DETECTION OF THYROID DISORDERS USING MACHINE LEARNING APPOARCH

P. Shamiulla

KV SubbaReddy Engineering College, Kurnool, A.P, India

H Ateeq Ahmed, M. Harish, P. Praveen Kumar, G. Madhu Krishna, A. Sreekanth

KV SubbaReddy Engineering College, Kurnool, A.P, India

Abstract

Classification based Machine learning plays a major role in various medical services. In medical field, the salient and demanding task is to diagnose patient's health conditions and to provide proper care and treatment of the disease at the initial stage. Let us consider Thyroid disease as the example. The normal and traditional methods of thyroid diagnosis involve a thorough inspection and also various blood tests. The main goal is to recognize the disease at the early stages with a very high correctness. Machine learning techniques play a major role in medical field for making a correct decision, proper disease diagnosis and also saves cost and time of the patient. The purpose of this study is prediction of thyroid disease using classification Predictive Modelling followed by binary classification using Decision Tree ID3 and Naive Bayes Algorithms. The Thyroid Patient dataset with proper attributes are fetched and using the Decision Tree algorithm the presence of thyroid in the patient is tested. Further, if thyroid is present then Naïve Bayes algorithm is applied to check for the thyroid stage in the patient.

Keywords : *Thyroid Disease, Machine Learning, Classification, Decision Tree (ID3), Naive Bayes Algorithm*

I.INTRODUCTION

Motivation

Thyroid disease diagnosis is not a simple task. It involves many procedures. The normal traditional way includes a proper medical examination and many blood samples for blood tests. Therefore,there is a necessity for a model which detects the thyroid disease at a very early stage of development .

In medical field machine learning plays an important role for thyroid disease diagnosis as it has various classification models based on which we can train our model with proper train dataset of the thyroid patient and can predict and give the results in an accurate manner with higher degree of correctness.

Some recent studies from Mumbai have suggested that congenital hypothyroidism is common in India. The disease occurs in 1 part of 2640 new born children, when compared to the worldwide average range of 1 in 3800 considered. Congenital hypothyroidism can lead to serious complications if not detected in early stages. Therefore, the proposed model serves the goal in early detection of thyroid disease.

Based on the obtained test values the health care staff can easily examine the condition of the patient and also skip further clinical examinations if not necessary. Hence, this approach proves to be very much beneficial to the healthcare field. A proper train dataset results into an accurate predicting model therefore reducing the overall cost of the thyroid patient treatment and also saving the time [2]. Classification algorithms are most suitable in decision making and also solving the real world problems.

ABOUT THYROID

The Thyroid is butterfly-shaped endocrine gland which is situated at the base of the human neck. The vital role of the thyroid gland is maintaining and balancing human metabolism and also the growth and development of the human body. The vital tasks performed by thyroid gland are blood circulation, body temperature control, muscle strength and brain functioning [1]. Any damage or improper functioning of the gland may seriously affect the normal human body functioning [2].Therefore, proper thyroid hormone secretion results into a healthy human body. If there is either low or high secretions of the hormone it will adversely affect the human health.

Various Thyroid Harmones and their effects.

The Thyroid gland mainly produces tri-iodothyronine (T3), thyroxine (T4) and Thyroid stimulating hormone (TSH). The Thyroid stimulating hormone (TSH) [3] is released by the pituitary gland which mainly stimulates the thyroid gland to produce T3 and T4 which further stimulate the metabolism of almost every tissue present in the human body. Therefore, the pituitary gland plays a vital role in controlling the production of the required amount of thyroid hormones. If the TSH production level is less then T3, T4 secretion will be more and vice versa.

The Thyroid disorder is the most common endocrine disease across the world. In a survey carried out in India, around 42 million people are suffering from this disease [1]. Thyroid disease is different from other type of endocrine diseases in terms of the mode of treatment relative attainability and the ease of predicting the disease

The high thyroid hormone secretion leads to Hyperthyroidism and low secretion leads to Hypothyroidism. Both the conditions adversely affect the human physiology and the symptoms shown for hyperthyroidism are dry skin, hair thinning, loss of weight, high blood pressure, neck enlargement and short menstrual periods.

The symptoms show for hypothyroidism include the thyroid gland inflammation, weight gain, low blood pressure, heavy menstrual periods and loss of appetite.

