Enhanced prediction of thyroid disease using machine learning method

Conference Paper · December 2023

CITATIONS
0

READS
24

3 authors:

Madhumita Pal
Government College of engineering, Keonjhar
18 PUBLICATIONS **120** CITATIONS
SEE PROFILE

Smita Parija
C. V. Raman College of Engineering
49 PUBLICATIONS **1,430** CITATIONS
SEE PROFILE

Ganapati Panda
C V Raman Global University
225 PUBLICATIONS **5,820** CITATIONS
SEE PROFILE

Enhanced prediction of thyroid disease using machine learning method

Madhumita Pal
Dept. *of Electrical Engg.*
Govt.college of Engg.
Keonjhar-758002, Odisha, India
madhumitapal@gcekj.ac.in

Smita Parija
Dept. of Electronics and Telecommunication
Cvraman Global Univesity
Bhubaneswar,Odisha,India
smita.parija@gmail.com

Ganapati Panda
Dept. of Electronics and Telecommunication
Cvraman Global Univesity
Bhubaneswar,Odisha,India
ganapati.panda@gmail.com

Abstract— Thyroid disease is becoming increasingly in men, women and children but commonly occurring among women over the age of 30. It causes heart problem, eye problem, fertility and pregnancy problems over its effect for long time. As a result, it is critical to evaluate the thyroid information in order to forecast the early prediction of disease and take steps to avoid the deadly condition of thyroid cancer. This study is based upon designing a model for timely detection of thyroid disease by observing the features from thyroid disease dataset which was accessed from UCI repository site by using machine learning algorithms. We have used three machine learning models such as K-Nearest Neighbors(K-NN), decision tree(DT) and multi-layer perceptron (MLP) for prediction of thyroid disease and measure the performance of these models in form of accuracy and area under the curve. Comparative analysis of these three models reveals that MLP performs better in classifying thyroid disease with an accuracy value of 95.73 and Area Under the curve with value of 94.23. The planned experiment was carried out on 3163 cases and 24 thyroid characteristics.

Keywords—Machine learning, Thyroid, multi-layer perceptron, decision tree, K-NN.

I. INTRODUCTION

Thyroid disease is a leading source of medical analysis and prediction, with an onset that is a challenging concept in medical study. One of our body's most vital organs is the thyroid gland. It is a small butterfly shaped organ that is found at neck just in front of the box. Thyroid hormone secretions are to blame for metabolic control [1]. It rises beneath the Adam's apple at the lower region of the human neck, assisting in the secretion of thyroid hormones, which affects the pace of metabolism and protein synthesis. The thyroid gland's production of thyroid hormones aids in the control of the body's metabolism. It contains two active thyroid hormones: levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3). These hormones are required in the manufacture, as well as in the overall construction and supervision, to regulate the body's temperature [2]. Thyroid gland commonly produced two types of hormones namely Thyroxin (T4) and triiodothyronine (T3) [3][4][6]. If the amount of thyroid hormone in the blood is not properly balanced then it affects heartbeat, energy levels, digestion, body temperature, thoughts and feelings of human body [5]. Thyroid illnesses such as hyperthyroidism and hypothyroidism have been caused by deficiency of thyroid hormones.

Types of thyroid disorder

i) Hypothyroidism

An underactive thyroid gland can lead little amount of thyroid being made which is called hypothyroidism. sign of hypothyroidism is tiredness, weight gain, constipation, muscle weakness and acres, hoarse voice, pins and needles in the glands, slow speech, low mood/anxiety, memory problems.

ii) Hyperthyroidism

Overactive thyroid gland where the thyroid hormones levels are too high called hyperthyroidism [7]. Symptoms of Hyperthyroidism are racing heartbeat, weight loss, feeling sweaty and shaky, diarrhea, dehydration, crawling, agitation, disturbed and irascible, obsession and nervousness. Data purification procedures were used to prepare the data so that analytics could be run on it to determine the likelihood of patients developing thyroid cancer. In the healthcare industry, computational biology is being used to advance. It allowed for the collection of stored patient data in order to anticipate medical disease. Recently machine learning plays a crucial role in disease prediction, which this study addresses. There are many data sets in the medical information system, but no intelligent systems that can quickly analyze the disease. Machine learning algorithms have been increasingly important in handling complicated and nonlinear problems in the development of prediction models over time. In any illness prediction model, the features that may be selected from many datasets and utilized as a classification in healthy patients as precisely as feasible must be prioritized. Otherwise, a misdiagnosis could lead to an otherwise healthy patient receiving unneeded treatment. As a result, the cardinality of forecasting any disease in association with thyroid disease is paramount. Application of machine learning helps the medical practitioner for timely detection of the disease and it also helps in treating patient accordingly. The main objective of the study is prediction of thyroid disease using machine learning classifiers such as K-NN, decision tree and MLP which helps the health care professionals to detect the disease at early stage instead of going several blood tests. The organization of the paper has been done in the subsequent manner. Section II represent the related work done by the researchers in prediction of different diseases using ML models. Section III focused on the methods used for the experimental work. Section IV represents the result part of the research work. Section V conclude the article.

