International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Prediction of Thyroid Disease using Advanced Machine Learning

***[1]V Vasavi Sujatha, [2]Bhupathi Harshini, [3]Ankathi Chinmayi**

[1]Assistant Professor, Department of Information Technology, Bhoj Reddy Engineering College for Women, Hyderabad, India

[2,3]Students, Department of Information Technology, Bhoj Reddy Engineering College for Women, Hyderabad, India

ARTICLEINFO

Article History:

Accepted: 10 April 2023
Published: 30 April 2023

Publication Issue
Volume 9, Issue 2
March-April-2023

Page Number
655-660

ABSTRACT

Thyroid disorder leading cause of medical diagnosis and prediction development, which medical science is a complicated axiom. Thyroid gland is one of our body's main organs. Thyroid hormone secretions are responsible for regulating metabolism. Hyperthyroidism and hypothyroidism are the two prominent thyroid disorders that produce thyroid hormones for control of body metabolism. The machine learning is critical in the disease prediction process and in the study and classification models used for thyroid disease on the basis of data obtained from hospital datasets. A decent knowledge base must be ensured, built and used as a hybrid model to solve dynamic learning tasks like medical diagnosis and prediction tasks. Basic techniques of machine learning are used for the identification and inhibition of thyroid. The SVM is used to predict the approximate probability of a thyroid patient. If the patient has risk of getting thyroid our system has to give suggestions like recommending home remedies, precautions, medication etc.

Keywords : Machine Learning Algorithm, Thyroid disease, Support Vector Machine (SVM), K-NN, Decision Trees Prediction model system.

I. INTRODUCTION

Advanced machine biology is used in the area of healthcare. It required data to be collected for medical disease prediction. For early-stage disease detection, various intelligent prediction algorithms are used. The Medical Information System is good with data sets, but intelligent systems are not available for the fast diagnosis of diseases. Eventually, machine learning algorithms play a key position in solving complex and non-linear problems during the creation of prediction model. The characteristics that can be selected from the various data sets that can be used as description in a healthy patient as specifically as possible are needed in any disease prediction models. Otherwise, misclassification can result in a good patient receiving inappropriate care. The reality of forecasting any condition associated with thyroid illness is also of the greatest cardinal number. Thyroid gland is endocrine in stomach. It is erected in lowered portion of human neck, under apple of Adam, and assists in secretion of thyroid hormones and which ultimately affects

metabolism rate and protein synthesis. To control body metabolism, these hormones count on how quickly heart beats and how quickly calories burn. The composition of thyroid hormones helps to control the body's metabolism. These glands consist of two mature levothyroxine (abbreviated T4) and triiodothyronine thyroid hormones (abbreviated T3). These thyroid hormones are essential for manufacturing and general construction and regulation in order to regulate body temperature. T4 and T3 are exclusively two activated thyroid hormones that usually compose of thyroid glands. These hormones are vital to the control of proteins; distribution at body temperature and energy- bearing and propagation in every part of the body. With T3 and T4 hormones, iodine is primary building block of thyroid glands and is prostrate in only some unique problems, which are exceedingly prevalent. Insufficient elements of these hormones to hypothyroidism and an inappropriate portion to hyper thyroidism. Hyperthyroidism and underactive thyroidism have multiple origins. There are a number of drugs. Thyroid surgery is weak to ionizing radiation, continuous thyroid softness, iodine deficiency, and loss of enzyme to produce thyroid hormones.