These symptoms may get even worse if they are not treated in an early stage. Hence, there is a need for a proper prediction model which helps in diagnosing the patient's condition in an early stage of the disease

II.LITERATURSURVEY

INTRODUCTION

Bibi Amina Begum et al. [1] have proposed different Thyroid prediction techniques using data mining approaches. They have considered different dataset attributes for prediction and have explained the classification techniques in data mining like Decision Tree, Backpropagation Neural Network, SVM and density based clustering. They have analyzed the correlation of T3, T4 and TSH with hyperthyroidism and hypothyroidism.

Ankita Tyagi et al. [2] have studied various classification based machine learning algorithms. They have considered train data set from UCI Machine Learning repository and compared and analyzed the performance metric of decision tree, support vector machine and K-nearest neighbor.

Aswathi A K et al. [3] have proposed a training model consisting of 21 thyroid causing attributes. They have proposed partial swarm optimization to optimize the support vector machine parameters.

M. Deepika et al. [4] have performed a general empirical study on various disease diagnosis like Diabetes, Breast Cancer, Heart disease, Thyroid prediction and have compared the accuracy rate by applying SVM, Decision tree and Artificial Neural Networks.

Sumathi A et al. [5] have considered Thyroid data preprocessing mainly by applying the decision tree algorithm. They have first calculated the mean values of T3, T4 and TSH and considered as the preprocessing stage. Later on they have applied machine learning based feature selection and feature construction. Further they have applied classification based J48 algorithm which is a continuation of ID3 algorithm and calculated the results.

I Md. Dendi Maysanjaya et al. [6] have analyzed a comparison on various classification methods used to diagnose thyroid disease. They have compared by using Artificial Neural Networks, Radial Based Function, Learning Vector Quantization, Back Propagation Algorithm and Artificial Immune recognition system and concluded the comparison results. Among that they found out that Multilayer Perceptron has the highest accuracy of 96.74%

Ammulu K et al. [7] have proposed a Thyroid Prediction System based on data mining classification algorithm. They have used random forest approach to predict the results using Weka open source tool used for data mining. Using this tool they have applied random forest algorithm with 25 thyroid data attributes and predicted the results accordingly.

Roshan Banu D et al. [8] have conducted a study on different data mining techniques to detect thyroid disease. They have done study on Linear Discriminant analysis, Kfold cross validation, and Decision tree. They have analyzed various splitting rules for the attributes of Decision tree. They have also compared the obtained values.

Dr.B.Srinivasan et al. [9] have conducted a study on diagnosis of the thyroid disease using different data mining approaches. They have explained the major cause of the thyroid disease and have also given description about Decision Tree, Naïve Bayes classification and SVM.

Sunila Godara et al. [10] have performed a Prediction on Thyroid Disease using various machine learning techniques. They have considered Logistic Regression and Support Vector Machine as the main Thyroid detection models. They have concluded that these two proposed classifier methods are the best when the number of classes increases in the thyroid prediction model

EXSISTING SYSTEM

In the data collection stage, small, memory-constrained and low energy-consumption sensors with a short-range communications capability are employed to collect information about the physical environment. Ethernet, WiFi, ZigBee, and wire-based technologies are combined with Transmission Control Protocol/Internet Protocol to connect the objects and users across prolonged distances during data transmission. During the data processing and utilization stage, applications process the data to obtain useful information, and may initiate control commands to act on the physical environment after making decisions based on the collected information. The coordination of diverse technologies, the heterogeneity, and the distributed nature of communications technologies proposed for the IoT by different standards development organizations [4] magnify the threat to end-to-end security in IoT applications

PROPOSED SYSTEM

The thyroid dataset is taken from Kaggle Machine Learning website [13]. The Database mainly includes the thyroid patient records having all the necessary patient details in it. The patient record has important attributes as mentioned in the Table I. Along with this, the proposed model also takes all the records of the patient's past clinical history showing in the Fig 1. These include whether the patient is allergic to any particular medicine, whether the patient has undergone any past thyroid surgery and also any recent thyroid test and genetic history of the patient. These also act as the major attributes since they ease the examination of the thyroid patient and reduce the thorough examination by the doctor. This saves time and eases the diagnosis process.

These attributes are stored in a dedicated cloud server which can be made private or hybrid based on the health organization's need and interest. Among the considered attributes a train dataset is prepared and is given as the input to the classification based machine learning model. This is a supervised learning method and the designed model will generate the results based on the train dataset values. The proposed model has Decision tree and Naive Bayes algorithm to generate the results. A decision tree is a tree based algorithm which follows a top down approach build. ID3 algorithm is used to construct the decision tree. It mainly eliminates any redundant element if present and improves the accuracy of the classification.

