2020 IEEE International Conference for Innovation in Technology (INOCON)
Bengaluru, India. Nov 6-8, 2020

A Machine Learning Approach to Predict Thyroid Disease at Early Stages of Diagnosis

Amulya.R.Rao
Electronics and Communication Department
Sri Jayachamarajendra College Of Engineering, JSS Science and Technology University
Mysuru, India
ramulya745@gmail.com

B.S.Renuka
Electronics and Communication Department
Sri Jayachamarajendra College Of Engineering, JSS Science and Technology University
Mysuru, India
renuka@sjce.ac.in

***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—Machine learning techniques, Thyroid disease, classification Predictive Modelling, Decision Tree ID3, Naïve Bayes

I. INTRODUCTION

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 [1].

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.

II. 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.

A. 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 [3].

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 [4].

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 [1].

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

978-1-7281-9744-9/20/$31.00 ©2020 IEEE

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 [9].

III. LITERATURE SURVEY

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, K-fold 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.

IV. DESCRIPTION OF THE DATASET

The Thyroid Dataset is taken from Kaggle Machine Learning Website [13]. The Database mainly consists of the basic patient information details like the name of the patient, personal contact and any past clinical history. These information will be stored in the database and will be considered as the patient record for the further clinical examination. The dataset attributes are considered based on priority. The attribute which is more responsible for causing thyroid disease are considered and the rest are neglected. The Attribute values are Boolean (True/False) or continuous. The main attributes considered are Age, Gender, Hyperthyroid, Hypothyroid, Pregnant, T3, T4 and TSH values.

The below Table I shows the attribute and type of attribute.

TABLE I. DATASET DESCRIPTION

SLNO.	Attribute Name	Value Type
1	Age	continuous
2	Gender	M,F
3	Hyper Thyroid	F,T
4	Hypo Thyroid	F,T
5	Pregnant	F,T
6	T3 value	continuous
7	T4 value	continuous
8	TSH value	continuous

Along with the Dataset Description, The patient's past clinical history is also considered which will further benefit to generate accurate results.

This mainly helps the Health care workers to examine the patient's condition in the best possible way.

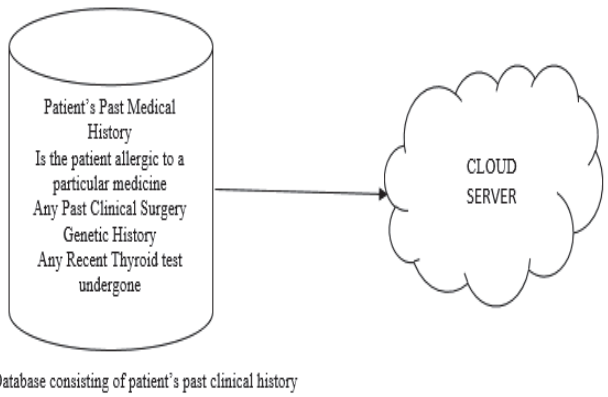

Fig. 1. Database consisting of various past clinical history of the patient.

V. CLASSIFICATION TASKS IN MACHINE LEARNING

A. Classification Predictive Modeling

The Classification Predictive Modeling works based on the given example input data and the corresponding class label is predicted to it [12].

For a classification model to be developed, it requires a good train dataset using which the model can learn the behavior and predict possible outcomes with higher accuracy [12].

The four main tasks in Classification Predictive Modeling are,

Binary Classification

Multi-Class Classification

Multi-Labeled Classification

Imbalanced Classification

A Binary Classification refers to a classification which has only two class tags.

Some of the examples are Decision Tree, Logistic Regression and Naïve Bayes.

Multi-Class Classification have more than two class tags.

This type of classification is very useful when the classifiers are more in number. The Binary Classification Algorithms can also be applied in Multi-Class Classification.

Random Forest can be an example for this type of classification task.

Multi-Labeled Classification includes two or more class tags wherein one or more class tags can be further used to predict each other's results.

Some examples include Multi-Label Decision Trees, Multi-Label Random Forests.

Imbalanced Classification refers to such classification sets wherein the number of the examples classes are not equally distributed [12].

VI. PROPOSED WORK

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.

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.

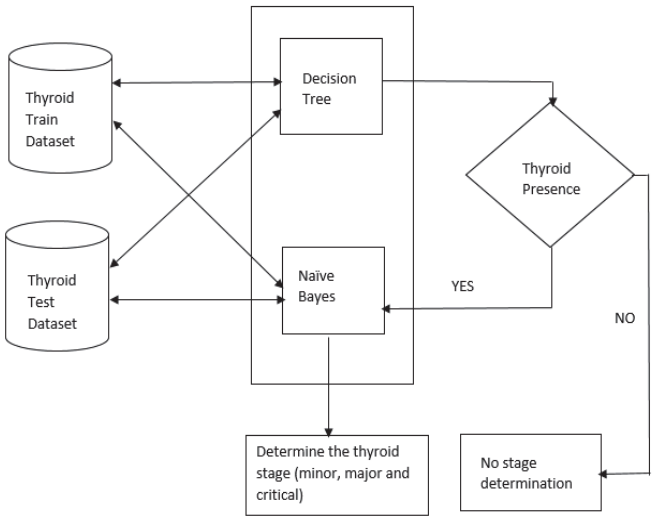

Fig. 2. Proposed machine learning classification model for thyroid disease diagnosis.

It is used for solving classification based problems. The model can be built fast and it is also cost effective one.

In this way our proposed system can make a major contribution in the healthcare field and also generate positive outcomes with good accuracy with a cost and time saving method for the thyroid patients

VII. 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.

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

[11] A. Colubri, T. Silver, T. Fradet, K. Retzepi, B. Fry, P. Sabeti, Transformingclinicaldata into actionable prognosis models: machine learning Framework and fielddeployableapp to predict outcome of Ebola Patients, PLoSNegl. Trop. Dis. 10 (3) (2016) e0004549.

[12] https://machinelearningmastery.com/types-of-classification-in-machine-learning

[13] https://www.kaggle.com/kumar012/hypothyroid

[14] Ali keles et al.,"ESTDD: Expert system for thyroid diseases diagnosis", Expert system with Applications,34,242-246,200

[15] http://www.thehealthsite.com/diseasesconditions/world- thyroid - day- 2012 – facts-you-should-know/