Performance Analysis of Machine Learning Algorithms for Thyroid Disease

Article *in* Arabian Journal for Science and Engineering · January 2021

DOI: 10.1007/s13369-020-05206-x

Abbad Ur Rehman
University of Guelph
10 PUBLICATIONS **203** CITATIONS
SEE PROFILE

Chyi-Yeu Lin
National Taiwan University of Science and Technology
158 PUBLICATIONS **3,524** CITATIONS
SEE PROFILE

Zohaib Mushtaq
University of Sargodha
55 PUBLICATIONS **1,069** CITATIONS
SEE PROFILE

Shun-Feng Su
National Taiwan University of Science and Technology
306 PUBLICATIONS **8,987** CITATIONS
SEE PROFILE

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-020-05206-x

RESEARCH ARTICLE-ELECTRICAL ENGINEERING

Performance Analysis of Machine Learning Algorithms for Thyroid Disease

Hafiz Abbad Ur Rehman[1] · Chyi-Yeu Lin[1] · Zohaib Mushtaq[2] · Shun-Feng Su[2]

Received: 13 February 2020 / Accepted: 7 December 2020

Abstract

Thyroid disease arises from an anomalous growth of thyroid tissue at the verge of the thyroid gland. Thyroid disorderliness normally ensues when this gland releases abnormal amounts of hormones where hypothyroidism (inactive thyroid gland) and hyperthyroidism (hyperactive thyroid gland) are the two main types of thyroid disorder. This study proposes the use of efficient classifiers by using machine learning algorithms in terms of accuracy and other performance evaluation metrics to detect and diagnose thyroid disease. This research presents an extensive analysis of different classifiers which are K-nearest neighbor (KNN), Naïve Bayes, support vector machine, decision tree and logistic regression implemented with or without feature selection techniques. Thyroid data were taken from DHQ Teaching Hospital, Dera Ghazi Khan, Pakistan. Thyroid dataset was unique and different from other existing studies because it included three additional features which were pulse rate, body mass index and blood pressure. Experiment was based on three iterations; the first iteration of the experiment did not employ feature selection while the second and third were with L_1-, L_2-based feature selection technique. Evaluation and analysis of the experiment have been done which consisted of many factors such as accuracy, precision and receiver operating curve with area under curve. The result indicated that classifiers which involved L_1-based feature selection achieved an overall higher accuracy (Naive Bayes 100%, logistic regression 100% and KNN 97.84%) compared to without feature selection and L_2-based feature selection technique.

Keywords Classification · Thyroid disease · KNN · SVM · DT · NB · LR · Feature selection

List of Symbols

k	Number of neighboring elements
L_1	L_1-norm
L_2	L_2-norm
a, b	Feature vectors
d	Distance

1 Introduction

Thyroid is a significant gland which resembles the shape of butterfly. It is placed in the lower part of the neck and helps to control the body metabolism [1]. This gland produces two active thyroid hormones which are levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3) [2, 3]. These hormones play a vital role in the production of proteins, in the regulation of body temperature, and in overall energy production and regulation [4, 5]. The thyroid gland is prone to many distinct diseases, some of which are especially common such as hypothyroidism and hyperthyroidism [3]. Production of deficient secretion in thyroid hormone causes hypothyroidism, and production of an excessive amount of secretion thyroid hormone causes hyperthyroidism [2, 6]. The former case refers to hypothyroidism condition which deals with deficiency or underproduction of thyroid hormones. The symptoms in this condition may involve a person experiencing weight gain, swelling in front of neck and low pulse rate, whereas hyperthyroidism refers to an excessive amount of thyroid hormone by the thyroid gland in which a person may suffer from elevated blood pressure and pulse rate while having reduced body weight [6, 7]. A commonly used method to identify thyroid disorders is the use of blood test, which can measure the TSH, T3 and T4 levels [8, 9].

✉ Hafiz Abbad Ur Rehman
d10603813@ntust.edu.tw

[1] Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

[2] Department of Electrical Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan

Published online: 23 January 2021

Health care industry produces a large part of complex data in the medical field that is very challenging to manage [5]. A fair amount of machine learning approaches has recently been used to examine and identify different types of diseases. Bayesian network (BN), SVM, neural network, ANN, decision tree (DT), Naive Bayes, K-nearest neighbor (KNN) and many more are the different classification methods used by researchers [9–11]. This literature review will highlight the different machine learning approaches carried out by researches in order to detect thyroid diseases.

The K-nearest neighbor (KNN) is an extremely popular and common machine learning algorithm and currently many techniques are based on achieving an effective KNN to diagnose thyroid disease [12]. A variety of classification methods such as KNN, neural network and Bayesian belief network discussed by Tomar and Agarwal [13] and the fuzzy logic using MATLAB described by Jahantigh [14] often play a vital role in the identification of diseases in the health care sector in order to obtain an appropriate thyroid classification. S. Sun and R. Huang discussed the limitation of KNN algorithm and proposed an adaptive KNN algorithm (AdaNN) for classification and showed that this is superior over traditional KNN algorithm. This is because, for each test case, the AdaNN algorithm finds the suitable *k* value. This determines the optimal value of *k* and takes the few number of neighbors closest to get the right class name [15]. Furthermore, the researcher Liu et al. presented an efficient computer-aided diagnostic (CAD) base system consisted of fuzzy K-nearest neighbor (FKNN) classifier for diagnosis of the thyroid disease. Two core parameters of FKNN, which are *k* value of neighborhood and fuzzy parameter m, are adaptively specified by particle swarm optimization (PSO) approach. The proposed PCA-PSO-FKNN system is then reported to use tenfold cross-validation (CV) with 99.09% accuracy to clearly distinguish and diagnose the different classes of thyroid diseases [16]. Acharya et al. addressed the CEUS-based thyroid nodule classification CAD system which is a contrast-enhanced ultrasound imaging to enhance the differential diagnosis of thyroid nodules as it gives a better representation of thyroid vascular pattern. Furthermore, discrete wavelet transform (DWT) and texture-based features were extracted from thyroid lesions 3D contrast-enhanced ultrasound images. K-nearest neighbor (KNN), probabilistic neural network (PNN) and decision tree (DT) classifiers were then used to test and train these resultant features by using ten cross-fold validation technique and achieved classification accuracy of 98.90% [17]. Researcher Nazari et al. used another approach to detect thyroid disease which was support vector machine classifier (SVM). This research study compared and analyzed two thyroid datasets taken one from UCI and another actual data from Imam Khomeini Hospital. For feature selection, sequential forward selection (SFS), sequential backward selection (SBS) and genetic algorithm (GA) scheme were used. In this case, GA-SVM showed the best classification accuracy of 98.26% among all proposed methods [18]. Moreover, another researcher Chen et al. developed a three-stage system to address thyroid disease. (FS-PSO-SVM) CAD method with particle swarm optimization demonstrated better performance than the existing methods and achieved the accuracy of 98.59 by using tenfold cross-validation (CV) [19]. A generalized discriminant analysis (GDA) and wavelet support vector machine (WSVM) (GDA–WSVM) approach consisted of feature extraction, and feature reduction classification phases were used by Dogantekin et al. for thyroid disease and obtained 91.86% classification accuracy [20]. In the study of fuzzy classifier, an expert system for thyroid disease called ESTDD (expert system for thyroid disease diagnosis) was introduced by two researchers Keleş and Keles. Fuzzy rules were applied on the bases of neuro fuzzy classification (NEFCLASS) algorithm and reported 95.33% accuracy [21]. Using several neural network methods like multilayer perception (MLP) through back-propagation, radial basis function and adaptive conic section function in neural network were proposed by Ozyilmaz et al. for thyroid diagnosis and resulted accuracies were 88.30%, 81.69% and 85.92%, respectively [22]. From the existing literature, it is revealed that classification is the imperative technique for detecting, predicting and diagnosing different diseases like heart disease, breast cancer, lung cancer and thyroid disorder. Figure 1 presents the role of classification techniques for detecting various diseases. The literature review revealed that thyroid disorders have been focused less compared to other diseases [6, 9].