II. RELATED WORK

We have seen that data mining techniques are used in the prediction of accuracy related to TD [5]. Numerous methods have been used in the past for knowledge abstraction by using recognized techniques of data mining for TD prediction. In one research, data mining techniques were used for thyroid disease prediction. These methods use dataset from UCI repository, where features were extracted for disease prediction. The dataset with support vector machine (SVM), Decision Tree is used for classification, where data set was chopped for training and testing purpose. The highest accuracy was achieved by SVM with 99.63% accuracy [6]. In another research Hetal Patel [7] came to conclusion that multiclass classifier algorithm achieved the highest accuracy of 99.5%. The dataset was taken from UCI machine learning repository, which is free, public accessible for research purposes. Ataide et al. [8] proposed soft computing techniques for thyroid prediction. The results on the UCI data set showed that multilayer perceptron (MLP) yielded an accuracy of 97.4%. However, after feature extraction the same classifier shown an accuracy of 91.7% which is less than previous results. Yadav et al. [9] generated ensemble methods (bagging+boosting) for thyroid prediction after comparing bagging, boosting and stacking methods. The dataset was downloaded from UCI machine learning repository. During experimentation the author found the performance measure of ROC=98.5, MAE=0.49, RMSE=0.07 and RAE=37.83 and RRSE=51.93. Sidiq et al. [10] implemented Decision Tree, Naïve Bayes, SVM and K nearest neighbour (KNN) in anaconda. 10-fold cross validation method was used to guarantee results. After experimentation on UCI dataset, it was concluded that Decision Tree obtained the highest accuracy of 98.89 than other classification techniques. Razia et al.employed SVM, Multiple Linear Regression, Naïve Bayes and Decision Tree on dataset collected from UCI. The results of these classifiers were compared and it was found that Decision Tree performed well and showed an accuracy of 99.23%. Deepita et al concluded that the use Decision Tree showed better results during the prediction of various diseases. The decision trees showed an accuracy of 95% on thyroid data set downloaded from UCI, however both the SVM and artificial neural network (ANN) performed well and obtained an accuracy of 98.6%. Gurram et al.compared logistic regression and SVM for thyroid disease prediction on the data set taken from UCI. The results showed that former performed well that latter and showed the accuracy of 98.82%. Shrivas et al.used ensemble approach with forward and

backward feature selection method for thyroid prediction on the thyroid dataset taken from UCI. The experiment was carried out in rapid minor tool with 70% of data for training and remaining 30% for testing purposes. The proposed ensemble model of Random Forest, Naïve Bayes and KNN achieved the accuracy of 97.61%. Ammulu K et al. took hypothyroid data set from UCI machine learning repository and applied random forest classifier for thyroid prediction. The results generated in WEKA tool showed an accuracy of 70.51%. Agarwal et al. proposed auto associative neutral network (AANN) on thyroid dataset. The data set collected from UCI was partitioned into 60-40% split for training and testing purpose. The resulting AANN approach yielded the accuracy of 95.1%.Mazin Abdul Rasool Hameed used multilayer forward feed neural network trained by back propagation algorithm for prediction of thyroid disease. The neural network contained only one hidden layer with five neurons which showed the classification rate of 99.2%. The proposed neural network was designed and tested in MATLAB. The dataset was collected from real patients containing three attributes as T3, T4 and TSH. Mahurkar et al. devised improvised k means algorithm for normalization of raw data. The normalized data set was fed to feed forward neural network, which achieved an accuracy of 98.21%. The data set collected from UCI contained 215 instances. Dewangan et al. developed classification and regression tree (CART) on the UCI thyroid data set. Initially info gain and gain ratio feature selection techniques were used with CART as base classifier. After comparison of feature selection techniques with CART as classification model, the best classifier (CART-info gain) achieved an accuracy of 99.47%. Bekar et al. compared the performance of various decision algorithms to find out best algorithm for thyroid prediction. The data set was collected from a general surgeon working at hospital (not mentioned). After experimentation it was concluded that Naïve Bayes tree showed the top accuracy of 75%.Ionitaet al compared radial basis function, Naïve Bayes, multi-layer perception and decision tree classifiers to find out the best classifier for thyroid prediction. The data set used to test and validate the classifiers was taken from the website containing Romanian data and UCI machine learning repository. During experimentation it was shown that decision tree showed the best accuracy of 97.35% with removal of three attributes in data set. Dash et al. proposed Naïve Bayes classifier by using ranker search as feature optimization technique for thyroid disease prediction. The dataset obtained from UCI repository was trained and tested by 10 fold cross validation. The results achieved an accuracy of 95.38%.As shown above, ensemble learning techniques are the most popular and effective machine learning methods used to diagnose TD. However, few studies have been conducted for using ensemble methods to diagnose TD. Ensemble learning methods are effective in diagnosing TD. Ensemble methods are effective in diagnosing TD as they combine the results of several classifiers into one prediction model. Ensemble methods are also known as meta-algorithms that reduce variance and bias to improve the results. It has been also seen that most of the researchers have taken data set from UCI machine learning repository in the field of thyroid disease prediction. The main objective of our research is to predict the thyroid disease of the real world patients. The data set present in the UCI is outdated because the data set was donated to UCI on 1st of January 1987 with plenty of missing values. We have discussed the parameters of the UCI dataset with the endocrine specialist working at Jaipur National University institute for medical sciences and research centre (JNUIMSRC) in India and we came up with conclusion that the data set doesn‟t meet the standards to predict the thyroid disease. In this work, we introduce a new dataset containing the pathological observations and the serological tests of the real patients.

