International Journal of Advanced Technology in Engineering and Science
Vol. No. 11, Issue No. 12, December 2023
www.ijates.com
ijates
ISSN 2348 - 7550

Early detection of thyroid disease with an improved random forest classifier with PCA based feature selection

Dr. S Jeyalaksshmi

Associate Professor, Department of BCA and IT,
Vels Institute of Science Technology & Advanced Studies
Email : jlakshmi.scs@velsuniv.ac.in

Abstract:

One of the most vital human organs, the thyroid, may develop vascular disease. The secretion of two hormones by this gland aids in the regulation of the body's metabolism. Hyperthyroidism and hypothyroidism are the two main thyroid diseases. The imbalance of the body's metabolism is brought about by the production of certain hormones when this issue develops in the body. A blood test for thyroid issues may be performed, however the results might be misleading due to the presence of noise and blurriness. To prepare the data for analytics that would reveal the likelihood of patients contracting this condition, data cleaning procedures were used. In the field of illness prediction, machine learning is crucial. In order to determine the likelihood that a patient may develop thyroid illness, we use feature selection methods such as Principal Component Analysis (PCA) in conjunction with classification Machine Learning algorithms, Random Forest, KNN, Support Vector Machine (SVM), Decision Tree, and others. It is possible to anticipate the kind of sickness by collecting user data via a web app.

Keyword: *Thyroid prediction, SVM, RF, DT, KNN*

I. INTRODUCTION

Thyroid disorders are common and diverse because of the powerful hormones secreted by the thyroid gland. Considering that the thyroid controls metabolism, there are a few obvious issues with it, some of which are basic. Many references to iodine atoms were made in connection with the hormones. An overactive thyroid gland causes hyperthyroidism, a condition in which the body produces an excess of hormones. Fatigue and weakness are signs of hypothyroidism, which is characterized by underactive thyroid hormone production. Hypothyroidism manifests itself in a variety of ways, some of which include melancholy, lethargy, ineffectiveness, dry

skin, spasms, extra weight, and sluggishness or exhaustion. Nervousness, shaking, weariness, tremors, anorexia, diarrhea, exophthalmia, and irregular menstruation are all signs of hyperthyroidism. Damage to the thyroid gland or the presence of blocking antibodies are additional risk factors for hypothyroidism.

The key to diagnosing and treating thyroid disease is, in most circumstances, the functional behavior associated with the condition. A thyroid disorder may be classified into three main forms: euthyroidism, hyperthyroidism, and hypothyroidism. These forms denote normal, excessive, and defective levels of thyroid hormones, respectively. When thyroid hormone production and secretion are within normal ranges, a medical condition known as euthyroidism is present. When thyroid hormone levels in the blood and cells are too high, a medical condition known as hyperthyroidism develops. Inadequate thyroid hormone production and treatment choices lead to hypothyroidism.

II. RELATED WORK

In this study, Gyanendra Chaubey et al. Here we are, in the midst of a scenario where models that benefit many sectors of life are being built using machine learning. Because this data is readily available and continues to expand, computer scientists have a better chance of generating predictions and analyzing datasets that can help people's lives. That component is the center of attention here. The classification and forecasting procedures use datasets and algorithms. If a better data set could be organized in real time and more machine learning and deep learning algorithms like SVM, Naive Bayes, auto encoders, ANNs, and CNNs were used, the results may be much better.

Many cherish Bibi Amina Begum [2]. There is a wide range of success rates among data mining classification methods used to categorize thyroid problems.By using these techniques, we can lessen the quantity of irrelevant information from patient data.Data mining methods such as ID3, KNN, Naive Bayes, and Support Vector Machine are considered.The accuracy, performance, speed, and cost of the model and treatment are the four factors that define the various results of the algorithms.Improved cost estimations and better care for thyroid patients may result from this beneficial data categorization.