Beside the clinical and essential investigation, proper interpretation of thyroid disease is also important for diagnosing purposes. Authors Chen et al. address the importance of feature selection technique for improving the classification accuracy beneficial for diagnosis purposes [19]. In this paper, effectiveness of different classification method was investigated with the implementation of L_1 and L_2 feature selection technique. Thus, it is hypothesized that new introduced features would provide accurate and precise measures for diagnosing thyroid disease. To carry-out the research,

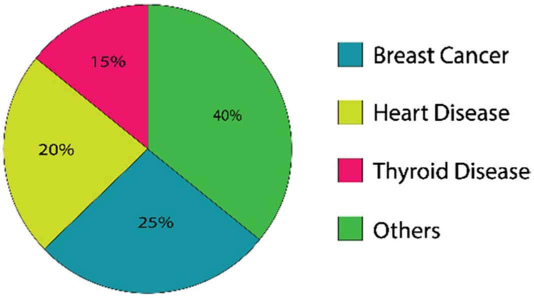


Fig. 1 Health care statistics using classification [6]

unique thyroid dataset was used. The performance of the proposed research is examined by using the confusion matrix, and the obtained results were also compared with existing studies reported in Table 3 focusing on the thyroid diagnosis. The complete paper is formed as follows: Sects. 1 and 2 include a literature review and dataset, respectively. Section 3 details the methodology adopted. Sections 4 and 5 include experimental and result analysis. Section 6 highlights the related existing studies. Section 7 concludes the paper with future scope.

2 Dataset Description

Thyroid disease dataset used in our experiment is taken from District Headquarters (DHQ) Teaching Hospital, Dear Ghazi Khan, Pakistan [23]. This hospital provides health care facilities to not only the inhabitants of the district but also to patients coming from neighboring provinces. The dataset used in this study is fully verified by two endocrinologists associated with well renowned teaching hospital based in Karachi, Pakistan. There are three classes and 309 patient samples in the dataset. Total patients were divided into three categories based on diagnosis results. The categories are as following:

Class (1): A total of 170 individuals with optimal range of hormonal values

Class (2): A total of 66 patients suffering with hyperthyroidism

Class (3): A total of 73 patients suffering with hypothyroidism

Thyroid dataset comprises of 309 entries with ten attributes column and one class column. This dataset has three new features, i.e., body mass index (BMI), measurement of blood pressure and pulse rate which make this dataset unique from others available on UCI and KEEL repository [24]. Thirteen missing values were reported in the T3 column shown in Table 1 and replaced by ‘?.’ Mean values were used as a replacement for missing entries. Three classes are “Hypo” for hypothyroidism, “Hyper” for hyperthyroidism and “Normal” for healthy individuals contributing 24%, 21% and 55% of the total, respectively.

3 Methodology

Methodology reported in this manuscript consists of few important steps as outlined in Fig. 2. Data processing is the initial step of our methodology which involves deletion and cleaning of useless columns or entries. Processing missing values and cleaning unnecessary data can potentially improve accuracy of overall result. Furthermore, processing missing values is very crucial because skipping the values would negatively impact the results as there is a risk of losing valuable information. Following this step, feature scaling based on min–max method is implemented in order to obtain the maximum and minimum entry values. To get an efficient accuracy and performance of the classifiers, first part of the experiment is implemented without feature selection techniques. L_1- and L_2-based feature selection techniques are implemented in the second and third phase of the experiment, respectively. Features such as blood pressure, pulse rate and BMI are included in this study because they directly correlate with thyroid disorders and played a vital role to achieve a best accuracy results. Various evaluation parameters like f_1-score, miss-rate Matthew correlation coefficient (MCC), error-rate, ROC curve with AUC, sensitivity, selectivity, fall-out and accuracy have been used for evaluation and comparison criteria of different classifiers and best algorithms for detecting thyroid disease.

Table 1 Thyroid dataset description

Thyroid dataset		
Attributes #	Features	Range description
1	Serial and hospital reference IDs	ID number
2	Pregnant	Yes, no
3	Body mass index (BMI)	Underweight–optimal–overweight
4	Blood pressure	High–healthy–low
5	Pulse rate	50–110
6	T3	0.15–3.7 (Missing values ‘?’)
7	TSH	0.05–100
8	T4	0.015–30
9	Gender	Male, female
10	Age	6–62
11	Class	0 ‘Hypo’, 1 ‘Hyper’ and 2 ‘Normal’

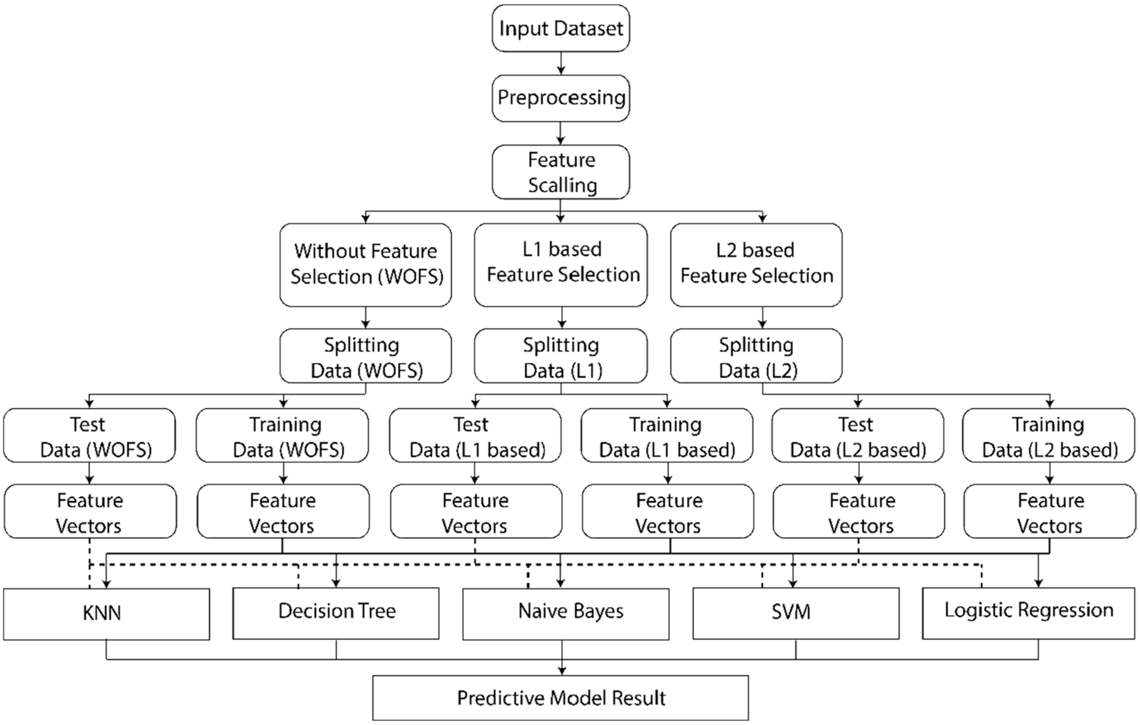


Fig. 2 General block diagram of proposed study

3.1 Feature Selection

Feature selection plays a vital role to increase the efficiency of a given classifier. Modern IoT devices send millions of information which create datasets with hundreds of unwanted features. In resultant, these features choke the model, exponentially increase the training time and increase overfitting risk. By using feature selection technique, a reduced average time for predicting and training can be achieved without loss of total information. Later, these important selected features were then used for training and testing in order to save cost and time. Such techniques play a large role in impacting the classification results [12].

