International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229398

A Novel Machine Learning Framework for Prediction of Early-Stage Thyroid Disease Using Classification Techniques

Annapurna Gummadi[1], D. Rammohan Reddy[2]
[1]M.Tech Student, CSE Department, Newton's Institute of Engineering, Macherla, India
[2]Assoc Prof, CSE Department, Newton's Institute of Engineering, Macherla, India

Article Info
Volume 9, Issue 3
Page Number : 467-479

Publication Issue
May-June-2022

Article History
Accepted : 01 June 2022
Published : 07 June 2022

ABSTRACT

Thyroid disease is one of the most common diseases among the female Population in Bangladesh. Hypothyroid is a common variation of thyroid disease. It is clearly visible that hypothyroid disease is mostly seen in female patients. Most people are not aware of that disease as a result of which, it is rapidly turning into a critical disease. It is very much important to detect it in the primary stage so that doctors can provide better medication to keep itself turning into a serious matter. Predicting disease in machine learning is a difficult task. Machine learning plays an important role in predicting diseases. Again distinct Predicting techniques have facilitated this process analysis and assumption of diseases. There are two types of thyroid diseases namely Hyperthyroid and Hypothyroid. Here, in this paper, we have attempted to predict hypothyroid in the primary stage. To do so, we have mainly used classification algorithms named Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Naive Bayes (NB). By observing the results, we could extrapolate that our Trained (Structured) Dataset provide's an (approx.) 97.05% accuracy for Random Forest (Bagging) classification algorithm.

Keywords : Machine learning, SVM, NB, Decision tree, Random Forest, classification, thyroid.

I. INTRODUCTION

At the current state, the thyroid is one of the most critical diseases of all and it has quite the potential to be transformed into a common disease among the female mass. In Bangladesh, according to experts, 50 million people suffer from thyroid disease. Among them, females are at 10 times more risk of being affected with thyroid disease. Though a vast majority of 50 million people are affected with thyroid disease, yet almost 30 million people among them are totally not aware of this condition. A study from the Bangladesh Endocrine Society(BES) depicts that around 20-30% of females are suffering from thyroid disease [14]. The thyroid is a gland that is situated in the middle of the neck in our body. It is butterfly-

shaped and small in size. It secretes several hormones that are mixed with blood and travel across the body to control various activities. The thyroid hormone is responsible for conserving metabolism, sleep, growth, sexual function, and mood. Depending on the secretion of thyroid hormone we can feel tired or restless and also may have weight loss.

There are two main thyroid hormones: Triiodothyronine (T3) and Thyroxin (T4). These two hormones are mainly responsible for maintaining the energy in our bodies. Thyroid Stimulating Hormone(TSH) is produced by the pituitary gland that helps the thyroid gland to release T3 and T4. There are two common thyroid diseases- 1) Hypothyroid 2) Hyperthyroid. Hypothyroid: When the thyroid gland cannot generate enough thyroid hormones the level of T3 and T4 becomes low and the level of TSH become high. Symptoms it presents are weight loss, tiredness, brain fog, etc. Hyperthyroid: When the thyroid gland produces more thyroid hormone than our body actually needs, the level of T3 and T4 becomes too high and the level of TSH becomes low. Symptoms it presents are- hair loss, anxiety, sweating, etc. In our research, we have concentrated on hypothyroid since it is the one that is most common among the females in Bangladesh. Therefore, our research mainly focused on detecting hypothyroid in the primary stage.

II. MACHINE LEARNING

Machine learning has become an important part of human lives that provides smart and affordable solutions to various problems. As such, healthcare is catching the attention of many researchers, as society relies upon healthy and performing individuals for its balanced functioning. It is obvious that a diseased person would spend much of his time in fretting about his health, thus leaving very little productive time left to complete the assigned duties, let alone perform well.

The reason being they might be suffering from a thyroid disorder, called hyperthyroidism. Some may feel drowsy and lethargic, which is a case of hypothyroidism. The thyroid malfunction is one of the common diseases affecting people from all age groups. The disease is not dangerous as other diseases like heart disease and cancer, but it may be the cause of other diseases with severe complications

Machine learning (ML) is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the extract information from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is in rise. Many industries apply machine learning to extract relevant data. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves without being explicitly programmed. Many mathematicians and programmers apply several approaches to find the solution of this problem which are having huge data sets.

