 & 7 \end{bmatrix}$	93.82	0.92	0.56
Random Forest	$\begin{bmatrix} 844 & 2 \\ 9 & 51 \end{bmatrix}$	98.79	0.99	0.92
SVC	$\begin{bmatrix} 843 & 3 \\ 19 & 41 \end{bmatrix}$	97.57	0.97	0.84
XGBoost	$\begin{bmatrix} 843 & 3 \\ 4 & 56 \end{bmatrix}$	99.23	0.99	0.96

As shown in Table 4, the Naïve Bayes Classifier had the lowest accuracy (93.82%), while the models with the highest accuracy were the XGBoost Classifier (99.23%). The Naïve Bayes Classifier scores the lowest in F1-score (0.92) and AUC value (0.56) as well. Conversely, besides accuracy, XGBoost Classifier models also achieved the best results in both F1-score (0.99) and AUC values (0.96), indicating that the model performed almost flawlessly in predicting the classes. Therefore, the champion model was the XGBoost Classifier.

Table 5
Results of Performance with Feature Selection Process

Classifier	Confusion Matrix	Accuracy (%)	F1-Score	AUC Value
DT	$\begin{bmatrix} 850 & 6 \\ 4 & 46 \end{bmatrix}$	98.90	0.99	0.96
KNN	$\begin{bmatrix} 845 & 1 \\ 22 & 38 \end{bmatrix}$	97.46	0.97	0.82
Logistic Regression	$\begin{bmatrix} 843 & 3 \\ 21 & 39 \end{bmatrix}$	97.35	0.97	0.82
Naïve Bayes	$\begin{bmatrix} 830 & 16 \\ 44 & 16 \end{bmatrix}$	93.38	0.92	0.62
Random Forest	$\begin{bmatrix} 844 & 2 \\ 3 & 57 \end{bmatrix}$	99.45	0.99	0.97
SVC	$\begin{bmatrix} 842 & 4 \\ 16 & 44 \end{bmatrix}$	97.79	0.98	0.86
XGBoost	$\begin{bmatrix} 843 & 3 \\ 3 & 57 \end{bmatrix}$	99.34	0.99	0.97

There were 16 significant features left after the feature selection procedure, down from 26, which were 'age', 'hypopituitary', 'psych', 'query_hyperthyroid', 'query_hypothyroid', 'on_thyroxine', 'SVI', 'SVHD', 'SVHC', 'STMW', 'other', 'TSH', 'T4U', 'T3', 'TT4', and 'FTI'. These selected features had a correlation limit of at least 0.8 with the target variable.

The results changed, as shown in Table 5. However, the weakest model was still the Naïve Bayes Classifier. It had the lowest accuracy (93.38%), F1-score (0.92), and AUC value (0.62). The model that performed best was Random Forest which had 99.45% accuracy, 0.99 in F1-score, and 0.97 in AUC value. Therefore, the champion model in this section was the Random Forest model.

VI. DISCUSSION

A. Models Comparison

As shown in Table 4, the classifiers with the highest accuracy for predicting thyroid disease were XGBoost (99.23%), DT (98.79%), and Random Forest (98.79%). Among these, XGBoost's performance was slightly more impressive, reflecting the advantages of boosting over the bagging method employed by Random Forest. Interestingly, DT yielded the same results as Random Forests, probably due to a few outliers in the data set.

Naïve Bayes had the worst performance compared to others, probably to the assumptions of class conditional independence since the variables in this work were correlated, as shown in Figure 2. This model also relies on categorical features, whereas this dataset comprised a mixture of both categorical and continuous variables. The algorithm scored an accuracy of 93.38%, a weighted average F1-score of 0.92 and AUC score at 0.56. Although the weighted average score was considered high, the low AUC value implied an imbalanced performance. As shown in Figure 3, Naïve Bayes predicted very badly for class 1/positive class/patient. It scored only 0.12 in recall indicating that this model produced many false negatives.

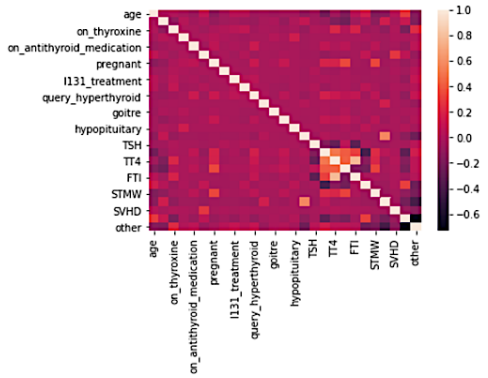

Figure 2. Correlation Heatmap for Features

	precision	recall	f1-score
0	0.94	1.00	0.97
1	0.70	0.12	0.20

Figure 3. F1-Score of Naïve Bayes Without Feature Selection

According to Table 5, the champion model had changed. After the feature selection process, Random Forest became the most accurate model (99.45%), while XGBoost was the second (99.34%). Nevertheless, the difference in the accuracy of the models was not significant. DT had become the third with an accuracy of 98.90%. The outcomes after feature selection made it more evident that the ensemble learning algorithms worked better in this thyroid disease data set. Meanwhile, the model with the lowest accuracy was still Naïve Bayes (93.38%).