3.1.1 L_1 and L_2 Norm-Based Model Feature Selection

For this report, the L_1- and L_2-based feature selection technique has been used with the help of a Python library known as scikit-learn. Compared to other existing libraries such as mlpy, pybrain and shogun, scikit-learn is a very user-friendly library with a remarkable response time of various algorithms and techniques [25]. These L_1- and L_2-based feature selection approaches can be used with classifiers to achieve dimensionality reduction for given datasets. L_1 feature selection techniques assign zero value to some coefficients. Therefore, due to estimation of target, certain features are removed because they do not contribute to the final prediction. However, in L_2 feature selection technique, the coefficient value is not assigned zero but rather is approached to zero. For this research, the linear support vector classifier (LSVC) was used, and to control the sparsity, a *C* parameter was selected. Upon observation, it can be noted that the value of *C* is directly proportional to the number of features selected; the larger the value of *C*, the more features will be selected and vice versa.

3.2 KNN

The K-nearest neighbors (KNN) is very common and most widely used supervised machine learning algorithm. KNN performs nicely for predictive analysis and pattern recognition purposes. One of the main use of KNN is to predict discrete values in classification problems [26, 27]. KNN uses two factors, namely the similarity measure or distance function and the selected *k* value to act as a classifier with the performance depending on the aforementioned factors. For any new data point, firstly KNN calculates the distance of all the data points and gathers the ones which are in close proximity to it. Then, algorithm organizes those closest data points based on their distance from arrival data point using different distance functions. Furthermore, the next step is to gather specific number of those data points which have the least distance among all and categorize them based on their distance. Figure 3 demonstrates the working principle of KNN. In the figure, the red

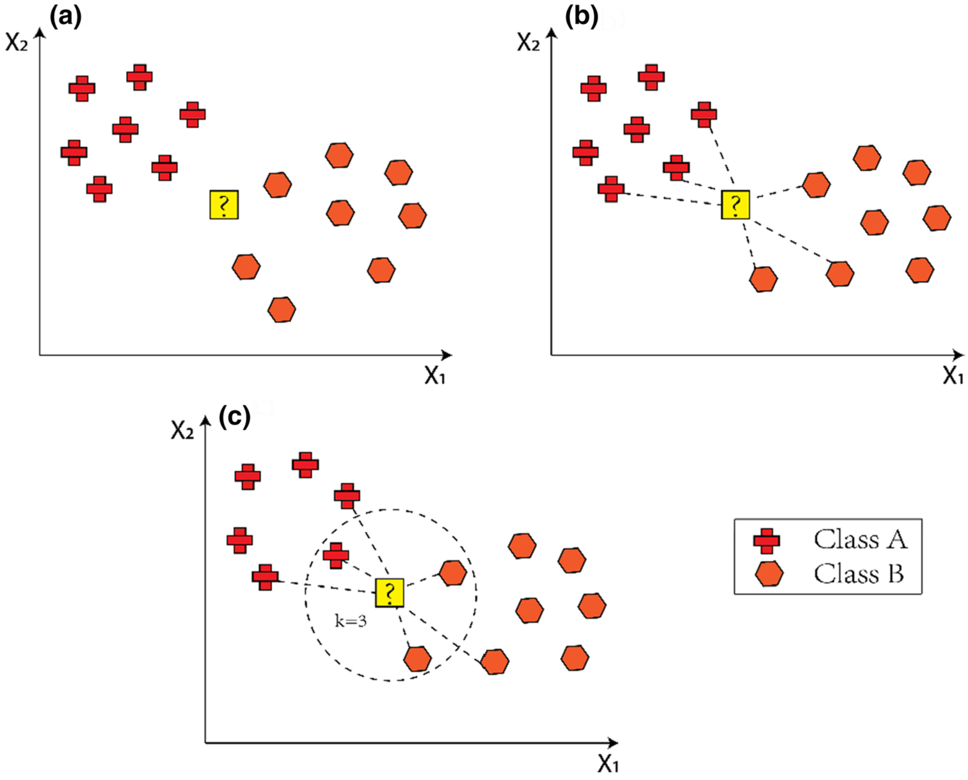


Fig. 3 Working principle of K-nearest neighbor method **a** initial data, **b** calculate distance and **c** find neighbors and vote

plus sign belongs to class 01 whereas the green sign belongs to class 02. The yellow box point "?" on the figure is either related to class01 or class02 which would be predicted by the algorithm. Let a and b be feature vectors $a = \left(a_1, a_2, \dots a_n\right)$ and $b = \left(b_1, b_2, \dots b_n\right)$. The considered distance functions are discussed as follows;

$$\text{Minkowski}_{d(a,b)} = \left(\sum_{i=1}^{n} \left|a_i - b_i\right|^p\right)^{\frac{1}{p}} \quad \text{where } p = 1, 2, \dots \infty. \tag{1}$$

$$\begin{aligned}\text{Euclidean}_{d(a,b)} &= \sqrt{\left(a_1 - b_1\right)^2 + \left(a_2 - b_2\right)^2 \cdots \left(a_n - b_n\right)^2} \\ &= \sqrt{\sum_{i=1}^{n} \left(a_i - b_i\right)^2}.\end{aligned} \tag{2}$$

$$\text{Manhattan}_{d(a,b)} = \sum_{i=1}^{n} \left(a_i - b_i\right)^2. \tag{3}$$

$$\text{Hamming}_{d(a,b)} = \text{def}\left\{\begin{array}{l} 0 \text{ if } a = b \\ 1 \text{ otherwise} \end{array}\right\}. \tag{4}$$

$$\text{Cosine}_{d(a,b)} = \frac{\sum_{i=1}^{n} \left(a_i\right)\left(b_i\right)}{\sqrt{\sum_{i=1}^{n} \left(a_i\right)^2} \sqrt{\sum_{i=1}^{n} \left(b_i\right)^2}}. \tag{5}$$

$$\text{Canberra}_{d(a,b)} = \sum_{i=1}^{n} \left(\frac{\left|a_i - b_i\right|}{\left|a_i\right| + \left|b_i\right|}\right). \tag{6}$$

$$\text{Correlation}_{d(a,b)} = \frac{\text{cov}(a, b)}{(\sigma a)(\sigma b)} = 1 - C_{ai,bi}. \tag{7}$$

3.3 SVM

Support vector machine (SVM) is a supervised machine learning algorithm which can be used for performing classification, regression and even outlier detection. The features of dataset are plotted in n-dimensional space. The two classes are differentiating by drawing a straight line called hyperplane [28, 29]. All the dataset points that lie on one side of the line will be considered as one class, whereas all the points that fall on the other side of the line will be labeled as second class. The strategy sounds simple enough, however, it is important to note that there is an infinite amount of lines to choose from. SVM helps with selecting the line that does the best job of classifying the data. The SVM algorithm not only selects a line that separates the two classes but also stays as far away from the closest samples as possible. In fact, the "support vector" in "support vector machine" refers to two position vectors drawn from the origin to the points which dictate the decision boundary [30]. Figure 4 shows the working principle of SVM.

