JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTION OF THYROID DISORDER USING MACHINE LEARNING APPROACH

[1] Meghana G, [2]R Siresha, [3]R Ajith Kumar, [4]Ritika G M, [5]Sathvika Patil

[1] Assistant Professor, [2345]Student
[12345]Department of Computer Science and Engineering,
[12345]Dayananda Sagar University, Kudlu gate, Bangalore, India

Abstract: In India, 42 million people suffer from diseases such as thyroid. Humans have a vascular gland called the thyroid that is one of the most important organs in their bodies. Two hormones are secreted by this gland, which function to control the body's metabolism. There is an imbalance in the body's metabolism when this disorder arises because specific hormones are secreted. Using Machine Learning, this project is designed to detect thyroid diseases in humans. Importing the dataset is accomplished through a User Interface. Three different machine learning algorithms are used to construct the model to detect thyroid disease. In this study, SVM, Random Forest, and Naive Bayes are machine learning techniques used to identify thyroid illness. Using three machine learning algorithms, the accuracy of the model is shown. The most effective machine learning algorithm for detecting thyroid disease has been chosen among the various algorithms that have been used.

Index Terms **- Thyroid, Machine learning, Random Forest, SVM, Naïve Bayes.**

I.INTRODUCTION

As we all know, nowadays, thyroid is a major andthe most frequent disorder mainly among women.Advanced computational biology is widely used in the healthcare industry. This involves gathering and collection of patient details for medical disease detection and prediction. Many intelligent algorithms are used in diagnosing disease at an early stage. The medical information system is rich with various datasets, but intelligent systems are not available for the easy analysis of diseases. Ultimately, machine learning algorithms play a major key role in solving the problems with high complexity and non-linear problems while developing a prediction model. The features are selected from various datasets which can be used as the patientsdetails in a healthy patient as accurate as possiblethat are necessary in any prediction models. Otherwise, misdiagnosis results in a healthy patient that undergoes necessary treatments and care. This thyroid disease is increasing and spreading rapidly all over the world. It is complexto detect this thyroid disorder from the laboratories and requires prior knowledge andgood experience. The thyroid gland is an essential hormone gland which plays an important role in the metabolism and development of the human body. It is a tiny, butterfly-shaped endocrine gland that secretes thyroid hormones that control metabolism and is situated in the neck region beneath the Adam's apple. These thyroid hormones help in maintaining the body's metabolism as well as the temperature. Thesehormones also play a role in processing protein and burning calories. The types of hormones are T4 (Thyroxine) and T3 (Triiodothyronine) thatare released by thyroid gland.

Machine learning plays a major role in detecting this thyroid disease. There are many machine learning algorithms which help in diagnosing thyroid in a human. This study involves three machine learning algorithms such as Naïve Bayes,Support Vector Machine and Random Forest to enhance the prediction accuracy of thyroid, to detect and identify thyroid problems. Thus, if any abnormal conditions of thyroid hormone levels are identified, patients may be prescribed for the treatment and medicine.

II.OBJECTIVE

The main objective of this study is as follows-

- A large number of data is used to estimate thelikelihood of a better result as increasing prediction accuracy will enhance the thyroid problem detection.
- Various pre-processing techniques are applied to enhance the model's performance.
- The accuracy, recall, precision and F1 scores are examined to evaluate the effectiveness of the machine learning algorithms.

III.DATASET

We have collected a dataset from Kaggle, thetotal number of records in the data set is 6962 andnumber of rows is 3772 and 31 columns. After thedata preprocessing, we divided the dataset into two i.e., training data and testing data. We have used 80 percent of the data for training the modeland we have used 20 percent of the data for testingthe model.