Studies conducted by Ankita Tyagi and colleagues [3] Finding new machine learning methods that might be used to the diagnosis of thyroid diseases is the objective of our continuing research. The diagnosis of thyroid disease has been improved in recent years with the introduction of many easily accessible algorithms. The research demonstrated that all of the publications use different technology, but to different degrees of accuracy. In comparison to alternative approaches, neural networks have shown superior performance in the majority of cases. This is true even if replacement approaches, such decision trees and support vector machines, have also shown encouraging outcomes. Despite the fact that scientists all around the world have come a long way in their capacity to identify thyroid issues, some have advised patients to cut down on the number of variables they use. Clinical trials need more resources (both time and money) to accommodate more patient characteristics. As a result, algorithms and models for predicting thyroid illness should be created that need minimal patient inputs to provide an accurate diagnosis, thereby minimizing wasted time and money.

III. PROPOSED METHOD

A blood report must be analyzed in order to diagnose and forecast thyroid ailment. To examine a database of thyroid blood test results, we will use several supervised machine learning classifiers approaches. Its accuracy in comparison to the other algorithms will be used to pick the optimal approach for getting the result. For the first analysis, thyroid data is retrieved from the repository at UCI. The dataset makes use of the terms hyperthyroidism and hypothyroidism as labels. You need to make sure this dataset is accurate before you use it for training. Because there may be omissions or unnecessary information in the data, data cleaning is necessary. Data that has been cleaned is used as input by the algorithm for both training and testing purposes. The algorithm has to get attributes from many datasets before it can label the data. The algorithm's predictive abilities are evaluated using test data. By comparing the extracted feature with the training data's properties, the likelihood of the test data may be determined. We will diagnose the patient with hyperthyroidism or hypothyroidism based on the probability value that is greater.

age	sex	on_thyrox	query_on_	on_antithy	sick	pregnant	thyroid_su	I131_treat	query_hyp	query_hyp	lithium	goitre	tumor	hypopituit	psych	TSH_meas	TSH	T3_measu	T3	TT4_meas	TT4	T4U_mea
41	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.3	t	2.5	t	125	t
23	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	4.1	t	2	t	102	f
46	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.98	f	?	t	109	t
70	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.16	t	1.9	t	175	f
70	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.72	t	1.2	t	61	t
18	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.03	f	?	t	183	t
59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	f	?	f	?	t	72	t
80	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.2	t	0.6	t	80	t
66	F	f	f	f	f	f	f	f	f	f	f	f	t	f	f	t	0.6	t	2.2	t	123	t
68	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.4	t	1.6	t	83	t
84	F	f	f	f	f	f	f	f	f	f	f	f	t	f	f	t	1.1	t	2.2	t	115	t
67	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.03	f	?	t	152	t
71	F	f	f	f	t	f	f	f	f	t	f	f	f	f	f	t	0.03	t	3.8	t	171	t
59	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	2.8	t	1.7	t	97	t
28	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	3.3	t	1.8	t	109	t
65	F	f	f	f	f	f	f	t	f	f	f	f	f	f	f	t	12	f	?	t	99	t
42	?	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.2	t	1.8	t	70	t
63	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.5	t	1.2	t	117	t
80	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	6	t	1.6	t	99	t
28	M	f	f	f	f	f	f	f	f	f	f	f	f	f	t	t	2.1	t	2.6	t	121	t
51	F	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.1	f	?	t	130	t
46	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.8	t	2.1	t	108	t
81	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	1.9	t	0.3	t	102	t
54	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	3.1	f	?	t	104	t
55	M	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.2	t	1.8	t	134	t
63	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.03	t	5.5	t	199	t
60	M	t	f	f	f	f	f	f	f	f	f	f	f	f	f	t	13	t	1.4	t	57	t
25	F	f	f	f	f	f	f	f	f	f	f	f	f	f	f	t	0.3	t	3.1	t	129	f

Figure 1. Thyroid Dataset

There is a new thyroid dataset available in the UCI machine learning repository; it has 215 samples and 5 features. A total of five laboratory tests are administered to patients in order to determine whether their thyroid function is normal, hyperthyroid, or hypothyroid. In order to arrive at the diagnosis, all medical records were reviewed. There are 155 instances of normal thyroid function, 35 cases of hyperthyroidism, and 35 cases of hypothyroidism in this sample. Class characteristic, the absorption of T3 resin, total T4, total T3, thyrotropin-releasing hormone (TSH), and the effect of a 200-microgram dose on TSH levels.