Table 1. ML algorithms for various model building approaches

Learning type	Model building	Examples
Supervised	Algorithms or models learn from labelled data (task-driven approach)	Classification, regression
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, associations, dimensionality reduction
Semi-supervised	Models are built using combined data	Classification, clustering

	(labelled + unlabeled)	
Reinforcement	Models are based on reward or penalty (environment-driven approach)	Classification, control

Supervised learning: It consists of a given set of input variables (training data) which are pre labelled and target data [5]. Using the input variables it generates a mapping function to map inputs to required outputs. Parameter adjustment procedure continues until the system acquired a suitable accuracy extent regarding the teaching data.

Unsupervised learning: In this algorithm we only have training data rather a outcome data. That input data is not previously labelled. It is used in classifiers by recognizing existing patterns or cluster in the input datasets [4].

Reinforcement learning: Applying this algorithm machine is trained to map action to a specific decision hence the reward or feedback Signals are generated. The machine trained itself to find the most rewarding actions by reward and punishment using past experience

There are massive numbers of algorithms used by machine learning are designed to erect models of machine learning and implemented in it [4]. All algorithms can be grouped by their learning methodology, as follows:

Regression algorithms: In Regression algorithms predictions are made by the model with modelling the relationship between variables using a measure of error[25]. Continuously varying value is predicted by the Regression technique. The variable can be a price, a temperature.

Instance based learning algorithms: In the algorithms which based on Instance, decision problem is a issue with illustration of training data build up a database and compare test data then form a prediction. Instance-based learning method is famous as lazy learner.

Algorithms using Decision Tree: Algorithms using Decision trees are used mainly in classification problem. They splits attributes in two or more groups by sorting them using their values. Each tree have nodes and branches [4]. Attributes of the groups are represented by each node and each value represented by branch [5].

Baysian algorithms: Machine Learning is multidisciplinary field of Computer Science like Statistics and algorithm. Statistics manages and quantifies the uncertainty and are represented by bayesian algorithms based on probability theory and Bayes' Theorem.

Data Clustering algorithms: This algorithm split items into different types of batches. It groups the item set into clusters in which each subset share some similarity. It is unsupervised learning method and its methods are categorized as hierarchical or network clustering and partitioned clustering.

Learning algorithms using Association Rule: Learning algorithms using Association rule are generally utilized by the organization commercially when multidimensional datasets are huge in size. They are used as extraction methods that can explore observed relationships between variables and data.

Algorithms using Artificial Neural Network: Artificial neural networks models are based on the biological neuron structure and uses supervised learning. It consists of artificial neurons which have weighted interconnections among units. They are also well known by parallel distributed processing networks.

Deep Learning algorithms: Deep Learning methods upgraded the artificial neural networks They are more complex neural networks are large in size.

Algorithms using Dimensionality Reduction: Dimensionality reduction method is widely used in case of large number of dimensions, large volume of space concerned. Then that problem requires a statistical significance. Dimensionality reduction methods used for minimizing the number of dimensions outlined the item and removes unrelated and un essential data which lessen the computational cost. Some of these methods are used in classifying and regression.

Ensemble Algorithms: They are based on unsupervised Learning. It groups the teaching data into many types of classes of data. Self-supporting models for learning are built for those groups. To make correct hypothesis all learning models are combined.

III. LITERATURE SURVEY

1). Early diagnosis of heart disease using classification and regression trees, Authors: Amir Mohammad Amiri, Giuliano Armano.

Early diagnosis of heart defects is very important for medical treatment. In this paper, we propose an automatic method to segment heart sounds, which applies classification and regression trees. The diagnostic system, designed and implemented for detecting and classifying heart diseases, has been validated with a representative dataset of 116 heart sound signals, taken from healthy and unhealthy medical cases. The ultimate goal of this research is to implement a heart sounds diagnostic system, to be used to help physicians in the auscultation of patients, with the goal of reducing the number of unnecessary echocardiograms and of preventing the release of newborns that are in fact affected by a heart disease. In this study, 99.14% accuracy, 100% sensitivity, and 98.28% specificity were obtained on the dataset used for experiments.

2).An Intelligent System for Thyroid Disease Classification and Diagnosis, Authors: A K Aswathi; Anil Antony

Data mining Techniques play a vital role in healthcare organizations such as for decision making, diagnosing disease and giving better treatment to the patients. Thyroid gland plays a major role in maintaining the metabolism of human body. Data mining in health care industry provides a systematic use of the medical data. Thyroid diseases are most common today. Early changes in the thyroid gland will not affect the proper working of the gland. By the early identification of thyroid disorders, better treatment can be provided in the early stage thus can avoid thyroid replacement therapy and thyroid removal up to an extent. This paper proposes a method for the classification and diagnosis of thyroid disease that a user is suffering from along with disease description and healthy advices. Support Vector Machine is used for classification. To optimize SVM parameters Particle Swarm Optimization is applied. User is provided with a window to enter the details such as the values of TSH, T3, T4 etc. There may be some values missing while the user entering the values. K-Nearest Neighbor algorithm is used for approximating the missing values in the user input.