EXISTING SYSTEM

This decision tree algorithm is applied to the thyroid patient's records consisting of age, gender, T3, T4, TSH values. The decision tree algorithm calculates the inputted values present in the thyroid patient record. The calculation is done based on the train dataset. Therefore more the number of records in the train dataset higher the accuracy of the algorithm. For example, if 3000 train thyroid dataset records are considered and trained to the decision tree algorithm then the generated accuracy rate will be high. Our proposed model has considered more than 3000 train dataset attributes resulting in 95% accuracy rate in the prediction results. The algorithm generates yes or no values i.e., whether the thyroid disease is present in the patient or not. If the patient's output value results true then further Naïve Bayes algorithm is applied to calculate which stage is the patient currently in. This adds as a major advantage to the health care staff in the easy analysis of the thyroid disease and also avoid certain lab tests if not necessary. The patient's thyroid stage here is divided into 3 stages i.e., minor, major and critical. The Naïve Bayes algorithm is applied if the Decision Tree returns thyroid true or positive value. The Naïve Bayes algorithm in machine learning is a supervised learning algorithm which is based on Bayes' Theorem.

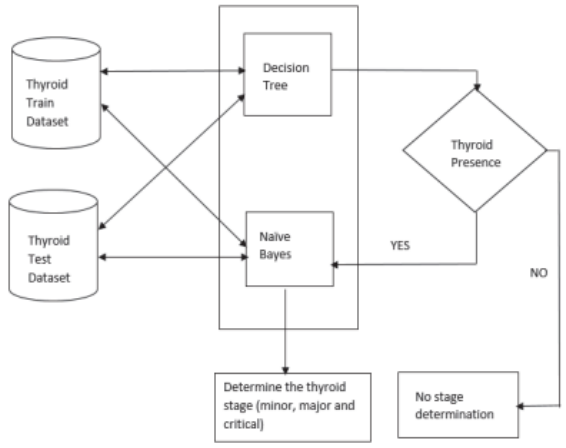


Fig. 2. Proposed machine learning classification model for thyroid disease diagnosis.

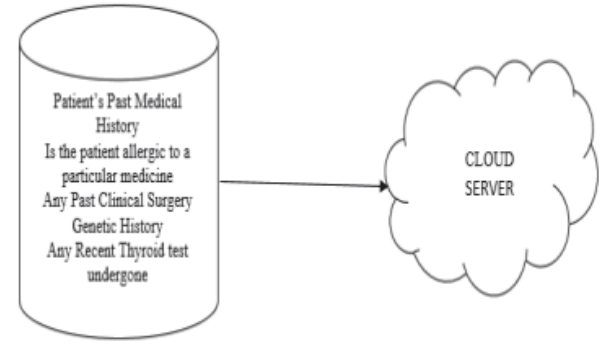
III.SYSTEM REQURIMENT SPECIFICATION

SOFTWARE REQUIREMENTS

- Operating system : Windows 10
- Coding language : Python
- Data base : Mysql

HARDWARE REQUIREMENTS

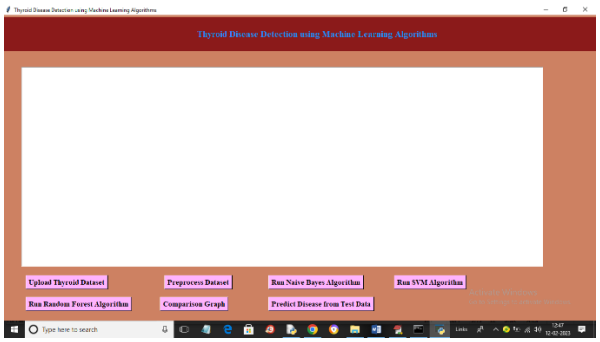
- Processor : intel i3 or above
- Hard Disk : 250 GB
- RAM : 4GB(min)
-



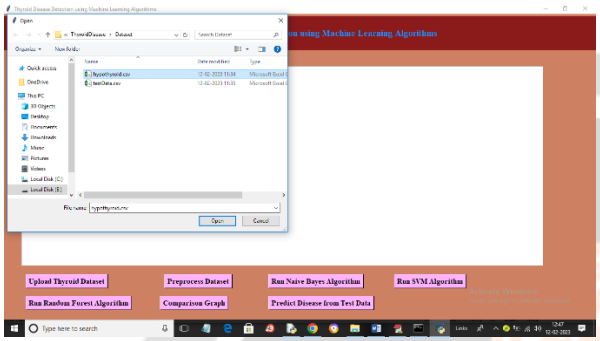
Database consisting of patient's past clinical history