We started by collecting data from the hospital, which includes serological tests and pathological observations of TD patients. To find the best classifier for TD prediction, we put the collected data through its paces using five popular classification methods, including ensemble approaches. In order to find the best classifier, we divided the dataset of pathological observations related to TD and used ML algorithms. This way, the patient may avoid spending time and money at the lab for needle tests. Thirdly, we benchmarked five distinct ML algorithms to find the best model for TD prediction using serological testing alone.

Fig. 2. Proposed Flow diagram

i. Support Vector Machine

This supervised learning strategy could be useful for regression and classification. This is accomplished by using hyperplane to partition the training data. Memory efficiency is improved by SVM since it requires a smaller subset of the training data. But it will need more space for storage as the number of attributes increases. This limitation may be circumvented by making use of the kernel function.

ii. Decision Tree

A decision tree is a supervised tree approach that uses a tree diagram to help in decision making. The root node represents output, the internal nodes represent characteristics, and the leaf nodes represent class labels. After considering all of the important elements, a decision tree selects the best branching strategy.

iii. K – Nearest Neighbors

One simple unsupervised learning method is K-Nearest Neighbors (KNN), which infers missing attribute values by comparing test data to an existing data set. Its main applications are in genomics, data reduction, and economic forecasting. The fact that it is easy to implement and tolerant of input data with noise is its primary advantage. Classification is slower, however, since the whole training data set is analyzed. By deferring all computation until the function is evaluated, k-nearest neighbors (k-NN) effectively performs a kind of classification. Normalizing the training data may substantially enhance this method's accuracy in cases when the attributes represent various physical units or come in drastically different sizes. Giving more weight to the contributions of near neighbors than further ones is an approach that might be useful for classification and regression. One common approach to weighing is to give each neighbor a value equal to one-fourth of the distance between them, denoted as d. The neighbors are selected from a pool of instances where the item's class or property value is known in k-nearest neighbor classification and regression. This may be thought of as the training set for the algorithm, even if no training step is actually executed. One unique aspect of the k-NN approach is its consideration of the data's geographical context.

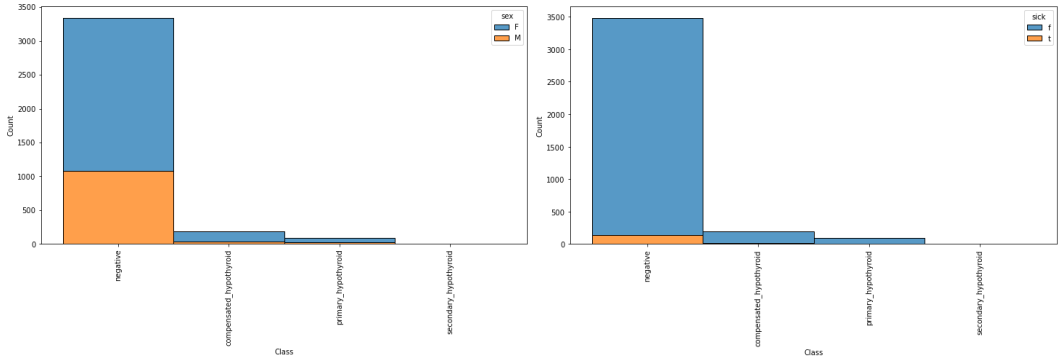
iv. Enhanced Random forest

Research conducted by Liaw et al. (2002) states that a random forest is comprised of several decision trees that collectively make choices. Every tree produces samples of data that are

completely unique. Based on user votes, it determines the highest prediction score. Furthermore, it provides a simple metric for the feature's relevance and identifies the most crucial aspects of a dataset. In order to reconstruct data and improve accuracy, feature selection is often used in classification research. The feature selection technique has several applications, such as filtering and encapsulation.

Instead of using the classification method, the filtering strategy depended on a dataset attribute to choose which features to filter. According to Lebedev et al. (2014), the degree to which a function is practical is a key component of its accuracy. One key difference between decision trees and random forests is the reliability of the ensemble forecasts generated by the former. The feature selection procedure is used to obtain the F value of the test statistic while testing a single function.

