2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) | 978-1-6654-1460-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT4SD50815.2021.9397052

Prediction Of Thyroid Disease(Hypothyroid) In Early Stage Using Feature Selection And Classification Techniques

Md Riajuliislam, Khandakar Zahidur Rahim, Antara Mahmud
Department of Computer Science and Engineering
Daffodil International University, Dhaka, Bangladesh
*riajul69990@gmail.com, kzrhridoy@gmail.com, antara.cse@diu.edu.bd

***Abstract*—Thyroid disease is one of the most common diseases among the female mass 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 feature selection techniques have facilitated this process prediction and assumption of diseases. There are two types of thyroid diseases namely 1. Hyperthyroid and 2.Hypothyroid. Here, in this paper, we have attempted to predict hypothyroid in the primary stage. To do so, we have mainly used three feature selection techniques along with diverse classification techniques. Feature selection techniques used by us are Recursive Feature Selection(RFE), Univariate Feature Selection(UFS) and Principal Component Analysis(PCA) along with 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 the RFE feature selection technique helps us to provide constant 99.35% accuracy for all four classification algorithms. Thus it's deduced from our research that RFE helps each classifier to attain better accuracy than all the other feature selection methods used.**

Keywords—Thyroid disease , Data mining , Feature selection , Recursive Feature Selection , Machine learning , Classification

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.

Nowadays, machine learning has become an immensely popular medium for detecting various diseases. It is very convenient and effective to presume diseases using machine learning techniques. Here, we have used feature selection and classification techniques to predict hypothyroid in the primary stage. We collected data from a registered diagnosis center in Dhaka, Bangladesh. Overall, we have collected a good number of data with a total of 9 attributes. Among these data, 77% are of females whereas the rest are males. We mainly use three feature selection techniques named Recursive Feature Elimination (RFE), Univariate Feature Selection (UFS) and Principal Component Analysis (PCA) along with distinct classification algorithms such as Support Vector Machine (SVM), Decision Tree, Logistic Regression (LR), Random Forest (RF) and Naive Bayes (NB). We finally deduced that the RFE feature selection technique helps us to attain better accuracy with any classification method used.

II. LITERATURE REVIEW

Our approach fundamentally proposes a model to detect hypothyroidism in the primary stage using the Feature selection technique and classification for prediction of hypothyroidism. Various related methodologies have been found in the past few years and some of them are discussed here.

In [1], the authors proposed an Early Diagnosis of Heart Disease Using Classification And Regression Trees. In this proposed work their ultimate goal is to implement a heart diagnosis system to reducing the number of unnecessary echocardiograms and of preventing the release of newborns that are in fact affected by heart disease. And in this work, they analyze PCG(phonocardiograms) signals by using Classification & Regression Tree(CART). In classification, they perform Feature extraction in time & frequency. Also, they use k means clustering. A CART regression tree is a binary decision tree that is constructed by either splitting each node on the tree into two daughter nodes. They achieve 99.14% accuracy, 100% sensitivity, and 98.28% specificity were obtained on the dataset used for experiments.

This work investigated An Intelligent System for Thyroid Disease Classification and Diagnosis [2]. They proposed a method for classification & diagnosis of thyroid disease to detect its early stage using Weighted SVM classification & to optimize SVM parameters such as TSH, T3, T4 they use particle swarm optimization. Also, they use KNN to approximate the missing value via user input.

In [3], the authors proposed work constructed the Prediction of Thyroid Disease Using Data Mining Techniques. It proposes a method for finding higher accuracy of predicting thyroid disease at a very early stage using a classification algorithm such as decision tree C4.5 & ID3 algorithm, KNN, SVM, Naive Bayes.

In [4], the authors proposed a work which identified Feature Selection Algorithms To Improve Thyroid Disease Diagnosis. The main purpose of this research 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. Four classifiers also used such as Multilayer Perceptron, Back Propagation Neural Network, Support Vector Machine, and Extreme Learning Machine. The wrapper-based algorithm produced maximum efficiency and produced a maximum accuracy of 98.14% with an ELM classifier.

In [5], the authors proposed work constructed a method of Thyroid Disease Diagnosis Based on Genetic Algorithms using PNN and SVM. This framework proposed a method to separate hypothyroid & hyperthyroid for diagnostics using Support Vector Machines(SVM) and Probabilistic Neural Network(PNN) for classification. For feature selection, they used a genetic algorithm. There accuracy using SVM&PNN with GA (FS) is 100%

In [6], the auhtors proposed work is Improved Ensemble Classification Method of Thyroid Disease Based on Random Forest. This work proposed a new method for thyroid disease classification based on random forest. They have used a random forest-based ensemble classifier method which achieves 96.16% accuracy.

Interactive Thyroid Disease Prediction System Using Machine Learning Technique [7]. The dataset from the UCI repository has been used for classification. In this proposed work they use Machine Learning Algorithms such as SVM(99.63), K-NN(98.62), Decision Trees(75.76), ANN(97.5) were used to predict the estimated risk on a patient's chance of obtaining thyroid disease.