IV.RESULT

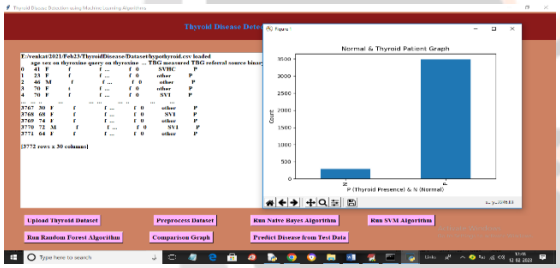
To run project double click on 'run.bat' file to get below screen



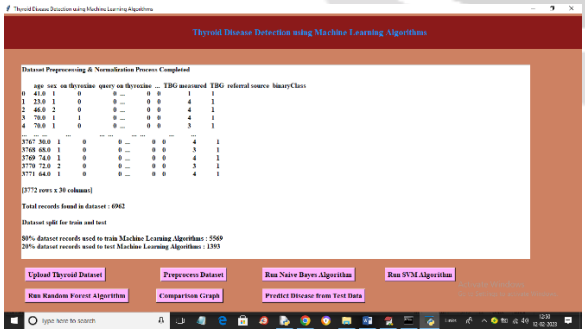
In above screen click on 'Upload Thyroid Dataset' button to load dataset and get below output



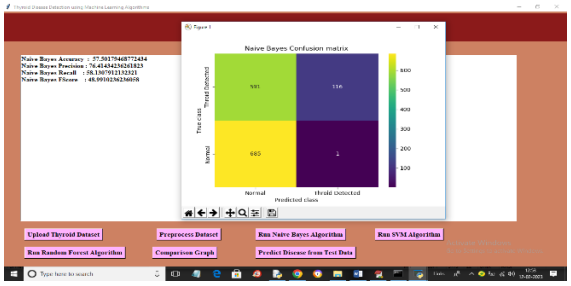
In above screen selecting and uploading thyroid dataset and then click on 'Open' button to load dataset and get below output



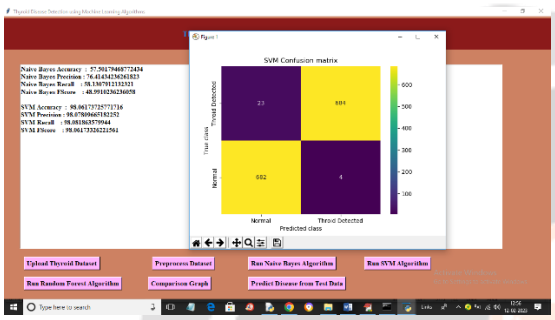
In above screen dataset loaded and in graph x-axis represents N (normal) and P (thyroid presence) and y-axis represents number of records and in above dataset values we can see some are non-numeric and some are numeric and machine learning algorithms accept only numeric values so we need to process dataset to encode non-numeric values to numeric values so click on 'Preprocess Dataset' button to get below output



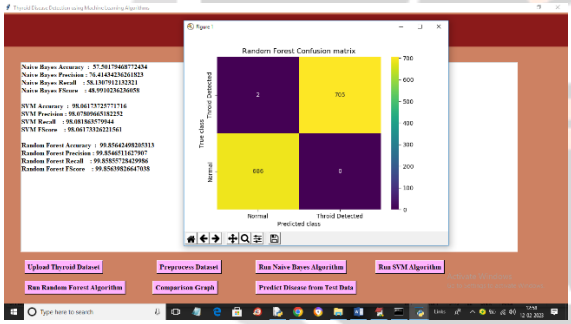
In above screen we can see all values are converted to numeric format and then we can see dataset contains 6962 records where application using 80% records (5569) for training and 1393 (20%) records for testing. Now click on 'Run Naïve Bayes Algorithm' button to train Naïve Bayes on 80% dataset and test on 20% data to get below prediction accuracy