3).Prediction of thyroid Disease Using Data Mining Techniques, Authors : Amina Begum; A Parkavi.

Classification based Data mining plays important role in various healthcare services. In healthcare field, the important and challenging task is to diagnose health conditions and proper treatment of disease at the early stage. There are various diseases that can be diagnosed early and can be treated at the early stage. As for example, Thyroid diseases. the traditional ways of diagnosing thyroid diseases depends on clinical examination and many blood tests. The Main task is to detect disease diagnosis at the early stages with higher accuracy. Data mining techniques plays an important role in healthcare field for making decision, disease diagnosis and providing better treatment for the patients at low cost. Thyroid disease Classification

is an important task. The purpose of this study is predication of thyroid disease using different classification techniques and also to find the TSH, T3,T4 correlation towards hyperthyroidism and hyporthyroidism and also to finding the TSH, T3,T4 correlation with gender towards hyperthyroidism and hyporthyroidism.

4).Feature selection algorithms to improve thyroid disease diagnosis, Authors : K. Pavya; B. Srinivasan.

Correct and early diagnosis of diseases is important and mandatory in healthcare industry for correct and timely treatment. This fact is more important in diseases like thyroid, which is very difficult to detect as its symptoms coincide with several diseases. Usage of machine learning algorithms for thyroid disease diagnosis is prominent. A typical thyroid disease diagnosis system uses three main steps, namely, feature extraction, feature selection and classification. The main goal of this paper is to analyze the use of filter-based (F-Score) and wrapper-based (Recursive Feature Elimination) feature selection algorithms on its effect on disease identification and classification. The analysis is also performed with Principle Component Analysis dimensionality reduction algorithms. Performance evaluation was performed with three metrics, namely, accuracy, sensitivity and specificity. Four classifiers, namely, MultiLayer Perceptron, Back Propagation Neural Network, Support Vector Machine and Extreme Learning Machine were used to analyze the selected algorithms. Experimental results showed that while both F-Score and Recursive Feature Elimination improved the performance of thyroid disease diagnosis, the wrapper-based algorithm produced maximum efficiency and produced a maximum accuracy of 98.14% with ELM classifier.

5).Thyroid Disease Diagnosis Based on Genetic Algorithms Using ANN and SVM, Authors: Fatemeh Saiti; Afsaneh Alavi Naini; Mahdi Aliyari Shoorehdeli; Mohammad Teshnehlab.

Thyroid gland produces thyroid hormones to help the regulation of the body's metabolism. The abnormalities of producing thyroid hormones are divided into two categories. Hypothyroidism which is related to production of insufficient thyroid hormone and hyperthyroidism related to production of excessive thyroid hormone. Separating these two diseases is very important for thyroid diagnosis. Therefore support vector machines and probabilistic neural network are proposed to classification. These methods rely mostly on powerful classification algorithms to deal with redundant and irrelevant features. In this paper feature selection is argued as an important problem via diagnosis and demonstrate that provide a simple, general and powerful framework for selecting good subsets of features leading to improved diagnosis rates. Thyroid disease datasets are taken from UCI machine learning dataset.

IV. RELATED WORK

Deepika Koundal et al.[6] have studied the existing the earlier automatic tools for diagnosis of disease at the easier stage in an efficient way. Also the metrics study about the different evaluation of performance and also investigations on the trends and future developments are
studied.

Nikita Sigh and Alka Jindal [7] have compared Support Vector Machine with K –Nearest Neighbor and Bayesian and concluded Support Vector Machine better then KNN and Bayesian with an accuracy about 84.62%.KNN found the nearest neighbourhood automatically. The results is represented by graph with object as each vertices. The probability classification is done using Bayesian which indicates the sample data belongs to a class.

Edgar Gabriel et al.[8] have proposed a texture-based segmentation i,e two parallel versions of a code for Fine Needle Aspiration Cytology thyroid images is the most important first step in identifying a fully automated Computer Aided Design solution. The code

is developed in MPI version to exploit computer resources such as PC clusters.

Preeti Aggawal et al.[9] listed the method for an automatic segmentation. The study shows the summary obtained by applying specific algorithm(automatic) segmentation and automatic tools on both thyroid US as well on lung CT [7]. For segmentation of thyroid US images they have used Analyze 10.0 and Mazda . Eystraints G[10]have provided system TND(Thyroid Nodule Detector) using a technique called computer aided diagnosis(CAD).During thyroid Ultra Sound examinations ,a nodular tissue detection is used in ultrasound(US) and thyroid images videos acquired.