From both results, the top three models are always the tree-based model. This was because tree-based ML algorithms did not need pre-processing features like standardization or normalization [13]. The ensembled ML algorithms patently perform better as they can reduce the generalization error. The result was compatible with the paper prepared by Islam S. et al in 2022 [13]. However, the poor performance of Naïve Bayes algorithms was unexpected, considering a paper published in 2021 showed that Naïve Bayes algorithms gained 100% accuracy in thyroid disease prediction [17].

B. Pairwise Comparison

According to Table 6, with the feature selection process, the accuracy of almost every model had enhanced, except Naïve Bayes and Logistic Regression.

Table 6
Accuracy Comparison of Each Model

Classifier	Accuracy (%)		$A_1 - A_0$ (%)
	Without Feature Selection (A_0)	With Feature Selection (A_1)	
DT	98.90	0.99	0.11
KNN	97.46	0.97	0.77
Logistic Regression	97.35	0.97	0
Naïve Bayes	93.38	0.92	-0.44
Random Forest	99.45	0.99	0.66
SVC	97.79	0.98	0.22
XGBoost	99.34	0.99	0.11

The performance of Logistic Regression was not affected by the feature selection process. The feature selection process was found to be advantageous in reducing the complexity by eliminating the irrelevant features and then improving the accuracy. However, it is uncertain that feature selection will necessarily enhance the accuracy of the model if the model can cope with the big and redundant data set. For example, the initial high-dimensional data did not affect the performance of the Logistic Regression algorithm because the algorithm itself already had a high resistance to overfitting. The execution of the Naïve Bayes Classifier is dependent mainly on all attributes being categorical. Hence, the decreasing accuracy in the Naïve Bayes model might be due to the lack of information in class prediction after some features, mostly categorical features, had been taken off from the process since its operation relied on categorical attributes.

On the contrary, KNN and Random Forest exhibited the most prominent improvement of 0.77% and 0.66%, respectively. The KNN algorithm is one of the simplest models that are non-parametric and effective. Nonetheless, it is sensitive to outliers and irrelevant attributes, which will lead to mislabeling of training distance. Therefore, the KNN model can function better and more accurately after removing irrelevant features. Although Random Forest already performed well without feature selection, it still showed a considerable improvement of 0.66% with feature selection. This is because the data set is less redundant after the noise and misleading data are eliminated. Hence, the Random Forest model can focus only on the significant independent variables when forming the nodes.

VII. CONCLUSION

Since thyroid disease is a common disease in the world's population, the use of machine learning in prediction is very beneficial as it will ease the work of the medical authorities to identify which patients are sick. This is crucial as any mistake might lead to not only a waste of money and time but also bring damage and risk to the healthy individual. In this study, a comprehensive approach was taken, employing a total of seven machine learning algorithms, namely Decision Tree, K-Nearest Neighbor, Logistic Regression, Naïve Bayes, Support Vector Classifier, Random Forest, and Extreme Gradient Boost. Since the dataset was redundant, the features that are important for predicting thyroid diseases were determined by utilizing feature selection. Feature selection was carried out by calling a new library in Python, named Featurewiz. It is one of the latest open-source libraries that are capable of identifying the significant features from a data set regard to the target variable. By implementing featurewiz, the important features with a correlation limit of 0.80 were revealed. Out of the initial 26 features, only 16 features were selected as significant: the demographic variable selected was 'age'; the categorical variables selected were 'hypopituitary', 'psych', 'query_hyperthyroid', 'query_hypothyroid', 'on_thyroxine', 'SVI', 'SVHD', 'SVHC', 'STMW', and 'other'; while the continuous features selected were 'TSH', 'T4U', 'T3', 'TT4', and 'FTI'.

This paper presents the results of a study on the performance of the newly developed Python library, Featurewiz. Our analysis reveals that Featurewiz is not only a speedy and efficient tool for feature selection but also offers the flexibility of setting desired correlation limits. The impressive accuracy gains observed after using Featurewiz for feature selection demonstrate its usefulness as a tool for predictive modeling. This is because when Featurewiz is applied, the selection of vital features becomes easier and less complex. Moreover, Featurewiz requires minimal training time and can be implemented using simple code, making it a widely accessible resource for researchers and practitioners alike. Therefore, the combination of Featurewiz with ML algorithms can be widely used in disease prediction.