In this work, focuses on the survey of the diagnosis of thyroid disorder [8] using Ranker Search as a Feature selection algorithm & Naive Bayes as a classifier algorithm which gives 95.38% accuracy.

This paper constructed Classification of Hypothyroid Disorder using Optimized SVM Method [9]. In this work, they proposed a method of detecting hypothyroid disorder level using classification machine learning techniques, namely KNN (K-Nearest Neighbor), SVM (Support Vector Machines), LR (Logistic Regression), and NN (Artificial Neural Network). Logistic Regression method achieved 96.08% accuracy among the other three classifiers but SVM provides the highest accuracy of 99.08% after standardizing the data and parameter tuning.

In [10] this work, they proposed a model to classify this thyroid data utilizing optimal feature selection and kernel-based classifier process. The novelty and objective of this proposed model as feature selection, it's used to enhance the performance of the classifying process with the help of improved gray wolf optimization. In this technique, MKSVM is used to distinguish the thyroid illness with high accuracy of 98.65%.

In [11] this work, the classification of thyroid disease which is one of the most important classification problems has been proposed. Two thyroid gland hypothyroid & hyperthyroid which is responsible for the metabolism of the body has been classified by using feature selection or extraction as preprocess such as Sequential forward selection and sequential backward selection & GA(Genetic Algorithm). SVM is used as a classifier to separate the thyroid disease. This study is based on datasets UCI machine learning repository & the second one is the real data which has been gathered by the Intelligent System Laboratory of the K.N.Toosi University of Technology from Imam Khomeini hospital.

In [12] this work, heart disease prediction in the early stage has been proposed as it already been done in the early stage but this paper increase the accuracy of prediction by using feature selection techniques as Rapid miner tool & algorithms are Decision Tree, Logistic Regression, Logistic Regression SVM, Naïve Bayes, and Random Forest & accuracy are 82.22%, 82.56%, 84.17%,84.24% and 84.85. A study has been applied to the datasets taken from the UCI data set.

In [13], the authors work proposed a model of thyroid disease diagnosis using feature selection techniques such as Univariate Selection(UFS), Recursive Feature Elimination(RFE), and Tree-Based Feature Selection and classification algorithm are Naive Bayes, Support Vector Machine(SVM) & Random Forest. Among them, SVM along with RFE provides them with the highest accuracy of 92.92%.

Dataset they used for this research is taken from the UCI Machine Learning Repository.

III. METHODOLOGY

In machine learning, there is a saying that if you input a junk value you will only get junk value in return. By using machine learning algorithm to predict something if the data set contains noisy data which is not important, as a result, it hampers the performance of algorithms to achieve the highest accuracy. To achieve the highest accuracy in the algorithm we want to feed those features which are really important and this is done by using the feature selection technique. In the first step, we have collected hypothyroid data from the registered diagnostic center then clean the data. In the second step, we applied feature selection in our dataset to find important attributes and the feature selection technique is RFE, UFS, PCA. In 3rd step-based, by using those feature selections we individually measure the performance of each algorithm. We studied our dataset based on these classification algorithms- Support Vector Machine(SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) & Naive Bayes (NB). The framework is shown in fig.1.

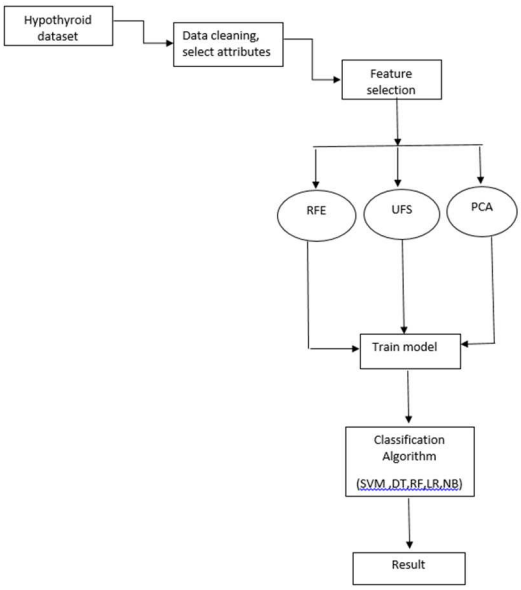

Fig-1: Data flow of the model

A. Dataset Description

In the pandemic situation in 2020 data collection was very tough job for us. We collected dataset from registered diagnostic center Dhaka, Bangladesh. The total number of data we collected are 519 with 9 attributes. Dataset contains the following attributes in the table.1-

Attributes	Type	Description
ID	Continuous	Patients ID
Age	Continuous	In years
Sex	Male , Female	Gender
FT3	Continuous	Free Triiodothyronine value
FT4	Continuous	Free Thyroxin value
T3	Continuous	Triiodothyronine value
T4	Continuous	Thyroxin value
TSH	Continuous	Thyroid Stimulating Hormone value
Result	categorical	0/1

Table-1: Attributes of Hypothyroid Dataset

B. Feature Selection Technique

The process of feature selection is to automatically select those features which are significantly important to help in predicted the output or variables we are interested in. There are some data that lies in our dataset that significantly decrease the accuracy of our model. And to eliminate these unwanted data feature selection technique plays an important role. The benefit of feature selection is-

1.**Reduction in Overfitting**- It makes the data less unnecessary as a result it maximize the possibility of making a decision based on relevant features.