III. PROPOSED SYSTEM

A system architecture diagram would be used to show the relationship between different components. Usually they are created for systems which include hardware and software and these are represented in the diagram to show the interaction between them.

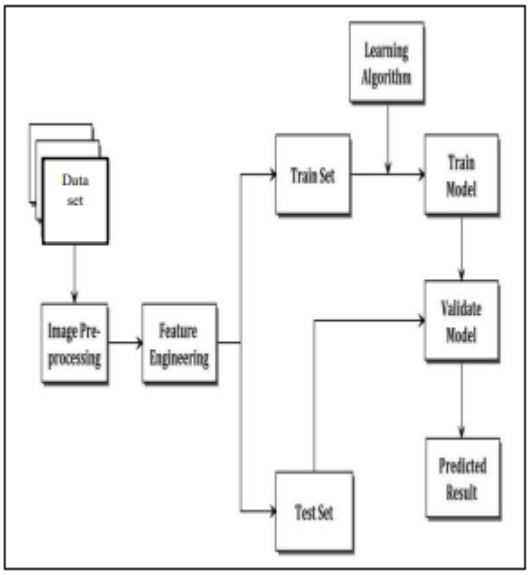

Figure 1: System Architecture

Random Forest Algorithm:

1. Select k data sets randomly from given training data.
2. For selected data sets build a decision trees for these subsets which is chosen.
3. Choose number N for decision trees that you want to build.
4. Repeat steps 1 and 2.
5. For new data sets, calculate each decision tree prediction and add new data set to class which has majority

IV. RESULTS

The Machine Learning Technology has become very easy to predict relation and patterns of various data's. This paper mainly involves in predicting the type of thyroid diseases. The model is built using training data set which have the data cleaning and data transformation. This model has 98% Accuracy in analysis the data set with the help of data visualization.

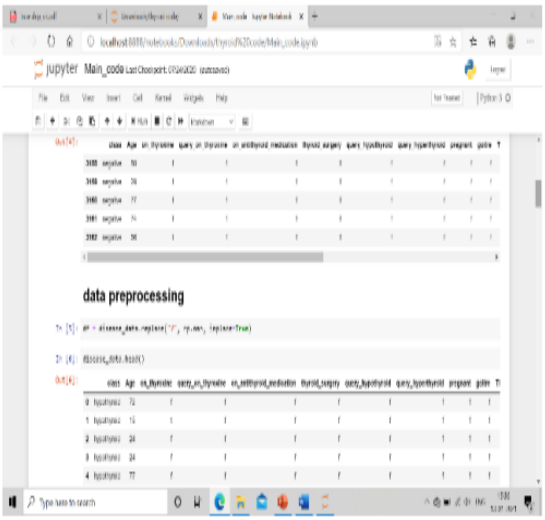

Figure 2 : Data of the Model

In the above diagram we have different parameters like age, Class, on_thyroxine, thyroid surgery, Pregnant, Goitre and outcome for prediction of thyroid for different Scenarios.