Won-Jin Moon et al.[8]have evaluated to differentiate between benign an d malignant thyroid nodules using the accuracy of diagnostic ultrasonography (USG). They concluded that the important criteria and presence of calcification is shape,margin,echogenicity from bening nodules is discrimination of malignant.

S.Yasodha et al.[11] have proposed hybridization of Class Attribute Contingency Coefficiant(CACC)-Support Vector Machine techniques. The combination of CACC and SVM classification techniques are applied on thyroid data when compared to other traditional models,the accuracy of the proposed model is better.

Alfonso Bastias et al.[4]have aimed at developing an machine learning classifier using AIS for diagnosis of health condition and of the proposed classifier for capability investigation. The proposed classifier successfully improved the thyroid gland disease identification process.

Gurmeet et al. [3] has proposed NN training diagnosis model for the of the thyroid disease. It aims in developing the general model for identifying any kind of disease. The objective of this paper is to thyroid disease diagnose by using three different artificial neural network algorithm having different framework,characteristics and accuracy Ali keles et al. [7] proposed an expert system for predicting of thyroid that is known as Expert System for Thyroid Disease Diagnosis(ESTDD).This expert system diagnose thyroid diseases through neuro fuzzy rules with 95.33% of accuracy.

V. DATASET DESCRIPTION

Dataset is taken from UCI machine learning repository [15]. Database consists of patients thyroid records. Each thyroid patients record is consists of 15 attributes listes below. Attribute can be Boolean (true /false) or continuous valued are in given below table

Table 1. Data Description

S.No	Attribute Name	Value Type
1	Age	Continuous
2	Sex	m,f
3	On_thyroxine	f t
4	Query_on_thyroxine	f t
5	Thyroid_surgery	f t
6	Query_hypothyroid	f t
7	Query_hyperthyroid	f t
8	Pregnant	f t
9	Goitre	f t
10	TSH value	Continuous
11	T3 value	Continuous
12	TT4 value	Continuous
13	T4U value	Continuous
14	FTT value	Continuous
15	TBG value	Continuous

VI. CLASSIFICATION TECHNIQUES

Decision Tree

A Decision tree[6][8] has 3 types of node such as internal node that represents test attribute, the classes or class attribute are denoted by the leaf node , the top most is denoted by the root node of the tree. To construct the decision tree C4.5 and ID3 algorithms

are used .The Advantages of using Decision tree is to identify and eliminate the redundant data known as "tree pruning" to improve the accuracy of the classification. The decisions are made on attribute with the highest normalized data also it can applied to both continuous and discrete
values. On the other hand the disadvantages includes , for large data bases the efficiency and scalability are low.

Back propagation Neural Network

Back propagation is a neural network algorithm. It consists of three different layers, input layer - the inputs are given here, hidden layer – the input to hidden layer can the outputs with weights [5], number of hidden layer's arbitrary, output layer- the input to the output layer is from hidden layers , which eliminates prediction of the network's. Thus the advantages includes high accuracy, Very flexible for noisy and when the data is inconsistence, easy update of weights. The Disadvantages of Back propagation Neural Network are representation of knowledge, it is difficult for humans to interpret, Knowledge .Decreases the accuracy of the network by the removal of weighted links. Selection of training dataset is difficult.

Support Vector Machine

One of the type of learning system algorithm is Support Vector Machine[8] ,which is used to perform classification in a better accurately and uses 2 class classifier, referred as hyper plane as "decision boundary or decision surface". The hyper plane separates positive training sample with the negative training data sample in an plan. The advantages includes an easy extend, used for pattern reorganization, quadratic optimization problem can be formulated .The some other disadvantages are suitable only for real valued space .It allow only 2 classes for classification using binary method and several strategies for multiple class classification. For user its very hard to understand Hyper plane .