IV.METHOD

The initial phase of this project is data collection.It is important to carefully select the data. The data depends on our study aims, objectives, and resource restrictions. The chosen data is nextevaluated to prepare it for the model selection procedure. Data pre-processing is done to clear all the unnecessary data from the raw data, whichmight contain missing, null, and duplicate values.So, to remove all these values, two methods are used. The methods used are Label Encoder and Standard Scaler from Python. After this, the cleaned data is separated into training and testingdatasets. This training and testing data split is used for analyzing the performance of a machine learning system. It is used for problems involving classification and regression and applies to all supervised learning techniques. This process includes separating the dataset into twosubgroups. The first subgroup is used to fit the model and is known as the training dataset. In thesecond subgroup, the input element of the datasetis presented to the model, then predictions are produced and compared to the predicted values. This second subgroup is known as the test dataset. The major goal of this is to evaluate the model's performance on new data. Then, a feature selection procedure is continued. It is the processof limiting the input variables while building a model. It improves the performance of the modeland helps to reduce the cost of computing for modeling. It is necessary for the detection and classification of thyroid disorders. Then, the processed data is utilized to implement all the machine learning algorithms Random Forest, Support Vector Machine (SVM), and Naïve Bayes algorithms. Random Forest is a classifier that uses many decision trees on different subsetsof the input dataset and averages the results to increase the dataset's predicted accuracy. TheSVM algorithm aims to construct the best decision boundary or line that can divide n- dimensional space into classes so that we can quickly classify new data points in the future. Naive Bayes makes predictions based on object probabilities because it is a probabilistic classifier. It uses the Bayes theorem. The formula used is given below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Here, the Confusion Matrix is used to easily understand the model's performance. The below figure represents the Confusion Matrix.

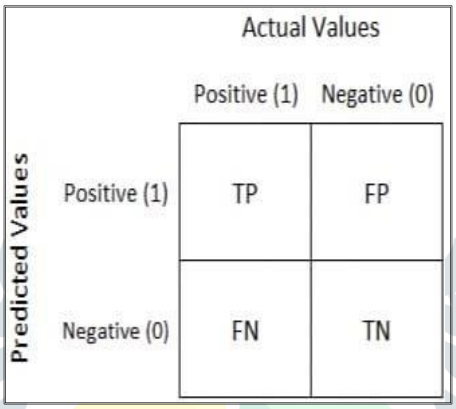

figure 1: Confusion Matrix

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

The above formulas are used to find the Accuracy, Precision, Recall, and F-score. The algorithm that provides the highest accuracy is finally selected as the best among the three.

V.RESULT AND DISCUSSION

The datasets were initially divided into train and test subsets. This method was employed to assess the effectiveness of machine learning models. Our categorization problem was solved inthis way. The three machine learning models we selected for our study were trained using the training dataset, and their performance wasassessed using the test dataset. In this work, we used the naive Bayes algorithm, the support vector machine algorithm, and random forest.

Accuracy, precision, recall, and F1-score are used to examine the experimental data. These assessments are based on information from the confusion matrix.