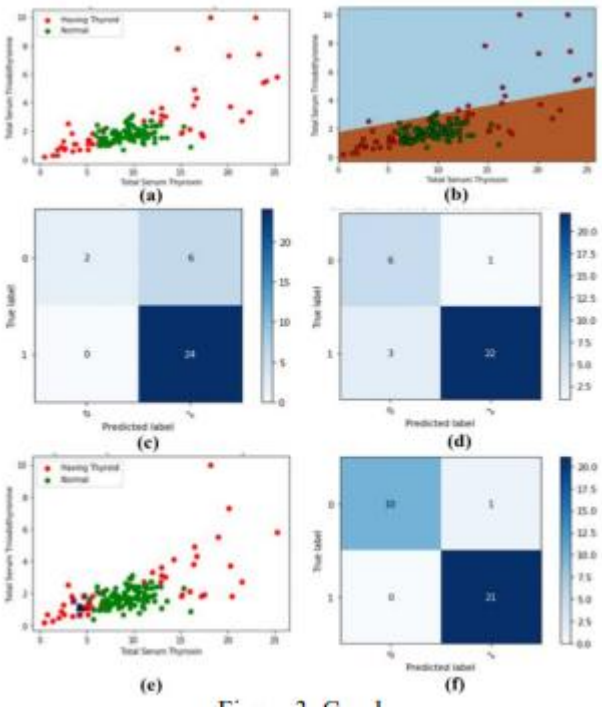

Figure 3: Graph

A dot chart or dot plot is a statistical chart consisting of data points plotted on a fairly simple scale. The numeric data is normally distributed to the left or right and it can also consider in the table. Age, Class, thyroid surgery, Pregnant and some more features.

V. CONCLUSION

In this work, we have used machine learning algorithms to predict thyroid disease. In this system, we have used data mining classification algorithms and regression algorithms. So, both regression and classification are combined to produce accurate diagnosis results. The logistic regression is more efficient and accurate compared to other classification techniques. But other recent techniques can be combined in future to give still more accurate results of thyroid diagnosis.

VI. REFERENCES

[1]. Bibi Amina Begum and Dr.Parkavi A "Prediction of thyroid disease using data mining techniques",5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019. (references)

[2]. Shaik Razia, A Comparative study of machine learning algorithms on thyroid disease prediction, International Journal of Engineering & Technology, 7 (2.8) (2018) 315-319

[3]. Thyroid: https://en.wikipedia.org/wiki/Thyroid

[4]. Aswathi A K and Anil Antony "An Intelligent System for thyroid disease classification and diagnosis" 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN:978-1-5386-1974- 2. (references)

[5]. Shaik Razia, Swathi Prathyusha, Vamsi Krishna, Sathya Sumana "A Comparative study of Machine Learning Algorithm on thyroid disease prediction", International Journal of Engineering & Technology, 7 (2.8) (2018) 315-319. (references)

[6]. Haria Viral, More Suraksha, Patel Bijal and Patil Harshali "Thyroid Prediction System using Machine Learning Techniques", International Journal of Scientific Research and Reviews (IJSRR) 2018, 7(4), 674-681. (references)

[7]. K. Rajam, R. Jemina Priyadarsini, "A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, Issue. 5, May 2016, pg.354 – 358.

[8]. D. T. Larose, Discovering knowledge in data: An introduction to data mining, John Wiley & Sons, (2005) 385

[9]. Yadav Dhyan, Pal Saurabh, "To Generate an Ensemble Model for Women Thyroid Prediction Using Data Mining Techniques," in Asian Pacific journal of cancer prevention, Vol. 20, Issue 4, pp.1275-1281, 2019.

[10]. Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22Nd ACM SIGKDD International Conferenceon Knowledge Discovery and Data Mining. KDD '16, pages 785–794, NewYork, NY, USA, 2016. ACM

[11]. John T. hancock, Taghi M. Khoshgoftaar, "CatBoost for Big Data : An interdisciplinary review" in Journal of Big Data, (2020) 7:94 https://doi.org/10.1186/s40537-020-00369-8

[12]. Guolin Ke, Qi Meng, Thomas Finley, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

[13]. Pratiksha Chalekar, Shanu Shroff, Siddhi Pise, SujaPanicker, "Use of k-Nearest Neighbor in Thyroid disease classification", in International Journal of Current Engineering and Scientific Research

[14]. Payam Refaeilzadeh, Lei Tang, and Huan Liu, “CrossValidation”, pages 532–538. Springer US, Boston, MA, 2009.

[15]. Ozyilmaz, Lale, and Tulay Yildirim, "Diagnosis of thyroid disease using artificial neural network methods" Neural Information Processing,. Proceedings of the 9th International Conference on. Vol. 4. IEEE, 2002.

Cite this article as :

V Vasavi Sujatha, B Harshini, A Chinmayi, "Prediction of Thyroid Disease using Advanced Machine Learning ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.655-660, March-April-2023.