Density-based clustering

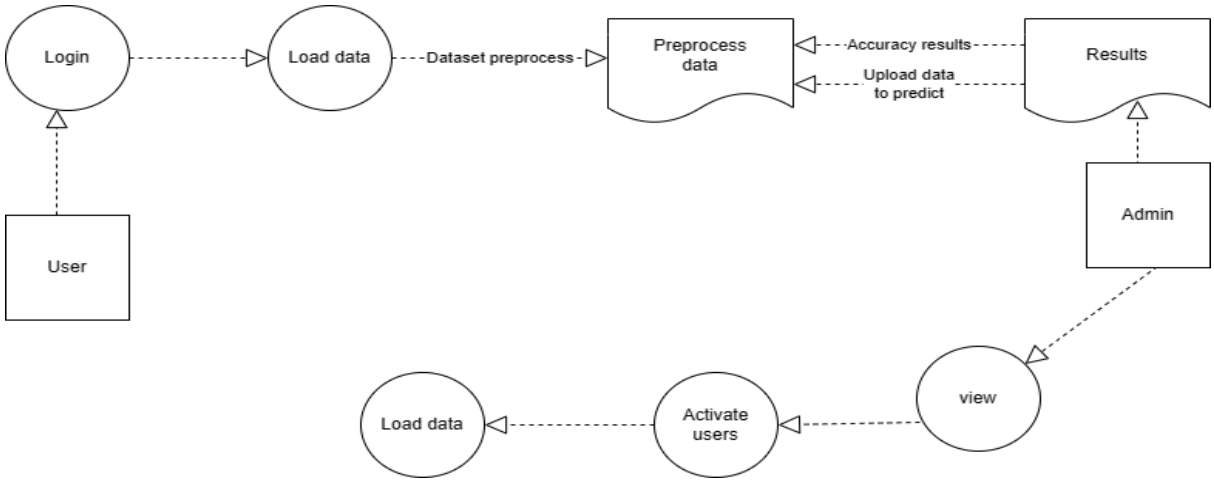
The density based clustering algorithm falls under data clustering algorithm: A space is considered with given set of points it groups together points that are closely packed together i.e., points with many closely neighbours. The most common clustering algorithms and also most cited for scientific literature is density based algorithm. It is opposite to k-means, using an R*tree. In Density based clustering algorithm ,an unassigned object is chosen from the given data set classification method like Hierarchical multiple classifier is used classify the given dataset. Thus it is an efficient way to classify an data with accurate information in reduced time and cost[13].

VII. PROPOSED WORK

The thyroid Dataset is taken from UCI data repository site. The Database consists of thyroid patient records. The Patients record is having different attributes described in the data set description and different data mining techniques are applied to get the predication of thyroid disease and then Linear regression is performed to obtain the which hormone among TSH,T3,TT4 affect the male and female. And also which among the TSH, T3,TT4 influence the hypothyroidism and hyperthyroidism. In healthcare services data mining technique is mainly used for making decision, disease diagnosing and giving better treatment to the patients at corporately low cost. Classification of thyroid disease plays is an important task in the prediction of disease. Dimensionality reduction may be done as a future work so that number of blood test the thyroid will be reduced and also time required diagnosing disease.

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the user. Once admin activated the user then user can login into our system. User can upload the dataset based on our dataset column matched. For algorithm execution data must be in float format. Here we took Dataset they used for this research is taken from the UCI Machine Learning Repository. for testing purpose. User can also add the new data for existing dataset based on our Django application. User can click the Classification in the web page so that the data calculated Accuracy based on the algorithms. User can click Prediction in the web page so that user can write the review after predict the review That will display results depends upon review like postive, negative or neutral.



Admin:

Admin can login with his login details. Admin can activate the registered users. Once he activate then only the user can login into our system. Admin can view the overall data in the browser. Admin can click the Results in the web page so calculated Accuracy based on the algorithms is displayed. All algorithms execution complete then admin can see the overall accuracy in web page.

Data Pre-processing:

A dataset can be viewed as a collection of data objects, which are often also called as a records, points, vectors, patterns, events, cases, samples, observations, or entities. Data objects are described by a number of features that capture the basic characteristics of an object, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as variables, characteristics, fields, attributes, or dimensions. The data pre-processing in this forecast uses techniques like removal of noise in the data, the expulsion of missing information, modifying default values if relevant and grouping of attributes for prediction at various levels.

Selected attribute

Name: age | Type: Numeric
Missing: 1 (0%) | Distinct: 93 | Unique: 5 (0%)

Statistic	Value
Minimum	1
Maximum	455
Mean	51.736
StdDev	20.085

Fig 2. Selected Attributes

Machine learning:

Based on the split criterion, the cleansed data is split into 60% training and 40% test, then the dataset is subjected to four machine learning classifiers such as Machine learning plays an important role in predicting diseases. algorithms named Support Vector Machine(SVM), Decision Tree(DT), Random Forest(RF), Logistic Regression(LR) and Naive Bayes(NB). The accuracy a of the classifiers was calculated and displayed in my results. The classifier which bags up the highest accuracy could be determined as the best classifier.

```
Correctly Classified Instances        3481               92.2853 %
Incorrectly Classified Instances       291                7.7147 %
Kappa statistic                          0
Mean absolute error                      0.0729
Root mean squared error                  0.1904
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              3772
```