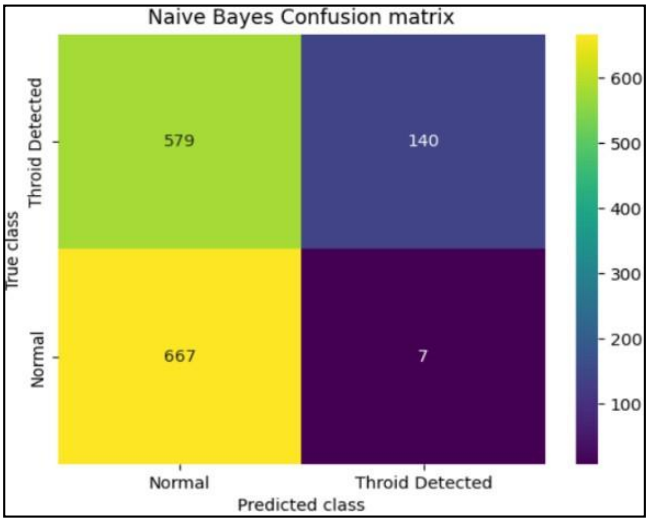

figure 2: Naïve Bayes confusion matrix

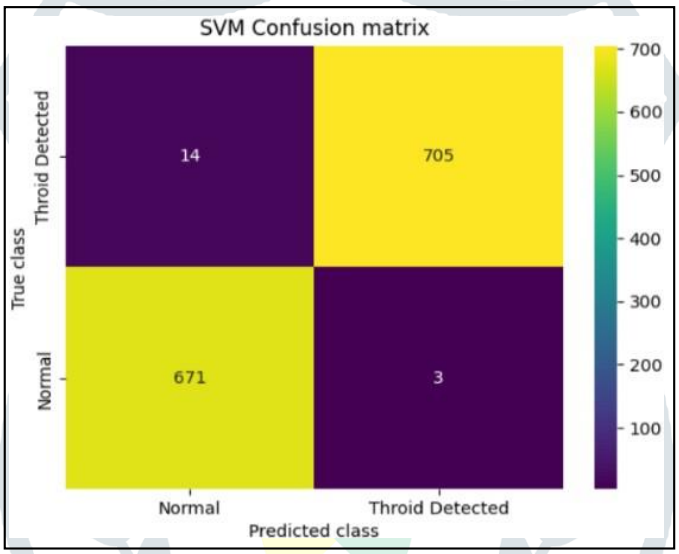

figure 3: SVM confusion matrix

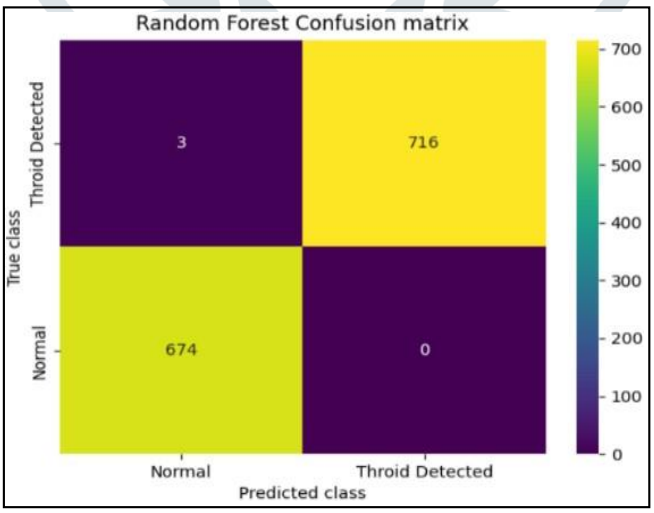

figure 4: Random Forest confusion matrix

The experimental results are summarized below.

```
Naive Bayes Accuracy  :  57.93251974156497
Naive Bayes Precision : 74.38469769930444
Naive Bayes Recall    : 59.216456255184625
Naive Bayes FScore    : 50.90586508852964

SVM Accuracy  :  98.77961234745155
SVM Precision : 98.76623778300136
SVM Recall    : 98.80387366231537
SVM FScore    : 98.77888488227411

Random Forest Accuracy  :  99.85642498205313
Random Forest Precision : 99.85207100591715
Random Forest Recall    : 99.86091794158554
Random Forest FScore    : 99.85628804291757
```

figure 5: Accuracy, precision, recall and F1-scoresfor different classification methods.

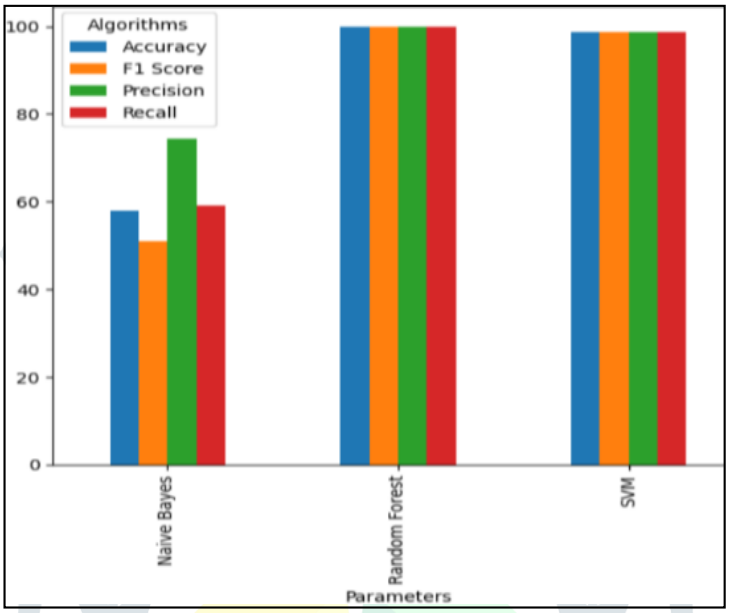

figure 6: Comparison graph of different algorithms

Using **Python** and the **Tkinter** package, we developed graphical user interface (**GUI**). The window's appearance is depicted in the images below, along with few tests that were run to ensure its functionality.