Hyperplane Equation

$w = a_i S_i, y = wx + b$

3.4 Naive Bayes

Naive Bayes is a very simple algorithm to classify various classification problems. It is easy to build and can make very powerful and accurate predictions for large amount of data. This classifier is the probabilistic learning method based on Bayesian theorem [28, 31]. The working principle depends on three steps. In first step, dataset is converted into frequency table. The second step involves creating a likelihood table after finding out the probabilities. In the last step, the posterior probability is calculated with help of Naïve Bayes equation for each class. The class of highest posterior probability rate is the outcome of prediction [30]. Then, Bayes theorem is as follow

$$P\left(\frac{h}{D}\right) = \frac{\left[\left(P\frac{D}{h}\right) \cdot P(h)\right]}{P(D)}. \tag{8}$$

Perior Probability $= P(h)$,

Conditional Probability $= P\left(\frac{D}{h}\right)$, Mixture Density $P(D)$

3.5 Decision Tree

A famous method for decision making is a decision trees. A unique strategy of 'divide-and-conquer' is used by creating decision regions by dividing the instance space. Through a testing process, a root node is established. Then, dataset is broken by the value of related test attribute. It is a repeated process halted by providing a predefined stopping criterion. The class is indicated by a leaf node which is a node at the end of a tree. The decision rule is defined by the branch or the path of the node. Each new sample has its unique decision rule for classification purposes [10]. These classifications occur over three steps. Firstly, training data are used to train the model in the learning process. Secondly, a test is conducted to calculate the accuracy of the model and depending on this value, the model is either accepted or rejected. In order to use the model for further classification of a new datum, the value has to be accurate and have considerable acceptance. Thirdly and finally, the utilization of the model is decided by either using it for classification purposes or predicting new data [30, 32]. The Entropy and Gini equation are defined below in Eqs. 9 and 10 whereas decision tree working principle is shown in Fig. 5;

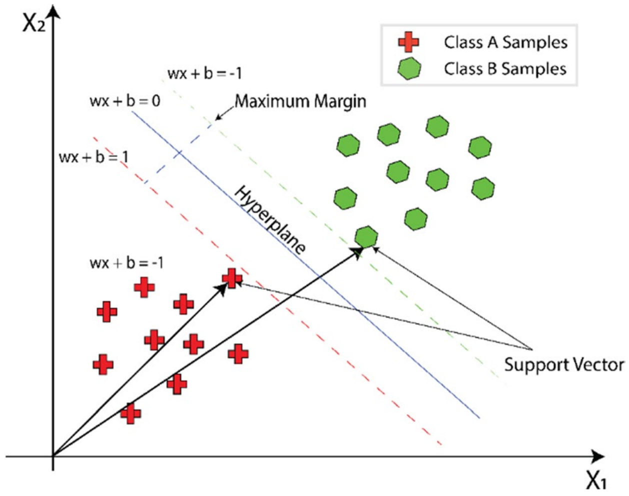


Fig. 4 Working principle of SVM

$$\mathrm{Ent}(D) = -\sum_{y\in Y} P(y|D)\log P(y|D). \tag{9}$$

$$G_{\mathrm{gini}}\left(D;\, D_1 \ldots, D_k\right) = I(D) - \sum_{i=1}^{k} \frac{\left|D_k\right|}{D} I\left(D_k\right)$$
$$\text{where } I(D) = 1 - \sum_{y\in Y} P(y|D)^2. \tag{10}$$

3.6 Logistic Regression

Logistic regression (LR) is a classification model in machine learning, which is widely used in the fields like medicine social science [30, 33]. Logistic regression has been used in many types of analysis to not only explore the risk factors of certain diseases but also for prediction of the probability of diseases. These predictions are discrete which refers to as specific values or categories. They can also view probability scores underlying the model's classifications. The logistic function is defined in Eq. 11 and its working principle is shown in Fig. 6.

$$\mathrm{Prob(event)} = P\left(\vec{x}\right) = \frac{1}{1 + e^{-g\left(\vec{x}\right)}} = \frac{e^{g\left(\vec{x}\right)}}{1 + e^{g\left(\vec{x}\right)}}. \tag{11}$$

where $P\left(\vec{x}\right)$ is the probability of some output event, $\vec{x}\left(x_1, x_2, \ldots x_k\right)$ is an input vector corresponding to the independent variables (predictors) and $g\left(\vec{x}\right)$ is the logit model.

3.7 Performance Evaluation Metrics

Classification algorithms can be evaluated in several ways. For evaluating various learning algorithms, the analysis of metrics should be interpreted correctly. For evaluating a diagnostic test, some of the measures derived from

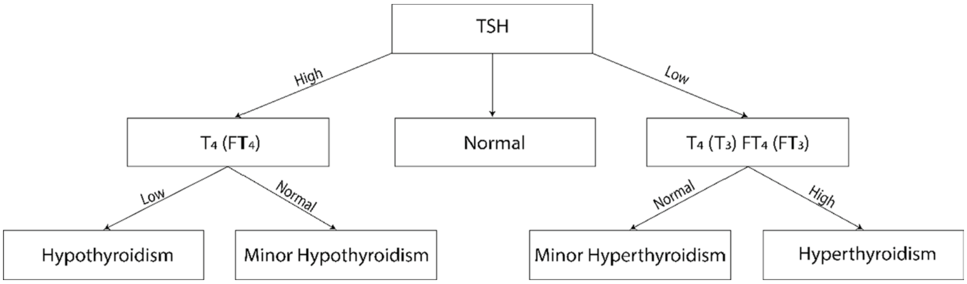


Fig. 5 Working principle of decision tree

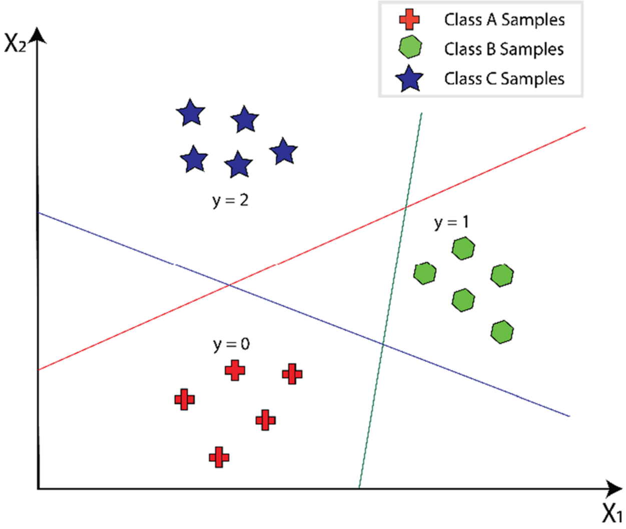


Fig. 6 Working principle of logistic regression

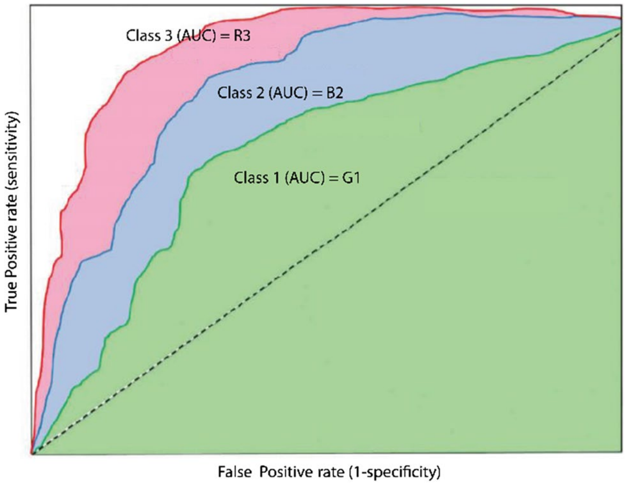


Fig. 7 Example of receiver operating characteristic (ROC) and area under curve (AUC) [12]