2. **Improvement in Accuracy**- It purifies our data to make it less misleading to improve the model accuracy.

3. **Reduction in Training Time**- Eliminating unnecessary data means reducing the time to train the algorithm and its complexity to train it faster.

C. Method of Feature Selection

C1. Recursive Feature Elimination (RFE):

Recursive feature elimination (RFE) is a method of feature selection that works by fits a model and eliminating the fragile feature and those features are ranked by 'coef_' and 'feature_importances_' attributes. The importance of features is provided by 'fit' method and the least important features are eliminated recursively until it reached the desired features. We performed RFE in several algorithms to see which features are appropriate for a particular algorithm and we find out three important features per algorithm. The estimated accuracy using RFE for each algorithm is SVM(99.35%), Decision Tree(99.35%), Random Forest(99.35%), Logistic Regression (99.35%) and Naive Bayes (94.23%).

Feature Selection Technique	Algorithm	Importance feature
RFE	SVM	T3,T4,TSH
RFE	Decision Tree	T3,T4,TSH
RFE	Random Forest	Age,FT4,TSH
RFE	Logistic Regression	FT3,T3,TSH
RFE	Naïve Bayes	T3,T4,TSH

Table-2: RFE Feature Selection

C2. Univariate Feature Selection(UFS):

UFS is another feature selection method that uses 'SelectKBest' with a statistical test chi-squared test(score_func=chi2) to find out the highest-scoring features. Statistical test chi-squared test(score_func=chi2) measure the strength of the relationship of feature individually in accordance with the response variable. This method finds out three important features from our dataset. The estimated accuracy using UFS for each algorithm is SVM (98.71%), Decision Tree (99.35%), Random Forest (99.35%), Logistic Regression (99.35%) and Naive Bayes (96.79%).

Feature Selection Technique	Algorithm	Importance feature
UFS	SVM , Decision Tree, Random Forest, Logistic Regression, Naïve Bayes	Age,T4,TSH

Table-3: UFS Feature Selection

C3. Principal Component Analysis(PCA):

PCA is basically called a data reduction technique which is a very important feature selection that converts the high-dimensional data into low dimensional to select the most important feature that can capture the maximum information about the dataset. Important features are ranked by the 'explained_variance_ratio_' attribute and the feature that causes the highest variance in PCA consider as the first principal component and the feature that causes the second variance to consider as the second principal component and so on. The estimated accuracy using PCA for each algorithm is SVM(89.74%), Decision Tree (87.17%), Random Forest (88.46%), Logistic Regression (89.74%) and Naive Bayes (89.74%).

Feature Selection Technique	Algorithm	Importance feature
PCA	SVM , Decision Tree, Random Forest, Logistic Regression, Naïve Bayes	Age,FT3,FT4

Table-4: PCA Feature Selection Algorithm

IV. RESULT ANALYSIS

We applied three feature selection methods in our model to predict thyroid disease(hypothyroid). It also showed that by using a machine learning algorithm we can also predict hypothyroid in a very early stage. We applied RFE, UFS and PCA feature selection to find out the important attribute that will help to better the performance of the algorithm and which feature selection technique is best for our model. According to table-5, we can see that the RFE feature selection technique helps algorithms by selecting suitable attributes for those. As a result, the RFE feature selection technique performs better with constant 99.35% accuracy with four algorithms. On the other hand, PCA is giving the lowest accuracy of these algorithms and it has a wide difference inaccuracy.

Serial Number	Algorithm	Feature Selection Technique (RFE) Accuracy%	Feature Selection Technique (UFS) Accuracy%	Feature Selection Technique (PCA) Accuracy%
1	SVM	99.35%	98.71%	89.74%
2	Decision Tree	99.35%	99.35%	87.17%
3	Random Forest	99.35%	99.35%	88.46%
4	Logistic Regression	99.35%	99.35%	89.74%
5	Naïve Bayes	94.23%	96.79%	89.74%

Table-5: Result Analysis

V. 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.

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 (ICICCT 2018), 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] V. S. Vairale, and S. Shukla, "Classification of Hypothyroid Disorder using Optimized SVM Method", Second International Conference on Smart Systems and Inventive Technology (ICSSIT 2019), pp. 258-263, 10 February, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Dhaka Tribune(2018), 50 million people suffer from thyroid disease in Bangladesh. Available: https://www.dhakatribune.com/feature/health-wellness/2018/05/25/experts-50-million-people-suffer-from-thyroid-disease-in-bangladesh/.