II. Literature Survey

There has been innumerable work done in recent years to diagnose the different illnesses of the thyroid. Various types of data mining techniques have been employed by many writers The authors demonstrated that their study, which comprises numerous datasets and algorithms related to future work to get effective and improved outcomes, provided an adequate methodology and assurance for finding disorders similar to thyroid disease. The objective of this work is to. explain various data mining algorithms and statistical

qualities that have been popular in recent years for the evaluation of thyroid illnesses with the assurance of many researchers to achieve diverse outcomes and approaches Decision tree, K-NN, and MLP are only a few of the machine learning algorithms that are widely utilized in common diseases and prognostic situations.

Authors [8] predict thyroid disease in infants using data mining technique. They performed the experiments over 4812 infants record collected from health center of Alborz provenance, Iran in 2016. They obtained maximum accuracy, precision, recall, F-measure 99.58%,100%.73.33% and 84.62% respectively using support vector machine model. The back-propagation technique was employed by authors [9]in a systematic strategy for earlier diagnosis of thyroid illness. ANN is delicate and relies on reverse propagation of an error that has already been used to forecast disease. The influence of ANN is trained using experimental features and testing methods that are supported by data that was not used during the training process. Kaur and co-workers [10] reported an accuracy of 97.26% on dermatology data set using random forest machine learning model. Dogantekin et al. [11] proposed a model for diagnosis of thyroid illness using the approach of Principle Component Analysis (PCA) and least square support vector machine. Wei-Chang Yeh used swarm optimization (SSO)technique to detect thyroid using close interval encoding scheme [12]. To determine the appropriate number of clusters, Azar et al (2013) compared hard and fuzzy clustering techniques on a data set of thyroid illnesses. When comparing the results, various scalar validity measures are applied. The Sammons mapping method is used to find a low-dimensional representation of a set of points using K-means clustering, KMedoid clustering [13]. Steg Mayer et al. (2012) suggested an integrated computational intelligence approach for biological data mining [14]. Rehman et al. [15] proposed a model for diagnosis of thyroid disease using machine learning classifiers. They have taken the thyroid dataset from DHQ teaching hospital, Pakistan and obtained an accuracy of 100% using logistics regression and naïve bayes machine learning classifiers. Geetha et al. [16] classify thyroid disease using evolutionary multivariate Bayesian classifier with an accuracy value of 97.97%. Prasad et al. [17] used particle swarm optimization for detection of thyroid disease and obtained an accuracy of 93%. Prediction of breast cancer diseases was performed by Mushtaq and coworkers [18] using K-nearest neighbor machine learning algorithm with k `values ranges from 1-9. Tomaro and co-workers [19] discussed the various application of data mining technique on health care domain. A new fuzzy logic technique has been developed for diagnosis of the kidney disease such as kidney stone [20]. Rajkumar et al. present a survey on patient suffering from coronary artery disease along with diabetes [26]. Chaubey and coworkers [28] obtained maximum accuracy of 96.87% using KNN model in prediction of thyroid disease. Yadav and coworkers [30] obtained accuracy 99% for detection of thyroid using random forest ml method.