Fig 3. Classification

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.923      1.000    0.960      ?        0.498     0.923     negative
                 0.000    0.000    ?          0.000    ?          ?        0.493     0.051     compensated_hypothyroid
                 0.000    0.000    ?          0.000    ?          ?        0.486     0.025     primary_hypothyroid
                 0.000    0.000    ?          0.000    ?          ?        0.100     0.001     secondary_hypothyroid
Weighted Avg.    0.923    0.923    ?          0.923    ?          ?        0.498     0.855
```

Fig 4. Accuracy by class in Classification

```
=== Confusion Matrix ===

   a    b    c    d   <-- classified as
3481    0    0    0 |    a = negative
 194    0    0    0 |    b = compensated_hypothyroid
  95    0    0    0 |    c = primary_hypothyroid
   2    0    0    0 |    d = secondary_hypothyroid
```

Fig 5. Confusion matrix

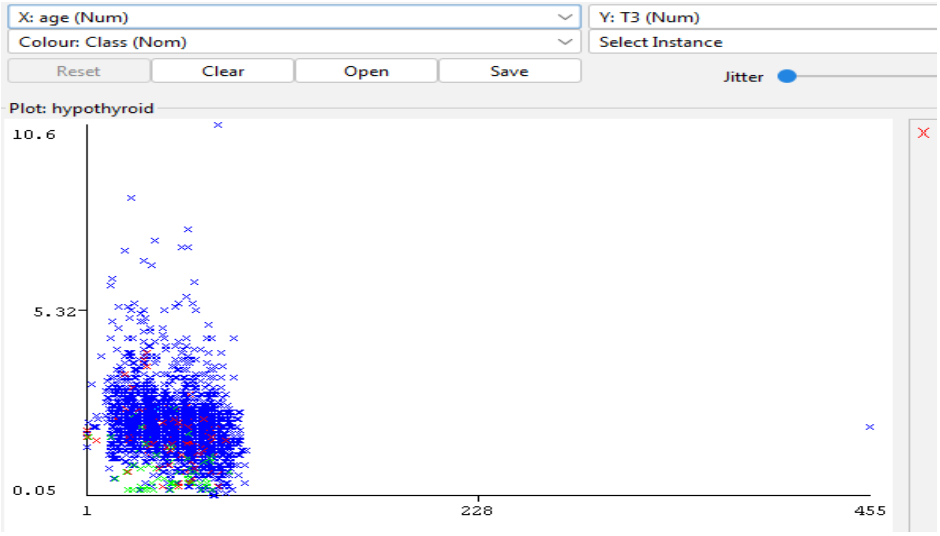


Fig 6. Visualization of plot matrix

```
Correctly Classified Instances        3600               95.4401 %
Incorrectly Classified Instances       172                4.5599 %
Kappa statistic                          0.6197
Mean absolute error                      0.0351
Root mean squared error                  0.1353
Relative absolute error                 48.1912 %
Root relative squared error             71.0671 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.992    0.467    0.962      0.992    0.977      0.654    0.938     0.993     negative
                 0.335    0.006    0.756      0.335    0.464      0.487    0.910     0.565     compensated_hypothyroid
                 0.832    0.003    0.868      0.832    0.849      0.846    0.996     0.874     primary_hypothyroid
                 1.000    0.001    0.400      1.000    0.571      0.632    1.000     0.583     secondary_hypothyroid
Weighted Avg.    0.954    0.432    0.949      0.954    0.947      0.650    0.938     0.968

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3454   14   10    3 |    a = negative
  127   65    2    0 |    b = compensated_hypothyroid
    9    7   79    0 |    c = primary_hypothyroid
    0    0    0    2 |    d = secondary_hypothyroid
```

Fig 7. By using Naive Bayes classification

```
Correctly Classified Instances        3772               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0055
Root mean squared error                  0.0249
Relative absolute error                  7.568  %
Root relative squared error             13.0704 %
Total Number of Instances             3772

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     negative
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     compensated_hypothyroid
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     primary_hypothyroid
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     secondary_hypothyroid
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 3481    0    0    0 |    a = negative
    0  194    0    0 |    b = compensated_hypothyroid
    0    0   95    0 |    c = primary_hypothyroid
    0    0    0    2 |    d = secondary_hypothyroid
```

Fig 8. By using Random Forest tree classification

