ISSN: 2366-1313

THYROID DISEASE PREDICTION USING PRINCIPAL COMPONENT ANALYSIS WITH MACHINE LEARNING ALGORITHMS

[1]N.SUDHA RANI, [2]R.VEERA REDDY

[1]Assistant Professor, Dept. of CSE, Megha Institute of Engineering and Technology for Women,
sudhareddy.nareddy93@gmail.com

[2]Assistant Professor, Dept. of CSE, Megha Institute of Engineering and Technology for Women,
veerareddy.redabothu@gmail.com

Abstract: *The Thyroid disease is a vascular disease and one of the most important organs of a human body. This gland secretes two hormones which help in controlling the metabolism of the body. The two types of Thyroid disorders are Hyperthyroidism and Hypothyroidism. When this disorder occurs in the body, they release certain type of hormones into the body which imbalances the body's metabolism. Thyroid related Blood test is used to detect this disease but it is often blurred and noise will be present. Data cleansing methods were used to make the data primitive enough for the analytics to show the risk of patients getting this disease. Machine Learning plays a very deciding role in the disease prediction. Feature selection techniques used by us Principal Component Analysis (PCA) along with classification Machine Learning algorithms, SVM - support vector machine, Random Forest, Decision tree, Logistic regression, Naïve Bayes are used to predict the patient's risk of getting thyroid disease. Web app is created to get data from users to predict the type of disease.*

Keywords: *Thyroid disease, Machine learning algorithm, Feature selection, Disease prediction, support vector machine.*

I. INTRODUCTION

The computational biology of evolution is used in the healthcare business. It allows the collection of stored data of infected people to predict the disease. There are predictive algorithms available for early diagnosis of the disease. Scientific information systems are rich in datasets, but there are only a few intelligent structures that can easily diagnose the disease. Over time, system control algorithms have begun to play an important role in resolving complex and nonlinear problems within the development version. In any case, the

ISSN: 2366-1313

prediction model is used to override features selected from specific datasets that can be used to classify healthy people as accurately as possible [1].

If this is not done, incorrect classification can cause an infected healthy person to receive unnecessary treatment. The thyroid gland is the gift of the endocrine gland under the human apple in the human neck, which helps release thyroid hormone, which affects the rate of metabolism and protein synthesis. Thyroid hormones help determine how fast the heartbeats and how fast we burn energy. The thyroid releases various active hormones called levothyroxine (T4) and triiodothyronine (T3). These hormones help control body temperature. They also contribute to the production and transmission of energy to all body parts and play an important role in protein control. Iodine is considered the primary building block of the thyroid gland. He prostrates on certain issues. An inadequate supply of these hormones can lead to hyperthyroidism. There are many causes associated with hyperthyroidism and dysfunctional thyroid. There are several types of medications, such as thyroid surgery, ionizing radiation, chronic thyroid allergy, iodine deficiency, and enzyme deficiency, to produce thyroid hormones [2].

In the current state, the thyroid is one of the most important diseases, and it has the potential to become a common disease for different women. According to experts, 50 million people in Bangladesh suffer from thyroid disease. Of these, girls are 10 times more likely to develop thyroid disease. Although most of the 50 million people with thyroid disease suffer from it, about 30 million are unaware of it. The Bangladesh Endocrine Society (BES) estimates that about 20-30% of women suffer from thyroid disease.

The thyroid is a gland located in our frame's middle of the neck. It is shaped like a butterfly and is small in length. It releases many hormones that travel with the blood to the body for manipulation in various sports. The thyroid hormone maintains metabolism, sleep, growth, sexual function, and mood. Depending on the release of the thyroid hormone, we may feel tired or stressed and lose weight. Triiodothyronine (T3) and thyroxine (T4) are the main thyroid hormones. Both of these hormones are primarily responsible for maintaining energy in our bodies. The pituitary gland produces thyroid-stimulating hormone (TSH), which makes it easier for the thyroid gland to secrete T3 and T4. There are two common thyroid diseases: 1) hypothyroidism, and 2) hyperthyroidism [3].

ISSN: 2366-1313

The thyroid disease analysis and treatment concept is presented through the deliberate practice of thyroid disease and is important in most thyroid diseases. Types of thyroid disease are based on euthyroidism hyperthyroidism and hypothyroidism, which indicate regular, excessive or poor levels of thyroid hormones. Rural euthyroidism represents normal production of normal thyroid hormone and normal levels through the thyroid gland at the cellular stage. National hyperthyroidism is a medical symptom due to the excessive flow of intracellular thyroid hormones. Hypothyroidism is on the rise due to a lack of thyroid hormone technology and negative change therapy.