III. DESCRIPTION OF THE DATA SET

The dataset used for the experimental work has been assesse d from University of California, Irvine (UCI) site which cont ains 3163 samples and 24 features as shown in fig 1. From w hich 2870 samples show patient doesn't suffered from thyroi d disease and 293 samples contain patient having hypothyroi d disease as shown in Fig 1. The features of the dataset have been described in table 1. In this research correlation matrix is used for feature selection of the dataset. Correlation matri x establish a relationship between dependent features and i ndependent features that is whether they are positively correl ated or negatively correlated with each other. Those features which are strongly correlated they can be dropped from the dataset for reducing the overfitting problem. The diagonal el ements of correlation matrix are one which represents same f eatures are strongly correlate with each other. Suppose if tw o features with correlation value greater than 0.67 then we c an consider only one feature and another variable may be dr opped from the dataset. The correlation matrix of the datase t has been shown in Fig.2.The experiment was conducted usi ng python open-source software in jupyter notebook.

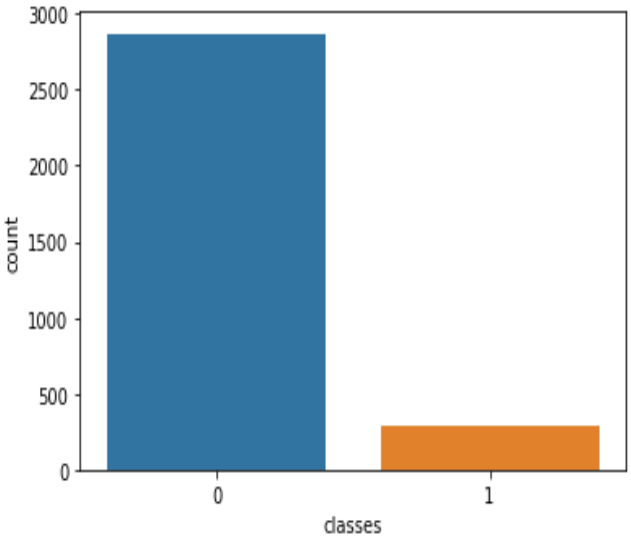

Fig.1.Number of patients suffered from hypothyroidism (1) and number of patients doesn't have thyroid disease (0)

Table 1. Features of the dataset

F1	on_thyroxine	Range of the fe atures
F2	query_on_thyroxine	False, true
F3	on_antithyroid_medication	False, true
F4	thyroid surgery	False, true
F5	query hypothyroid	False, true
F6	query hyperthyroid	False, true
F7	pregnant	False, true
F8	Sick	False, true
F9	Tumor	False, true
F10	Lithium	False, true
F11	Goitre	False, true
F12	ThyroidStimulatinghormonemeasured	False, true
F13	T3_measured	False, true
F14	TT4_measured	False, true
F15	T4U_measured	False, true
F16	FTI_measured	False, true
F17	Age	1 to 94
F18	Sex	Male, Female
F19	TSH	0.005 to 530
F20	T3	0 to 11
F21	TT4	2 to 430