confusion matrix are reported in Sahu et al. [34] and Islam et al. [35]. There are four distinct terms used in a confusion matrix which are true positive (TP), false positive (FP), true negative (TN) and false negative (FN). True positive means that the system predicts the outcome to be a correct value and the result is also correct. False positive means that the system predicts the outcome to be a correct value however the result is false. True negative means that the system predicts the outcome to be a false value and the result is also a false value. False negative means that the system predicts the outcome to be a false value, whereas the result is a correct value. Another parameter to consider the performance of the classifier is 'ROC curve with Area Under Curve' (AUC). The receiver operating characteristics (ROC) curve is a two-dimensional graph in which the TPR represents the *y*-axis and FPR is the *x*-axis. The ROC curve has been used to evaluate many systems such as diagnostic systems, medical decision-making systems and machine learning systems [36]. In Fig. 7, it describes ROC curve with AUC values of three classes separated using colors and initialized as G1, B2 and R3. Class 3 has a large AUC value so its performance is better than class 2 and 1. If the classifier value is below the threshold line, then it would indicate poor performance of the class/model [12]. Some more derived measures from confusion matrix [37] are discussed as follows.

3.7.1 Accuracy and Error

The most important and commonly used factor to measure the performance of the classifier is accuracy. Accuracy (ACC) is calculated by the ratio of correct prediction samples to the total samples in the dataset.

$$\text{Accuracy(Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%. \tag{12}$$

However, error rate (ERR) represents the number of wrongly classified samples in both negative and positive class and calculated as follows.

$$\begin{aligned}\text{Error rate(ERR)} &= (1 - \text{Acc}) \times 100\% \\ &= \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%. \end{aligned} \tag{13}$$

3.7.2 Sensitivity and Specificity

Sensitivity (TPR) or recall is defined as the ratio of true predicted positive sample to the total number of positive sample. However, specificity (TNR) or selectivity is the ratio of true predicted negative sample to the total number of negative samples. Equations (14) and (15) represent TPR and TNR, respectively.

$$\begin{aligned}\text{Sensitivity or Recall(TPR)} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\% \\ &= 1-(\text{FNR}) \times 100\%.\end{aligned} \tag{14}$$

$$\begin{aligned}\text{Specificity, Selectivity (TNR)} &= \frac{\text{TN}}{\text{TN}+\text{FP}} \times 100\% \\ &= 1-(\text{FPR}) \times 100\%\ .\end{aligned} \tag{15}$$

3.7.3 False Positive and False Negative Rate

False positive rate (FPR) or fall-out in Eq. (16) represents the false positive prediction in the total number of negative samples. While, false negative rate (FNR) or miss-rate is the proportion of positive samples that were incorrectly classified in Eq. (17).

$$\text{Fall-Out(FPR)} = \frac{\text{FP}}{\text{FP}+\text{TN}} \times 100\% = 1-(\text{TNR}) \times 100\%. \tag{16}$$

$$\text{Miss-Rate(FNR)} = \frac{\text{FN}}{\text{FN}+\text{TP}} \times 100\% = 1-(\text{TPR}) \times 100\%. \tag{17}$$

3.7.4 Matthews Correlation Coefficient

Brain W. Matthews in 1975 introduced the Matthews correlation coefficient (MCC) [r]. This coefficient shows the relationship between observed and predicted classification. MCC is calculated from the confusion matrix and their + 1 value represents perfect prediction while − 1 value indicated the conflict between prediction and true values. Equation (18) defined MCC as

$$\begin{aligned}&\text{Matthews Corelation Coefficient (MCC)} \\ &= \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \times 100\%.\end{aligned} \tag{18}$$

3.7.5 *F*-Measure

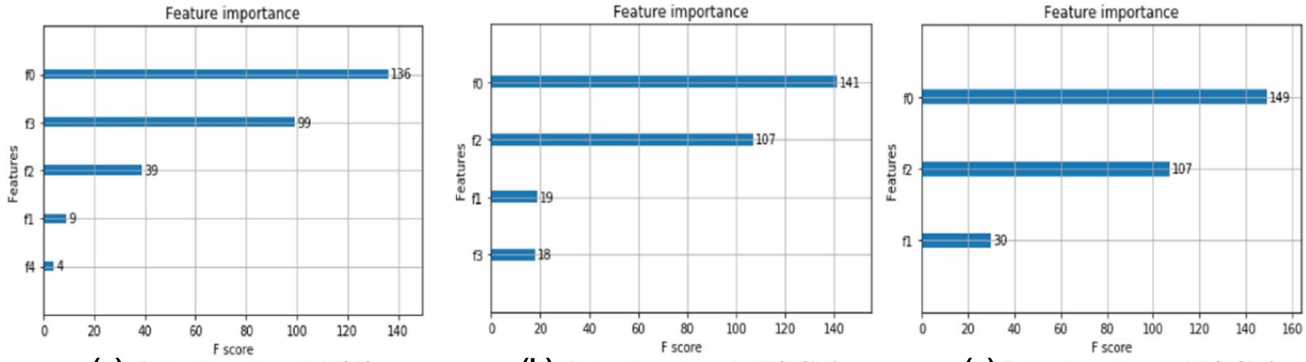
F-measure is also known as F_1-score. It described the harmonic mean between precision and recall. A model is considered good if its value is one or it have low false positive or low false negative, whereas zero value showed poor performance. Equation of F_1-score is as follows.

$$\begin{aligned}F1\text{ - Score} &= 2 \times \frac{\text{Precision } \times \text{recall}}{\text{Precision } + \text{ recall}} \times 100\% \\ &= \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \times 100\%.\end{aligned} \tag{19}$$

4 Experimental Data

The experimental analysis of this research study involved dependence on hardware and software performance of the system. The hardware system specification used in this experiment was Intel(R) Core i7-7700HQ CPU @ 2.80 GHZ, with 512 GB SSD, 2 TB HDD, 16 GB RAM and Nvidia 6 GB GTX 1060 GPU. On the other hand, the software description included the usage of scikit-learn [25] and Anaconda [38]. Scikit-learn was a great choice for its accessibility, simplicity and its great performance for analyzing data. Splitting method for training and testing dataset was used with five machine learning algorithms which were KNN, decision tree, SVM, logistic regression and Naive Bayes. The two well-known L_1 and L_2 feature selection techniques were used for the five machine learning algorithms. The experiment on thyroid dataset was repeated three times. The first iteration was without feature selection, abbreviated as (WOFS). The second attempt of the experiment was done by L_1-based feature selection denoted as WLSVC(L_1). The third iteration of the experiment was employed with L_2-based feature selection implemented denoted as WLSVC(L_2).

The training and testing time can be reduced if classifier uses important features. The importance of feature selection depends upon various parameters where one of them is F-score that determines the importance and usefulness of various features. Xgboost classifier has advantage to solve regression and classification problems. This technique prioritizes superior results using less resources in terms of computing and time. The main objective to use Xgboost in L_1 and L_2 feature selection technique is to prevent the model from overfitting. This study also selected various features depending upon their F-score values by using xgboost classifier [39] in which features were automatically named according to their index in the input array. Figure 8a indicates the result of experiment done using WOFS where algorithms gave weight to five features. According to the results, f_0 (TSH) has the highest importance and f_4 (pulse rate) has the lowest importance. Similarly, Fig. 8b with implementing WLSVC(L_1), four important features were selected based on their *F*-scores. In which WLSVC(L_1) f_0 (TSH) has the highest importance, whereas f_3 (BMI) has the lowest index. Lastly, with implementation of WLSVC(L_2) Fig. 8c,



(a) Feature Importance in WOFS **(b)** Feature Importance in WLSVC(L1) **(c)** Feature Importance in WLSVC(L2)

Fig. 8 Feature importance for thyroid data by using WOFS. **b** Feature importance for thyroid data by using WLSVC(L_1). **c** Feature importance for Thyroid Data by using WLSVC(L_2)

only three important features were selected based on their *F*-scores where f_0 (TSH) and f_1(T4) received the highest and lowest importance, respectively. The training and prediction time of thyroid dataset is shown in Table 2.