```
Number of clusters selected by cross validation: 11
Number of iterations performed: 2


                          Cluster
Attribute                      0       1       2       3       4       5       6       7       8       9       10
                          (0.07)  (0.09)  (0.14)  (0.07)  (0.03)  (0.14)  (0.15)  (0.06)  (0.08)  (0.09)  (0.07)
=============================================================================================================================
age
  mean                   53.8084 47.1407 51.8509 49.6842 61.9617 62.6699 44.4006 51.4335 39.3797  61.482  46.965
  std. dev.              16.6018 17.9927 18.9323 18.3302 15.5681 24.6491 16.8443 19.2612 16.5418 14.6802  19.449

sex
  F                     259.3569 224.8736 496.2405 16.5668 94.8961 534.5378 338.8497 197.8012 268.3926  3.2675 206.2172
  M                      22.6399 128.3807 26.3293 252.9731 36.4163  3.4423 238.9883 31.9267 23.9932 346.734 41.1761
  [total]               281.9968 353.2543 522.5698  269.54 131.3125 537.9801 577.838 229.7279 292.3858 350.0015 247.3933
on thyroxine
  f                      10.7374 313.5647 487.6185 239.0349 124.5366 523.2914 566.6008 218.9848 267.0289 343.212 224.3901
  t                     271.2594 39.6896 34.9513 30.5051  6.7759 14.6887 11.2372 10.7431  25.357  6.7895 23.0032
  [total]               281.9968 353.2543 522.5698  269.54 131.3125 537.9801 577.838 229.7279 292.3858 350.0015 247.3933
query on thyroxine
  f                     279.4309 335.1762 519.0453 261.1025 126.7686 531.9535 576.7092 225.0578  287.26 344.1028 246.3933
  t                       2.5659 18.0781  3.5245  8.4375  4.5438  6.0266  1.1288  4.6702  5.1258  5.8987       1
  [total]               281.9968 353.2543 522.5698  269.54 131.3125 537.9801 577.838 229.7279 292.3858 350.0015 247.3933
on antithyroid medication
  f                     278.5398 350.3786 504.8815 262.5372 130.2459 536.9413 574.7533 228.6239 282.7092 348.9961 241.3933
  t                        3.457  2.8757 17.6883  7.0027  1.0665  1.0388  3.0847  1.1041  9.6766  1.0054       6
  [total]               281.9968 353.2543 522.5698  269.54 131.3125 537.9801 577.838 229.7279 292.3858 350.0015 247.3933
sick
  f                     279.5796 345.8339 519.4476 265.762 27.7723 534.8971 576.5614 218.4573 286.3561 344.9393 236.3933
  t                       2.4172  7.4204  3.1222  3.778 103.5401  3.0831  1.2767 11.2706  6.0297  5.0622      11
```

Fig 9. By cross validation attributes

VIII. CONCLUSION

We see that the feature selection technique RFE helps us to get better accuracy with all other classifiers. In our findings, we have seen that RFE significantly helps us to predict hypothyroid in the primary stage by using a real-time dataset. It is very difficult for us to collect data in this current pandemic situation. As a result, we have collected only 519 data. So, considering the situation and the constraint we couldn't study on a larger dataset. In our study, we have seen that there have not been done any work in thyroid based on Bangladesh before. We have a limitation of data to work with. So, in the future, we want to work with a larger dataset and we hope that more people from our country will show interest to work on this disease that will help us to find a better solution and able to predict disease in the primary stage with better accuracy. Hope that will help the people of our country to maintain a healthy society. Thus this work is need full to identify how to predict the thyroid disorder at earlier stage using data mining techniques. Data mining classification algorithms are used to diagonise the thyroid problems and gives different level of accuracy for each techniques. These techniques help to minimize the noisy data of the patient's data from the data bases. Data mining Algorithms such as KNN, Naïve bayes, Support vector machine, ID3are considered for the study. These various algorithm results are based on speed, accuracy and performance of the model and cost for the treatment. Also these classifications of effective data are helps to find the treatment to the thyroid patients with better cost and facilitate the management.

IX. REFERENCES

[1]. A. M. Amiri, and G. Armano, "Early Diagnosis of Heart Disease Using Classification And Regression Trees", In The 2013 International Joint Conference on Neural Networks, pp. 1-4, 09 January, 2014.

[2]. A. K. Aswathi, and A. Antony, "An Intelligent System for Thyroid Disease Classification and Diagnosis", 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT2018), pp. 1261-1264, 27 September, 2018.