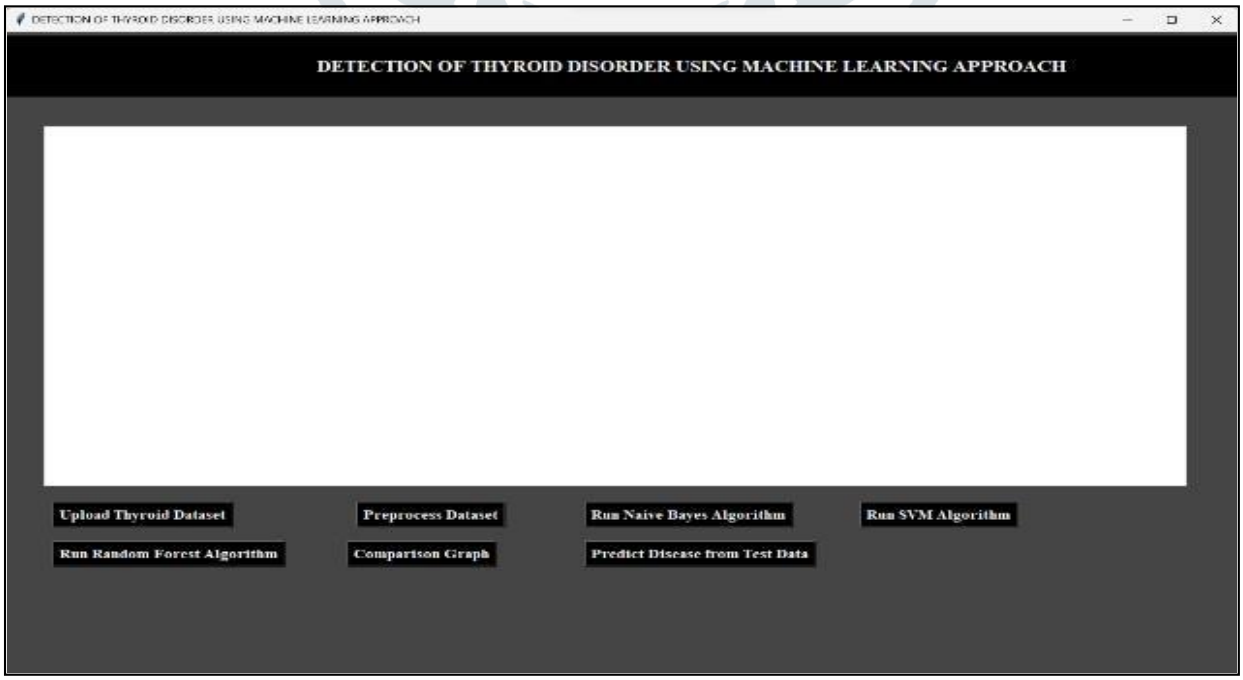

figure 7: opening window

On the opening window*(fig.7)* after performing the task i.e., uploading of dataset, pre-processing it, followed by running the algorithms, then comparison study with comparison graph, then prediction of thyroid diseases using test data.