5 Result Analysis

Table 3 outlines the detailed performance of different classifiers for thyroid disease. Various performance evaluation metrics like accuracy, recall, fall-out and error-rate were used for this comparative study on the bases of the output confusion matrix. In Table 3, it is explicitly shown that accuracy gets improved when feature selection technique was applied. In part (a) without applying feature selection, Naive Bayes is providing more precise results. KNN also performed well and achieved 91.39% accuracy by using Minkowski distance function. SVM and logistic regression both attained satisfying performance with having accuracy of 80.46% and 90.32%, respectively. Decision tree had the lowest accuracy of 74.19% among all the classifiers and had a five level depth. Furthermore, results after implementing the WLSVC(L_1) feature selection were significantly improved. KNN accuracy jumped to a 97.84% accuracy with same distance function. SVM increased to 86.02% accuracy while decision tree with five level depth also showed a little improvement of having 75.34% accuracy. Logistic regression indicated 100% accuracy which was the best improvement among all. Part (c) demonstrated the result after applying WLSVC(L_2) feature selection. In this part, the algorithms with the highest accuracies were logistic regression and KNN, SVM and decision tree also demonstrated some improvement, whereas Naive Bayes showed the maximum accuracy of 100% in both part (b) and (c).

Another crucial parameter is the ROC curve with the area under curve (AUC) value, which is used to check the classifier's performance. Range of the AUC from '0' to '1' demonstrates that a classifier has a better performance if its value is or close to '1.' ROC curves are constructed in Origin Pro 8.5 software, and AUC is calculated with the help of trapezoid rule. Naive Bayes achieved the highest value of AUC 1.00 in all three experiments as indicated in all parts of Fig. 9. KNN and logistic regression came at the second place by achieving 0.98 and 0.97 AUC values, respectively, in without feature selection. Furthermore, both of these classifiers showed 1.00 AUC in both WLSVC(L_1) and WLSVC(L_2). SVM overall performance was satisfactory, indicating AUC value of 0.94 in WOFS whereas AUC values of 0.95 and 0.98 in WLSVC(L_1) and WLSVC(L_2), respectively. Lastly, decision tree had comparatively the lowest performance in all three parts of the experiment.

Table 2 New thyroid dataset training and prediction time in seconds

Classifier	Training time WOFS (s)	Predicting time WOFS (s)	Training time WLSVCL$_1$ (s)	Predicting time WLSVCL$_1$ (s)	Training time WLSVCL$_2$ (s)	Predicting time WLSVCL$_2$ (s)
KNN	0.695	0.42	0.53	0.361	0.51	0.369
Decision tree	0.763	0.422	0.629	0.360	0.681	0.372
Naïve Byes	0.659	0.388	0.549	0.358	0.574	0.367
SVM	0.601	0.398	0.506	0.359	0.511	0.361
Logistic regression	0.510	0.339	0.449	0.142	0.439	0.152

Table 3 Performance evaluation metrics for thyroid dataset

Classifier	Performance factor	(a) Without feature selection (WOFS)		(b) With feature selection WLSVC(L_1)		(c) With feature selection WLSVC(L_2)	
		Values	Confusion matrix	Values	Confusion matrix	Values	Confusion matrix
KNN	Accuracy (%)	91.39	$\begin{vmatrix} 18 & 8 & 0 \\ 0 & 43 & 0 \\ 0 & 0 & 24 \end{vmatrix}$	97.84	$\begin{vmatrix} 14 & 2 & 0 \\ 0 & 54 & 0 \\ 0 & 0 & 23 \end{vmatrix}$	96.77	$\begin{vmatrix} 17 & 3 & 0 \\ 0 & 51 & 0 \\ 0 & 0 & 22 \end{vmatrix}$
	Recall (%)	90		96		95	
	Fall-out (%)	5		2		2	
	Specificity (%)	95		98		98	
	F_1-score (%)	92		97		96	
	Error rate (%)	8.61		2.16		3.23	
	MCC (%)	78.6		92.4		90	
	Miss-rate (%)	10		4		5	
Decision tree	Accuracy (%))	74.19	$\begin{vmatrix} 26 & 0 & 0 \\ 0 & 43 & 0 \\ 0 & 24 & 0 \end{vmatrix}$	75.34	$\begin{vmatrix} 16 & 4 & 0 \\ 0 & 39 & 0 \\ 0 & 14 & 0 \end{vmatrix}$	76.92	$\begin{vmatrix} 20 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 21 & 0 \end{vmatrix}$
	Recall (%)	67		67		67	
	Fall-out (%)	49		20		17	
	Specificity (%)	51		80		83	
	F_1-score (%)	60		62		61	
	Error rate (%)	25.81		24.66		23.08	
	MCC (%)	100		100		100	
	Miss-rate (%)	33		33		33	
Naïve Bayes	Accuracy (%)	100	$\begin{vmatrix} 26 & 0 & 0 \\ 0 & 43 & 0 \\ 0 & 0 & 24 \end{vmatrix}$	100	$\begin{vmatrix} 16 & 0 & 0 \\ 0 & 54 & 0 \\ 0 & 0 & 23 \end{vmatrix}$	100	$\begin{vmatrix} 20 & 0 & 0 \\ 0 & 51 & 0 \\ 0 & 0 & 22 \end{vmatrix}$
	Recall (%)	100		100		100	
	Fall-out (%)	0		0		0	
	Specificity (%)	100		100		100	
	F_1-score (%)	100		100		100	
	Error rate (%)	0		0		0	
	MCC (%)	100		100		100	
	Miss-rate (%)	0		0		0	
SVM	Accuracy (%)	80.46	$\begin{vmatrix} 9 & 17 & 0 \\ 0 & 43 & 0 \\ 0 & 6 & 18 \end{vmatrix}$	86.02	$\begin{vmatrix} 7 & 9 & 0 \\ 0 & 54 & 0 \\ 0 & 4 & 19 \end{vmatrix}$	86.02	$\begin{vmatrix} 10 & 10 & 0 \\ 0 & 51 & 0 \\ 0 & 3 & 19 \end{vmatrix}$
	Recall (%)	70		76		79	
	Fall-out (%)	15		20		85	
	Specificity (%)	85		80		90	
	F_1-score (%)	78		84		85	
	Error rate (%)	19.54		13.98		13.98	
	MCC (%)	52.6		62.6		66.3	
	Miss-rate (%)	30		24		21	
Logistic regression	Accuracy (%)	90.32	$\begin{vmatrix} 17 & 9 & 0 \\ 0 & 43 & 0 \\ 0 & 0 & 24 \end{vmatrix}$	100	$\begin{vmatrix} 16 & 0 & 0 \\ 0 & 54 & 0 \\ 0 & 0 & 23 \end{vmatrix}$	98.92	$\begin{vmatrix} 19 & 1 & 0 \\ 0 & 51 & 0 \\ 0 & 0 & 22 \end{vmatrix}$
	Recall (%)	88		100		98	
	Fall-out (%)	6		0		1	
	Specificity (%)	94		100		99	
	F_1-score (%)	91		100		98	
	Error rate (%)	9.68		0		1.08	
	MCC (%)	80		100		97	
	Miss-rate (%)	12		0		2	