[3]. A. Begum, and A. Parkavi, "Prediction of thyroid Disease Using Data Mining Techniques", 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 342- 345, 06 June,2019.

[4]. K. Pavya, and B. Srinivasan, "FEATURE SELECTION ALGORITHMS TO IMPROVE THYROID DISEASE DIAGNOSIS", IEEE International Conference on Innovations in Green Energy and Healthcare Technologies(ICIGEHT'17), pp. 1-5, 02 November, 2017.

[5]. F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM",3rd International Conference on Bioinformatics and Biomedical Engineering, pp. 1-4, 14 July, 2009.

[6]. Q. Pan, Y. Zhang, M. Zuo, L. Xiang, and D. Chen, "Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest", 8th International Conference on Information Technology in Medicine and Education, pp 567-571, 13 July, 2017.

[7]. 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, 27 June, 2019.

[8]. S. Dash, M. N. Das, and B. K. Mishra, "Implementation of an Optimized Classification Model for Prediction of Hypothyroid Disease Risks",International Conference on Inventive Computation Technologies(ICICT) ,pp. 1-4, 19 January, 2017.

[9]. Ravindra Changala, "Development of Predictive Model for Medical Domains to Predict Chronic Diseases (Diabetes) Using Machine Learning Algorithms and Classification Techniques", ARPN Journal of Engineering and Applied Sciences, VOL. 14, NO. 6, March 2019, ISSN 1819-6608.

[10]. G.Vinoda Reddy,"A Review on Machine Learning Algorithms and Classification Techniques in Diabetes Medical Diagnosis and Healthcare Systems", Journal of Emerging Technologies and Innovative Research (JETIR),June 2022, Volume 9, Issue 6,ISSN-2349-5162

[11]. K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, V. H. C.D. Albuquerque, "Optimal feature-based multi-kernel SVM Approach for thyroid disease classification", Springer Science +Business Media, LLC, part of Springer Nature 2018, pp. 1128-1143, 2 July, 2018.

[12]. M. R. N. Kousarrizi, F.Seiti, and M. Teshnehlab, "An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification", International Journal of Electrical & Computer Sciences IJECS-IJENS, pp. 13-19, February, 2012.

[13]. S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches", 16th International Bhurban Conference on Applied Sciences & Technology(IBCAST), pp. 619-623, 18 March, 2019.

[14]. 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, 09 April 2020.

[15]. Ravindra Changala, "Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods In Text Mining" in ARPN Journal of Engineering and Applied Sciences,Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.

[16]. Roshan Banu D, and K.C.Sharmili "A Study of Data Mining Techniques to Detect Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2017.

[17]. Irina IoniŃă and Liviu IoniŃă" Prediction of Thyroid Disease Using Data Mining Techniques"The Classification Technique for Talent Management using SVM, (ICCEET), 978-1- 4673-02l0-4/12, pp. 959- 963, 2017.

[18]. Khushboo Taneja, Parveen Sehgal, Prerana "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network" International Journal of Research in Management, Science & Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016.

[19]. Hanung Adi Nugroho, Noor Akhmad Setiawan, Md. Dendi Maysanjaya," A Comparison of Classification Methods on Diagnosis of Thyroid Diseases"IEEE International Seminar on Intelligent Technology and Its Applications,2017.

[20]. M.N Das ,Brojo Kishore Mishra ,Shreela Dash," Implementation of an Optimized Classification Model for Prediction of Hypothyroid Disease Risks" 2017 IEEE Conference on Big Data and Analytics (ICBDA).

[21]. Ravindra Changala, "Statistical Models in Data Mining: A Bayesian Classification" in International Journal of Recent Tredns in Engineering & Research (IJRTER), volume 3, issue 1, pp.290-293. in 2017.

[22]. Vinoda Reddy, "Recurrent Feature Grouping and Classification for action model prediction in CBMR", International Journal of Data Management and Knowledge Process, Vol.7,

No.5/6, November 2017, pp. 63 74.http://dx.doi.org/10.5121/ijdkp.2017.7605.

[23]. Rasitha Banu,G.— "Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data MiningTechnique", IJTRA 7Izdihar Al-muwaffaq and Zeki Bozkus" MLTDD : Use of machine Learning Techniques For Diagnosis Of Thyroid Gland Disorder" Proceeding of Annual International Conference.

Cite this article as :

Annapurna Gummadi, D. Rammohan Reddy, "A Novel Machine Learning Framework for Prediction of Early-Stage Thyroid Disease Using Classification Techniques", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 3, pp. 467-479, May-June 2022. Available at doi : https://doi.org/10.32628/IJSRST229398
Journal URL : https://ijsrst.com/IJSRST229398