For a variety of tasks, including classification and regression, the ensemble learning method known as random forests (sometimes called random choice forests) constructs many decision trees during the training process. Random forests are great for classification issues since their output is the class that most trees choose. After a regression task is finished, you will get back the mean or average forecast from all the trees.



(a) (b)

Figure 3. Feature comparison on (a) male and female patients and (b) true and false cases

Random decision forests avoid the problem of decision trees overfitting their training data. While decision trees are better, gradient enhanced trees outperform random forests in terms of precision [citation needed]. Data attributes may, however, affect its operation. The first method using random decision trees was created in 1995 by Tin Kam Ho using the random subspace approach to apply Eugene Kleinberg's "stochastic discrimination" technique for classification. Leo Breiman and Adele Cutler developed an algorithm and registered "Random Forests" in 2006; Minitab, Inc. still owns the trademark as of 2019. In order to construct a collection of

decision trees with minimal variance, the extension combines Breiman's "bagging" idea with random feature selection, which was first proposed by Ho[1] and then refined independently by Amit and Geman. Because they can reliably predict outcomes from various datasets with little configuration, random forests find widespread usage in business as "blackbox" models.

PCA

In machine learning, there may be a notification stating that all we would get back for entering a garbage value is rubbish value. In order for machine learning algorithms to operate at their best, it is necessary to remove noisy, non-critical data from the dataset before making any predictions. There is a hurdle. The feature choosing approach has been used to choose the most significant characteristics to input into the algorithm in order to get the greatest possible accuracy. After collecting data on hypothyroidism from a recognized diagnostic center, the first step included statistical clearing. Finding the necessary properties in our dataset is the second stage in applying feature selection. RFE, UFS, and PCA are the feature selection techniques we use.

Feature selection is a method to automatically choose the qualities that are crucial for predicting the output or variables that interest us. Some of the data in our dataset is inaccurate, which drastically lowers our model's accuracy. In order to get rid of these extraneous pieces of data, feature selection is crucial. The advantage of selecting features is

1. Reduction in over fitting- Making decisions based on important aspects becomes more likely as a consequence of the data being less unneeded.
2. Improvement in Accuracy- By removing any potentially misleading information, it improves the model's predictive power.
3. Reduction in Training Time- Reducing the amount of time and complexity needed to train the algorithm is possible by removing unneeded input.

Primary data analysis, or PCA, is a data reduction approach that plays a crucial role in feature selection by reducing high-dimensional data to low-dimensional data in order to identify the most informative features for a dataset. Using the 'explained_variance_ratio_' attribute to rank the importance of features, the first principal component is the feature that accounts for the most variance in principle component analysis (PCA), the second principal component is the feature that accounts for the second most variation, and so on.

Table.I. Accuracy comparison

Algorithm	Accuracy of prediction in percentage
SVM	83.74
Decision Tree	85.17
KNN	87.44
PCA-Random Forest	89.46