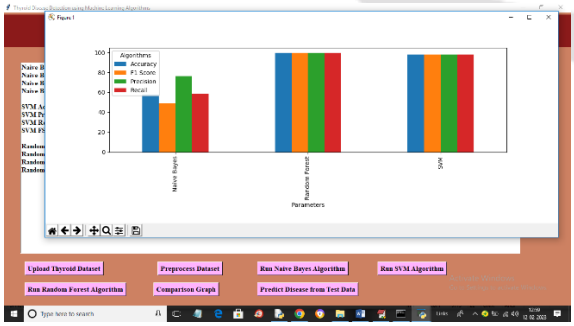
In above screen with Naïve Bayes we got 57% accuracy and we can see other metrics like precision, recall and FSCORE. In confusion matric graph x-axis represents Predicted Labels and y-axis represents True Labels and yellow and light blue colour in diagnol represents correct prediction and dark blue and green box contains incorrect prediction count. Now close above graph and then click on 'Run SVM Algorithm' button to get below output



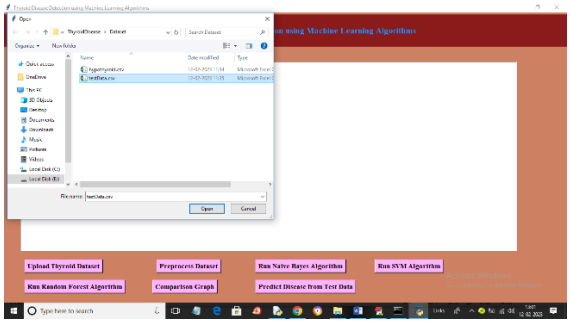
In above screen with SVM we got 98% accuracy and in confusion matrix graph yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count. Now close above graph and then click on 'Run Random Forest Algorithm' button to get below output



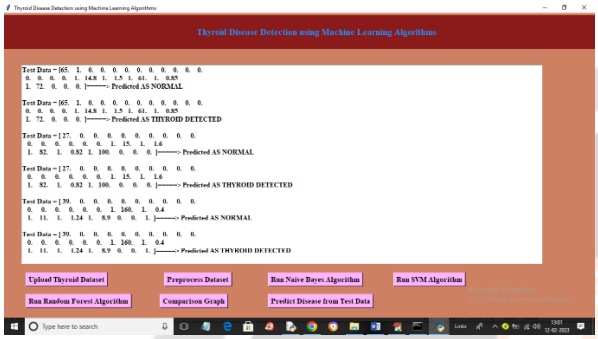
In above screen with Random Forest we got 99% accuracy and now click on 'Comparison Graph' button to get below graph



In above graph x-axis represents algorithm names and y-axis represents metric values like accuracy, precision, recall in different bar colour and in all algorithms Random Forest got high accuracy. Now close above graph and then click on 'Predict Disease from Test Data' to upload test data and get prediction output



In above screen selecting and uploading testData.csv file and then click on 'Open' button to get below output



In above screen in square bracket we can see test data and after arrow symbol we can see predicted output

V.CONCLUSION

Thus the proposed work will be very much useful to identify the thyroid disease in a patient at an early stage using classification based machine learning techniques. These algorithms give various levels of precision and accuracy. These methods also aid in decreasing the unwanted redundant data from the patient's database. The algorithms used in the proposed model are cost effective and also have good output performance and speed. These classification methods make the treatment of the thyroid patient simple by reducing further complex procedures with an affordable price.

VI.REFERENCES

[1] Bibi Amina Begum and Dr.Parkavi "Prediction of thyroid Disease Using Data Mining Techniques" 5Th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019

[2] Ankith 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

[3] Aswathi A K and Anil Antony "An Intelligent System for Thyroid Disease Classification and Diagnosis" Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

[4] M Deepika and Dr. K. Kalaiselvi "A Empirical study on Disease Diagnosis using Data Mining Techniques." Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974-2

[5] Sumathi A, Nithya G and Meganathan S "Classification of Thyroid Disease using Data Mining Techniques" International Journal of Pure and Applied Mathematics, Volume 119 No. 12 2018, 13881-13890

[6] Md. Dendi Maysanjaya, Hanung Adi Nugroho and Noor Akhmad Setiawan "A Comparison of Classification Methods on Diagnosis of Thyroid Diseases" 2015 International Seminar on Intelligent Technology and Its Applications

[7] Ammulu K. and Venugopal T. "Thyroid Data Prediction using Data Classification Algorithm" IJIRST – International Journal for Innovative Research in Science & Technology| Volume 4 | Issue 2 | July 2017

[8] 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)

[9] Dr. Srinivasan B, K.Pavya "Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study" International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov - 2016

[10] SunilaGodara and Sanjeev Kumar "Prediction of Thyroid Disease Using Machine learning Techniques" International Journal of Electronics Engineering (ISSN: 0973-7383) Volume 10 • Issue 2 pp. 787-793 June 2018