II. REVIEW OF LITERATURE

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.

Ankita Tyagi et al.[4] This work used unusual category algorithms: decision tree, support vector machine, artificial neural network, and k-nearest neighbour algorithm. Based on a set of data from the UCI repository, type and prediction were completed, and accuracy was obtained based on the output produced. They analysed the accuracy of the algorithms used and created a contrast to find a good technique with high accuracy.

Sunila Godara et al.[5] They have used logistic regression and SVM device learning techniques to investigate thyroid datasets. Comparisons between these algorithms are mainly based on accuracy, recovery, F-measure, ROC, and RMS errors. Logistics regression turned out to be a satisfactory rating.

YongFeng Wang et al.[6] The thyroid nodule is identified by ultrasound imaging of the thyroid with complete information of imaging evolution radionics and basic procedures for benign or malignant type. A comparison is made between the two tactics. The accuracy, sensitivity and specificity of the application type of radioman approach are 66.81%, 51.19% and 75.77%, respectively. The evaluation rates for in-depth knowledge of the technique based on the test samples are 74.69% and 63.10.% and 80%.respectively. Gain a deep understanding of growth to become the best strategy.

Hitesh Garg. [7] The feed-forward neural network is used to extract and isolate the features of ultrasound images for tumour prognosis. Accuracy and various factors

ISSN: 2366-1313

were measured, and all common values were above 86%.

In [8], The authors suggest an early diagnosis of heart disease using classified trees and regression. This proposed work its ultimate goal is to implement a cardiac prognosis system to reduce the large variety of useless echocardiograms and to prevent the discharge of new-borns who may be affected by coronary heart disease. And in this work, they analyse PCG signals using a classification and regression tray (CART). In class, they draw features in time and frequency. Also, they use K-mean classification. A cart regression tree is a binary selection tree made by dividing each node in a tree into child nodes. They achieved 99.14% accuracy, 100% sensitivity and 98.28% specificity on the data set used for the experiments. The work explores an intelligent system for classifying and diagnosing thyroid diseases. They suggested a technique for diagnosing the type and diagnosis of thyroid disease using the weighted SVM class to detect its early stages and improve SVM parameters, including TSH, T3, and T4; they used particle congestion. Use optimization. In addition, they use KNN to estimate missing costs through customer information.

In [9], the authors proposed built-in paintings that predict thyroid disease using data mining techniques. C4.5 suggests finding the best accuracy for early-stage thyroid disease prediction using a set of principles such as selection tree and ID3, KNN, SVM, and Naïve Bayes algorithms.

III. PROPOSED METHODOLOGY

There may be an announcement in machine learning that if we enter a junk value, we will only receive a junk value in return. Using machine learning algorithms to predict something if the dataset contains noisy data that is not critical, as a result of which, in the algorithm's performance to achieve the best accuracy. There is an obstacle. To get the best accuracy in the algorithm, we need to feed the important features and that have been done using the feature choice technique. We collected hypothyroidism information from a registered diagnostic centre and then cleared the statistics in the first stage. In the second step, we apply feature selection to our dataset to find the required attributes, and the feature selection technique is RFE, UFS, and PCA. In the third step, using these feature options, we rate the performance of each algorithm. We studied our dataset based on these categories of algorithms: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression

ISSN: 2366-1313

(LR), and Naive Bayes (NB). The framework has been showed below figure.1.

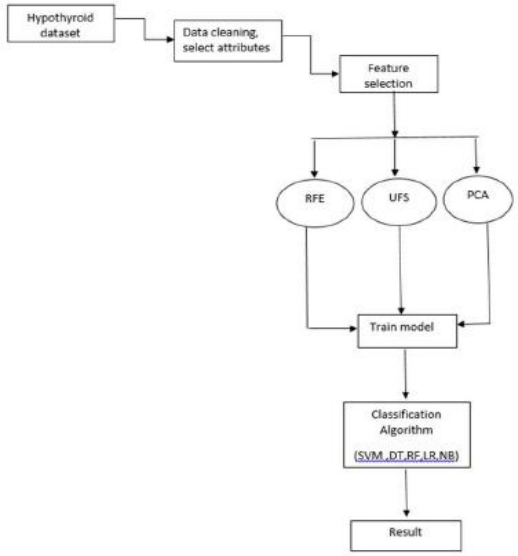

Fig.1 Proposed work data flow model

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

Table.1 Attributes of Hypothyroid Dataset

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

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 Over fitting- 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.