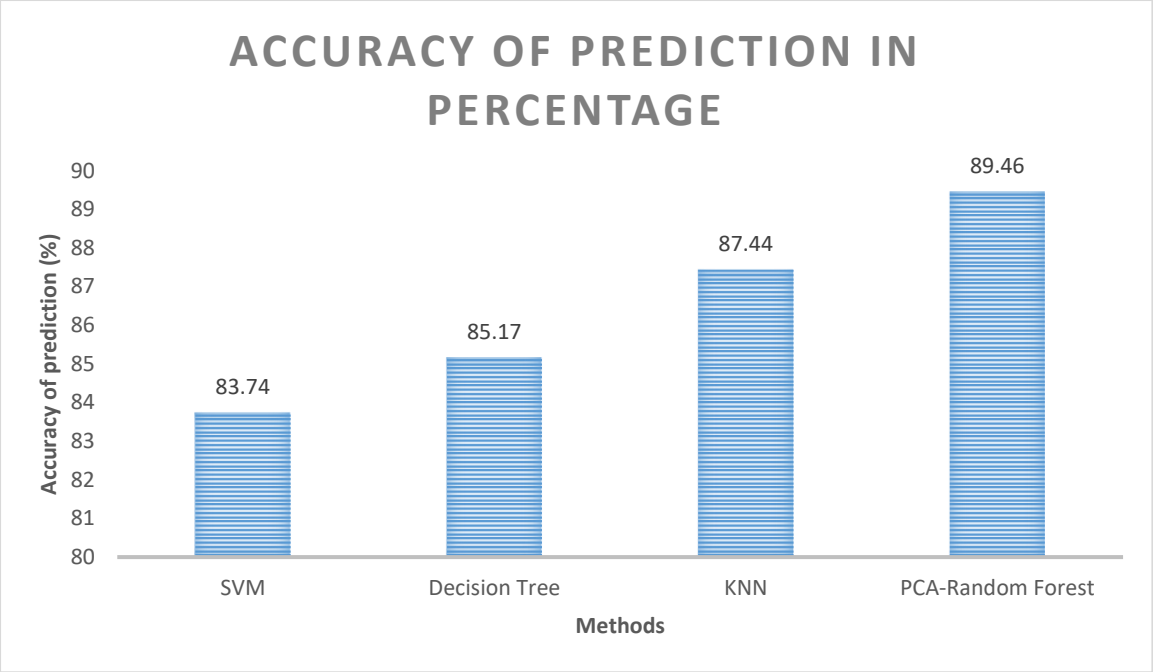


Figure4. Accuracy comparison

IV. CONCLUSION

Using machine learning to diagnose thyroid sickness is a state-of-the-art approach that achieved good accuracy. After data was extracted from a repository at UCI, it was subjected to exploratory analysis to extract insights, and finally, it was cleaned and updated for predictive modeling. We have executed SVM, KNN, DT, and RF and evaluated their performance using F1 score, recall, accuracy, and precision. With regard to precision, the RF classifier delivered satisfactory results.Each algorithm's predicted accuracy using PCA is as follows: SVM (83.74%), Decision Tree (85.17%), KNN (87.44%), and Random Forest (89.46%).

REFERENCES

1. Unnikrishnan, Ambika Menon Usha. Thyroid disorders in India: An epidemiological perspective Review Article. Indian Journal of Endocrinology and Metabolism, Vol. 15, pp. 78-81, 2011.

2. Mirza Shuja, Mittal Sonu, Zaman, Majid. Applying Decision Tree for Prognosis of Diabetes Mellitus. International Journal of Applied Research on Information Technology and Computing, Vol. 9, Issue 1, pp. 15-20, 2018.
3. Shrivas, A. K. Ambastha, Pallavi. An Ensemble Approach for Classification of Thyroid Disease with Feature Optimization. International Education and Research Journal, Vol. 3, Issue 5. pp. 112-113, 2017.
4. Mirza Shuja, Mittal Sonu, Zaman, Majid. Design and Implementation of Predictive Model For Prognosis of Diabetes Using Data Mining Techniques. International Journal of Advanced Computer Research, Vol. 9, Issue 2. pp. 393-398, 2018.
5. Roshan Banu D, 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. pp. 549-553, 2017.
6. TyagiAnkita, MehraRitika. Interactive Thyroid Disease Prediction System Using Machine Learning Technique. 5th IEEE International Conference on Parallel, Distributed and Grid Computing, pp. 689- 693, 2018.
7. Patel Hetal. An Experimental Study of Applying Machine Learning in Prediction of Thyroid Disease. International Journal of Computer Sciences and Engineering, Vol. 7, Issue 1, pp. 130-133, 2019.
8. Marrisa Lourdes De Ataide, AmitaDessai. Thyroid Disease detection using Soft computing Techniques. International Research journal of Engineering and Technology, Vol. 6, Issue 5, pp. 8015-8016, 2019.
9. Yadav Dhyan, Pal Saurabh. To Generate an Ensemble Model for Women Thyroid Prediction Using Data Mining Techniques. Asian Pacific journal of cancer prevention, Vol. 20, Issue 4, pp.1275-1281, 2019.
10. Sidiq U, Aaqib, S.M, Khan, R.A. Diagnosis of Various Thyroid Ailments using Data Mining Classification Techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 5, Issue 1, pp.131-136, 2019.