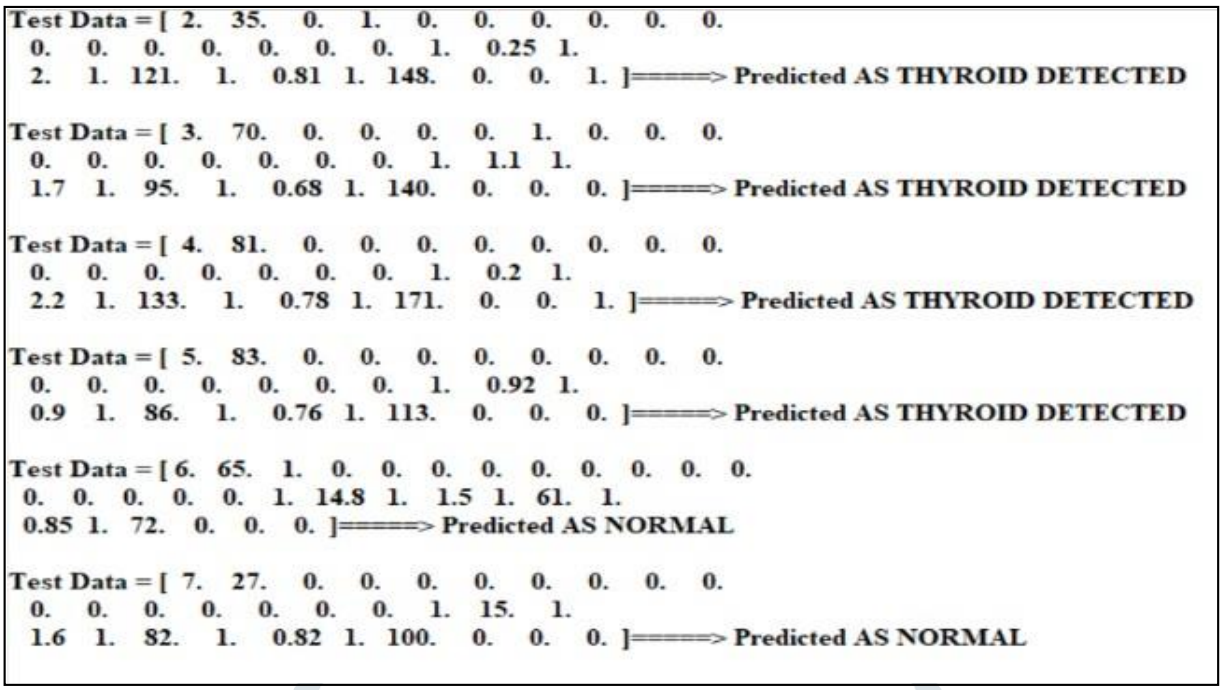

figure 8: Result (the prediction of thyroid diseases)

In the result window in square bracket, we can see test data and after arrow symbol we can see predicted output.

VI.CONCLUSION

The study for the Naive Bayes algorithm, Random Forest, and Support Vector Machine to predict thyroidillnesses is presented in this publication. In order to forecast thyroid illnesses, these algorithms are appliedto a dataset. Precision, recall, F1 score, and accuracy are calculated to evaluate the algorithms used. The Random Forest algorithm performs best overall, with an accuracy rate of 99.8. With accuracy of 98.7, the Support Vector Machine comes in second. Finallycomes the Naive Bayes method, which has the lowestaccuracy at 57.9. With the help of this result while predicting thyroid diseases Random Forest is used foraccurate result. We also draw the conclusion that medical personnel can use the suggested model as a tool to more precisely diagnose patients. This could aid in avoiding the physical labor that might producewrong findings and be quite time consuming.

VII.REFERENCE

[1] Lerina Aversanoa, Mario Luca Bernardia, Marta Cimitileb, Martina Iammarinoa, Paolo Emidio Macchiac, Immacolata Cristina Nettorec, Chiara Verdonea." Thyroid Disease Treatment prediction with machine learning approaches". 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science 192 (2021) 1031–1040.

[2] Priyanka Duggal, Shipra Shukla. "Prediction Of Thyroid Disorders Using Advanced Machine Learning Techniques ". 10th International Conference on cloud computing, data science, and Engineering (Confluence).

[3] Hebatullah Mohammad Almahshi, Esraa Abdallah Almasri, Hiam Alquran, Wan Azani Mustafa, Ahmed Alkhayyat. "Hypothyroidism Prediction and Detection Using Machine Learning". 5th International Conference on Engineering Technology and its Applications 2022- (5thIICETA2022) 978-1- 6654-7215-9/22/$31.00 ©2022 IEEE.

[4] Dr. G. Rasitha Banu, M. Baviya, Dr. Murtaza Ali3. "A STUDY ON THYROID DISEASE USING DATA MINING ALGORITHM". International Journal of Technical Research and Application e-ISSN: 2320-8163, www.ijtra.com volume 3, Issue 4 (July August 2015), PP. 376-379

[5] Ankita Tyagi, Ritika Mehra, Aditya Saxena. "Interactive Thyroid Disease Prediction System Using Machine Learning Technique ".5th IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC-2018), 20-22 Dec 2018, Solan, India

[6] Md Riajuliislam, Khandakar Zahidur Rahim, Antara Mahmud."Prediction of Thyroid Disease (Hypothyroid) In Early Stage Using Feature Selection and Classification Techniques". 2021 International Conference on Information And Communication Technology for Sustainable Development (ICICT4SD), 27-28 February, Dhaka.

[7] Shaik Razia1* and M. R. Narasinga Rao2." Machine Learning Techniques for Thyroid Disease Diagnosis – A Review". Indian Journal of Science and Technology, Vol 9(28), DOI: 10.17485/ijst/2016/v9i28/93705, July 2016.

[8] Tahir Alyas, 1 Muhammad Hamid,2 Khalid Alissa,3 Tauqeer Faiz,4 Nadia Tabassum, And Aqeel Ahmad6 "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach "Hindawi BioMed Research International Volume 2022, Article ID 9809932, 10 pages https://doi.org/10.1155/2022/9809932.

[9] Devansh Sirohi1, Deepanshu Kashyap2, Devendra Pal3, Gopal Goyal4, Bhumica Verma IMSEC Ghaziabad."Thyroid Disease Detection System", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 11 Issue I Jan 2023- Available at www.ijraset.comXXX.

[10] Marissa Lourdes De Ataide, Amita Dessai. "Thyroid Disease Detection using Soft Computing Techniques". International Research Journal of Engineering and Technology (IRJET)