The results of this study showed that Random Forest and XGBoost have a very high accuracy which is higher than 99%. This outcome indicates that implementing an ensemble machine learning model in thyroid disease prediction will give a better and more accurate result, especially after feature selection. However, the different methods of the ensembled ML algorithms do not have a great impact on the prediction of thyroid disease. Overall, the champion model for thyroid disease prediction is the Random Forest algorithm with feature selection. The accuracy of this model is 99.45%, and its weighted average F1-score is 0.99. This output reveals that ensembled ML algorithms are preferable in disease prediction.

In this study, a small dataset was used to compare the accuracy of ML algorithms in thyroid disease detection. Given the specific research question and the available data, we opted to use a small dataset in this study. While larger datasets are often preferred, the quality and diversity of the available data were such that a small dataset was deemed sufficient for our purposes. Moreover, previous research in similar areas has shown that smaller datasets can provide valuable and reliable results for the comparison of ML models [11][13][14]. While the use of a small dataset may limit the generalizability of our findings, we believe that our results provide important insights into the combination of Featurewiz and ML algorithms and that our approach can serve as a useful methodological framework for future research.

The outcomes of this work are satisfying, but future work is needed with a more balanced data set as the data set used is an imbalanced data set with a very contrasting class with a ratio of 1135:75. In the future study, the low imbalance problem of the data set can be handled first before further operations by implementing the resampling method. This can be done using a Python package called imbalanced-learn (imblearn). Besides, other techniques, such as the TOMEK link and SMOTE (Synthetic Minority Oversampling Technique) can also be considered. Next, more features, such as countries and common food intake can be included in the data set to improve the results. Besides, more models can be introduced to conduct thyroid disease prediction. Since, from the result, it was known that ensembled machine learning models were better, future work can focus on models developed by the ensemble method: bagging and boosting. Different deep-learning models can also be included. Besides, future work can include more extensive data collection. The techniques can also be implemented for other diseases, such as chronic disease.

REFERENCES

[1] Y. S. Khan and A. Farhana, Histology, Thyroid Gland. StatPearls Publishing, Treasure Island (FL), 2022.

[2] A. Tyagi, R. Mehra, and A. Saxena, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), pp. 689–693, 2018, Feb. 23, 2023.

[3] B. A. Begum and P. A, "Prediction of Thyroid Disease Using Data Mining Techniques," 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)., Feb. 23, 2023.

[4] A. Niazi, S. Kalra, A. Irfan, and A. Islam, "Thyroidology over the ages," Indian J Endocrinol Metab, vol. 15, no. 6, p. 121, 2011.

[5] Prerana, P. Sehgal, and K. Taneja, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network," 2015.

[6] R. Pal, T. Anand, and S. K. Dubey, "Evaluation and performance analysis of classification techniques for thyroid detection," 2018.

[7] D. David, "Automatic Feature Selection in Python: An Essential Guide," Hackernoon, Jul. 20, 2021.

[8] I. Ionită and L. Ionită, "Prediction of Thyroid Disease Using Data Mining Techniques," BRAIN. Broad Research in Artificial Intelligence and Neuroscience, vol. 7, no. 3, pp. 115–124, Aug. 2016.

[9] H. Abbad Ur Rehman, C. Y. Lin, Z. Mushtaq, and S. F. Su, "Performance Analysis of Machine Learning Algorithms for Thyroid Disease," Arab J Sci Eng, vol. 46, no. 10, pp. 9437–9449, Oct. 2021.

[10] K. Salman and E. Sonuc, "Thyroid Disease Classification Using Machine Learning Algorithms," in Journal of Physics: Conference Series, IOP Publishing Ltd, Jul. 2021.

[11] T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," Biomed Res Int, 2022.

[12] S. Sankar, A. Potti, G. Naga Chandrika, and S. Ramasubbareddy, "Thyroid Disease Prediction Using XGBoost Algorithms," Journal of Mobile Multimedia, vol. 18, no. 3, pp. 917–934, 2022,

[13] S. S. Islam, M. S. Haque, M. S. U. Miah, T. Bin Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," PeerJ Comput Sci, vol. 8, 2022.

[14] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid Disease Prediction Using Machine Learning Approaches," National Academy Science Letters, vol. 44, no. 3, pp. 233–238, Jun. 2021.

[15] K. Davagdorj, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, "Xgboost-based framework for smoking-induced noncommunicable disease prediction," Int J Environ Res Public Health, vol. 17, no. 18, pp. 1–22, Sep. 2020.

[16] Seshadri, Ram, "AutoViML/featurewiz: Use advanced feature engineering strategies and select the best features from your data set

fast with a single line of code", Accessed: 20 Oct 2022, [Online] Available: https://github.com/AutoViML/featurewiz.

[17] P. Duggal and S. Shukla, "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques," 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 670–675, 2020, Feb. 23, 2023.