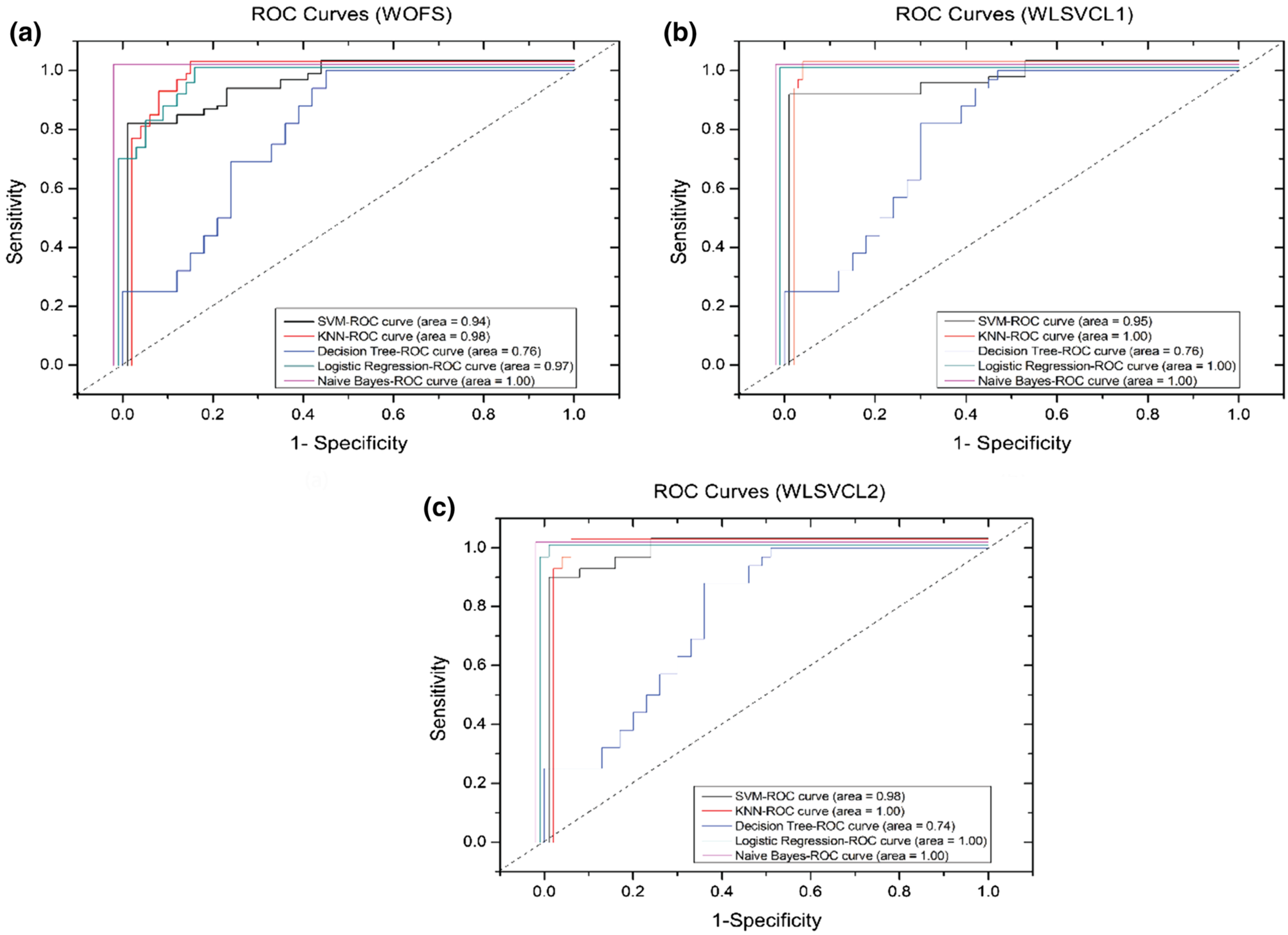


Fig. 9 ROC curves with AUC before using feature selection technique. **b** ROC curves with AUC after using WLSVC(L_1). **c** ROC curves with AUC after implementing WLSVC(L_2)

From Table 2, it is clearly shown that prediction time gets improved. Before applying feature selection technique, prediction time is comparatively higher in all classifiers. The algorithm which has the best accuracy, minimum error rate and the lowest prediction time is Naïve Bayes in all three parts of the experiment, whereas logistic regression and KNN both performed great in terms of minimum error rate and low prediction time in part (b) of the experiment. According to the original data, healthy individuals are 170, 66 are suffering from hyperthyroidism and 73 with hypothyroidism. After applying different classifiers, the result indicated that detection of Naïve Bayes (in all three parts of experiment) and logistic regression (in part b of experiment) is excellent with 100% accuracy. Moreover, KNN detection is closer to the original data. From KNN, it is determined that 146 are healthy, whereas 66 and 73 have hyperthyroidism and hypothyroidism, respectively.

6 Related Existing Studies

The approach utilized in this study has been investigated alongside with other related existing studies shown in Table 4. Our model dataset is distinguished with these existing studies because of three new features as described in Sect. 2. The proposed study results which were achieved by using different supervised classifiers. Higher accuracy, low training and prediction time were the significant goals of this research. Other existing models use hybrid approaches with a combination of different algorithms and complex models. Such methodologies are not only costly to achieve accurate data, but also take an increased time for training and validation.

7 Conclusion

Disease detection and its early diagnosis are very important for human life. By using machine learning algorithms, precise and accurate identification and detection have become

Table 4 Related others existing models

References	Methodology	Accuracy %	Dataset
Deepika et al. [30]	SVM	95.62	UCI Repository thyroid disease dataset
	DT	95.00	
	ANN	98.60	
Pal et al. [6]	Naïve Bayes	94.70	KEEL repository thyroid disease dataset
	SVM	92.70	
	KNN	96.90	
Chandel et al. [9]	KNN	93.44	KEEL repository thyroid disease dataset
	Naïve Bayes	22.56	
Turanoglu-Bekar et al. [10]	NBTREE	75.00	Local hospital
	LADTREE	66.25	
	REPTREE	62.50	
	BFTREE	65.00	
Chalekar et al. [27]	KNN	97.00	UCI repository thyroid disease dataset
Tyagi et al. [40]	ANN	97.50	UCI repository thyroid disease dataset
	KNN	98.00	
	DT	75.00	
This study	Naive Bayes	100	District Headquarters (DHQ) Teaching Hospital, Dear Ghazi Khan, Pakistan
	Logistic regression	100	
	K-NN	97.84	
	SVM	86.02	
	Decision tree	75.34	

more achievable. Thyroid disease is not easy to diagnosis because mix-up of their symptoms with other condition. The three newly introduced features in thyroid dataset in this research show the positive impact on classifier performance and results show that it gives best accuracies than the existing studies. After comparison and analysis of KNN, Naïve Bayes, SVM, decision tree and logistic regression, it was observed that 100% accuracy is achieved by Naïve Bayes in all three parts of experiment, while logistic regression gained second best accuracy 100% and 98.92% in L_1- and L_2-based feature selection, respectively. KNN also carried out excellent result accuracy of 97.84% with error rate of 2.16%. Upon analyzing the results, the advantages and robustness of new dataset are clearly seen and would allow doctors to get more precise and accurate results in less time. However, in the future classifiers with different distance functions of KNN and data augmentation techniques can be used for more precise results.