ISSN: 2366-1313

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

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 considers 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%).

Table.2 PCA Feature Selection Algorithm

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

IV. RESULTS AND DISCUSSIONS

We applied PCA 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 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-3, we can see that the PCA feature selection technique helps algorithms by selecting suitable attributes for those. As a result, the PCA feature selection technique performs better with constant 99.35% accuracy with four algorithms.

Table.3 Result Analysis for PCA

Algorithm	**Feature selection technique using PCA**
SVM	98.71%
Decision Tree	99.35%
Random Forest	99.35%
Logistic Regression	99.35%
Naïve Bayes	96.79%

ISSN: 2366-1313

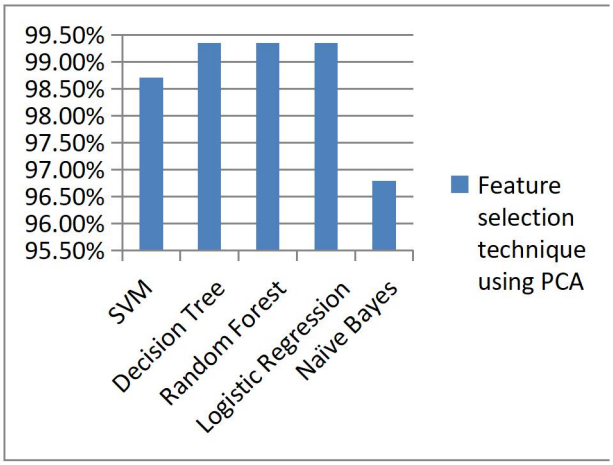

Fig.2 Result analysis

V. CONCLUSION

We implement 3 feature selection methods in our version to anticipate thyroid diseases (hypothyroidism). It also showed that we could expect hypothyroidism at an early stage by using a tool that learns about a set of rules. We implement RFE, UFS, and PCA features that are selected to discover key features along the way to help improve the performance of the set of rules and who to choose for our release. The feature selection method is first class. According to Table 5, we can see that the selection method of RFE functions enables the algorithm to select the appropriate attributes for them. As a result, the RFE feature works best with the Selection Approach 4 algorithm with a regular accuracy of 99.35%.

REFERENCES

[1] Tyagi A, Mehra R , 2018, "Interactive thyroid disease prediction system using machine learning technique", pp.20–22.

[2] YongFeng Wang, 2020, "Comparison Study of Radiomics and Deep-Learning Based Methods for Thyroid Nodules Classification using Ultrasound Images" published on IEEEAccess.

[3] Sunila Godara, 2018, "Prediction of Thyroid Disease Using Machine Learning Techniques" published on IJEE

[4] Sidiq U, Khan RA, 2019, "Diagnosis of various thyroid ailments using data mining classification techniques", pp.2456–3307

[5] Gorade SM, Purohit P, 2017, "A study of some data mining classification technique". Int Res J Eng Technol 4(4):3112–3115.

[6] S. Godara and R. Singh,(2016) "Evaluation of Predictive Machine Learning Techniques as Expert Systems in Medical Diagnosis", Indian Journal of Science and Technology, (Vol. 910).

[7] Sunila, Rishipal Singh and Sanjeev Kumar.(2016) "A Novel Weighted Class based Clustering for Medical Diagnostic Interface." Indian Journal of Science and Technology (Vol 9)

[8] Patel BN, Lakhtaria K, 2012, "efficient classification of data using decision tree". Bonfring Int J Data Min 2(1), pp. 6–12.

[9] Chen Ling, Li Xue, Sheng Quan Z, Peng W-C (2016) Mining health examination

ISSN: 2366-1313

records—a graph-based approach. IEEE Trans Knowl Discov Eng 28:2423–2437.

[10] Hitesh Garg,(2013). "Segmentation of Thyroid Gland in Ultrasound image using Neural Network" published on IEEE.

[11] Prasadu Peddi (2019), Data Pull out and facts unearthing in biological Databases, International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.

[12] Prasadu Peddi (2019), "AN EFFICIENT ANALYSIS OF STOCKS DATA USING MapReduce", ISSN: 1320-0682, Vol 6, issue 1, pp:22-34.