Acknowledgements I would like to extend my sincere gratitude to Dr. Abid Hussain, Dr. Zarnab Lashari and Dr. Aimen Javed for aiding in gathering Thyroid Data and contributing continuous support. I am thankful to them for their invaluable guidance.

References

1. Miller, K.D., et al.: Cancer treatment and survivorship statistics, 2016. CA Cancer J. Clin. **66**(4), 271–289 (2016)
2. Shroff, S.; Pise, S.; Chalekar, P.; Panicker, S.S.: Thyroid disease diagnosis: a survey. In: IEEE 9th International Conference on Intelligent Systems and Control, 2015 (ISCO 2015), pp. 1–6. IEEE (2015)
3. Thyroid Cancer: https://seer.cancer.gov/statfacts/html/thyro.html. Accessed 01 Jan 2020
4. Thyroid Problems: https://medlineplus.gov/thyroiddiseases.html. Accessed 01 Jan 2020
5. What Is Thyroid Cancer: https://www.cancer.org/cancer/thyroid-cancer/about/what-is-thyroid-cancer. Accessed 01 Jan 2020
6. Pal, R.; Anand, T.; Dubey, S.K.: Evaluation and performance analysis of classification techniques for thyroid detection. Int. J. Bus. Inf. Syst. **28**(2), 163–177 (2018)
7. Thyroid Patient Information: https://www.thyroid.org/thyroid-information/. Accessed 01 Jan 2020
8. Acharya, U.R.; Choriappa, P.; Fujita, H., et al.: Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images. Knowl. Based Syst. **107**, 235–245 (2016)
9. 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. **4**(2–4), 313–319 (2016)
10. Bekar, E.T.; Ulutagay, G.; Kantarcı, S.: Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms. Oxf. J. Intell. Decis. Data Sci. **2016**(2), 13–28 (2016)
11. Prasad, V.; Rao, T.S.; Babu, M.S.P.: Thyroid disease diagnosis via hybrid architecture composing rough data sets theory and machine learning algorithms. Soft Comput. **20**(3), 1179–1189 (2016)
12. Mushtaq, Z.; Yaqub, A.; Sani, S.; Khalid, A.: Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. J. Chin. Inst. Eng. **43**(1), 1–13 (2019)
13. Tomar, D.; Agarwal, S.: A survey on data mining approaches for healthcare. Int. J. Bio-Sci. Bio-Technol. **5**(5), 241–266 (2013)
14. Jahantigh, F.F.: Kidney diseases diagnosis by using fuzzy logic. In: 2015 International Conference on Industrial Engineering and Operations Management, 2015 (IEOM2015), pp. 2369–2375. IEEE (2015)
15. Durairaj, M.; Ranjani, V.A.: Data mining applications in healthcare sector: a study. Int. J. Sci. Technol. Res. **2**(10), 29–35 (2013)

16. Liu, D.Y.; Chen, H.-L.; Yang, B.; Lv, X.-E.; Li, L.-N.; Liu, J.: Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnostic system for thyroid disease. J. Med. Syst. **36**(5), 3243–3254 (2012)
17. Acharya, U.R.; Vinitha Sree, V.S.; Molinari, F.; Garberoglio, R.; Witkowska, A.; Suri, J.S.: Automated benign and malignant thyroid lesion characterization and classification in 3D contrast-enhanced ultrasound. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2012 (EMBS2012), pp. 452–455. IEEE (2012)
18. Kousarrizi, M.R.N.; Seiti, F.; Teshnehlab, M.: An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification. Int. J. Electr. Comput. Sci. **12**(1), 13–19 (2012)
19. Chen, H.L.; Yang, B.; Wang, G.; Liu, J.: A three-stage expert system based on support vector machines for thyroid disease diagnosis. J. Med. Syst. **36**(3), 1953–1963 (2012)
20. Dogantekin, E.; Dogantekin, A.; Avci, D.: An expert system based on generalized discriminant analysis and wavelet support vector machine for diagnosis of thyroid diseases. Expert Syst. Appl. **38**(1), 146–150 (2011)
21. Keleş, A.; Keles, A.: ESTDD: expert system for thyroid diseases diagnosis. Expert Syst. Appl. **34**(1), 242–246 (2008)
22. Ozyilmaz, L.; Yildirim, T.: Diagnosis of thyroid disease using artificial neural network methods. In: 9th International Conference on Neural Information Processing, 2002 (ICONIP2002), pp. 2033–2036, IEEE (2002)
23. Teaching Hospital - Dera Ghazi Khan: http://thdgkhan.org/. Accessed 15 Mar 2020
24. Alcalá-Fdez, J.; Sánchez, J.L.; Garc, S.; Jesus, M.J.D., et al.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Mult. Valued Log. Soft Comput. **17**, 255–287 (2011)
25. Pedregosa, F.; Weiss, R.; Brucher, M.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**(2011), 2825–2830 (2011)
26. Li, C.; Zhang, S.; Zhang, H.; Pang, L.; Lam, K.; Hui, C.; Zhang, S.: Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. Comput. Math. Methods Med. (2012)
27. Chalekar, P.; Shroff, S.; Pise, S.; Panicker, S.S.: Use of K-nearest neighbor in thyroid disease classification. Int. J. Curr. Eng. Sci. Res. **1**(2), 2394–2697 (2014)
28. Mushtaq, Z.; Yaqub, A.; Hassan, A.; Su, S.F.: Performance analysis of supervised classifiers using PCA based techniques on breast cancer. In: International Conference on Engineering and Emerging Technologies, 2019 (ICEET2019), pp. 1–6, IEEE (2019)
29. Aboudi, N.; Guetari, R.; Khlifa, N.: Multi-objectives optimisation of features selection for the classification of thyroid nodules in ultrasound images. IET Image Process. **14**(9), 1901–1908 (2020)
30. Deepika, M.; Kalaiselvi, K.: A empirical study on disease diagnosis using data mining techniques. In: International Conference on Inventive Communication and Computational Technologies, 2018 (ICICCT2018), pp. 615–620, IEEE (2019)
31. Zhou, Z.-H.: Ensemble Methods: Foundations and Algorithms—Zhi-Hua Zhou—Google Books. CRC Press, Boca Raton (2012)
32. Lavanya, D.; Rani, K.U.: Performance evaluation of decision tree classifiers on medical datasets. Int. J. Comput. Appl. **26**(4), 1–4 (2011)
33. Yang, Y.; Chen, G.; Reniers, G.: Vulnerability assessment of atmospheric storage tanks to floods based on logistic regression. Reliab. Eng. Syst. Saf. **196**, 106721 (2019)
34. Sahu, B.; Mohanty, S.; Rout, S.: A hybrid approach for breast cancer classification and diagnosis. ICST Trans. Scalable Inf. Syst. **6**(20), 2–8 (2019)
35. Islam, M.M.; Iqbal, H.; Haque, M.R.; Hasan, M.K.: Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: 5th IEEE Region 10 Humanitarian Technology Conference. 2017, pp. 226–229, IEEE (2017)
36. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006). https://doi.org/10.1016/j.patrec.2005.10.010
37. Tharwat, A.: Classification assessment methods. Appl. Comput. Inf. (2018). https://doi.org/10.1016/j.aci.2018.08.003
38. Anaconda: https://www.anaconda.com/. Accessed 05 Jan 2020
39. Feature Importance and Feature Selection with XGBoost in Python: https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/. Accessed 05 Jan 2020
40. Tyagi, A.; Mehra, R.; Saxena, A.: Interactive thyroid disease prediction system using machine learning technique. In: PDGC 2018–2018 5th International Conference on Parallel, Distributed and Grid Computing, pp. 689–693 (2018). https://doi.org/10.1109/PDGC.2018.8745910