F22	T4U	0.25 to 2.12
F23	FT1	2 to 395
F24	CLASSES	0=normal 1=Hypothyroid

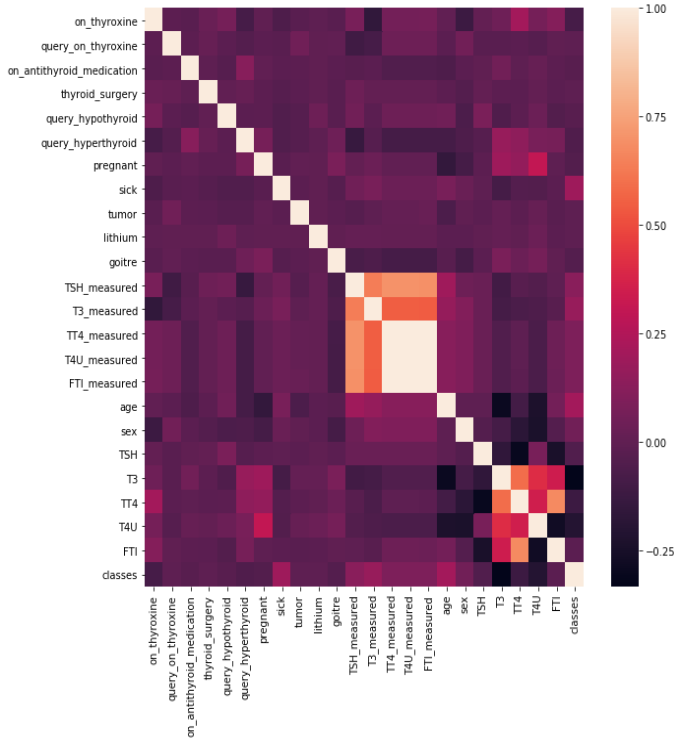

Fig.2.Correlation matrix between features of thyroid disorder dataset

Machine learning (ML) is a subset of artificial intelligence that is infiltrating scientific study in ever-increasing ways. Machine learning allows models to learn from their mistakes without being prioritized [21]. Classic epidemiology is an advanced integrated modern data science strategy to strap the capabilities of the cultured data [22]. Machine learning has been caused by the input explosion that is related with an enhancing computing ability. To analyze large amounts of data, the programmed investigates nearby clinically significant connections between input and output criteria. Machine learning allows computers to make precise prognosis on instant data based on previous data. The descriptive aspect creates very reliable prophecy methods that can replicate previously novel communication in large, complex data sets and adapt to effective data aura.

IV. Proposed System

The dataset used for the experiment is collected from UCI repository site and null values of the dataset has been removed in data pre-processing phase. Then features of the dataset were selected which are required for thyroid disease evaluation and three machine learning models are implemented on the dataset for predicting the thyroid patient as hypothyroidism or normal with significant features as shown in fig 4.

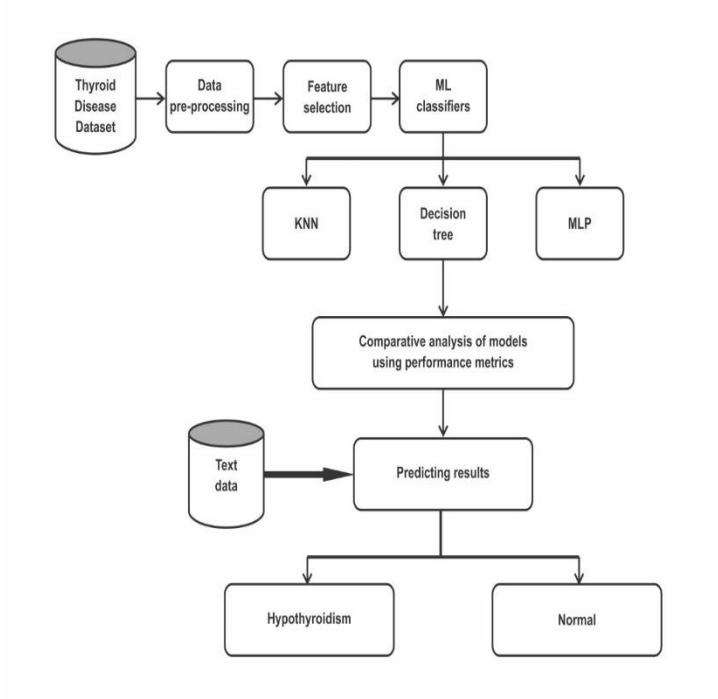

Figure 4. Block diagram for prediction of thyroid

Accuracy considers the percentage of correctly assigned positive and negative classes. AUC (Area under the ROC Curve) is a composite performance statistic that considers all possible categorization levels.

V. Machine Leaning Algorithms used in the study

The algorithms used for the experiment was explained in concise manner

A.K-NN

It stands for K-nearest neighbor which is based on supervised algorithm. It is used to solve classification and regression problems technique. The object is classified depending upon the nearest neighbor using the classification technique. The calculation of the nearest neighbor is measured using the Euclidean distance.

$$Euclidean\ Distance, d(a,b)^2 = (b_1 - a_1)^2 + (b_2 - a_2)^2$$

Here, the input consists of the closest or nearest neighbor in the dataset for model deploying. The classifier assumes the similar attributes existing in closer proximity. After loading data and choosing the nearest neighbor, the distance between query and original example is calculated and numbers of entries are sorted in the collection [24].

B.Multi layer perceptron

A multilayer perceptron network, as shown in Figure 3 frequently utilized for pattern recognition, input pattern classification, and other comparable tasks [23]. MLP consisting of three layers: an input layer, a hidden layer, and an output layer. The input is passed through the weight function in the input layer, and then the nonlinearity function is injected in the hidden layer. At each hidden layer neuron, a weight (wji) is multiplied by the value from each input neuron. After that, each hidden layer neuron weights are put together to form a composite value. The weighted sum is then fed into a transfer function, which produces the value outputs.

The neurons in the output layer (o1, o2) receive the summed outputs from the hidden layer neurons.

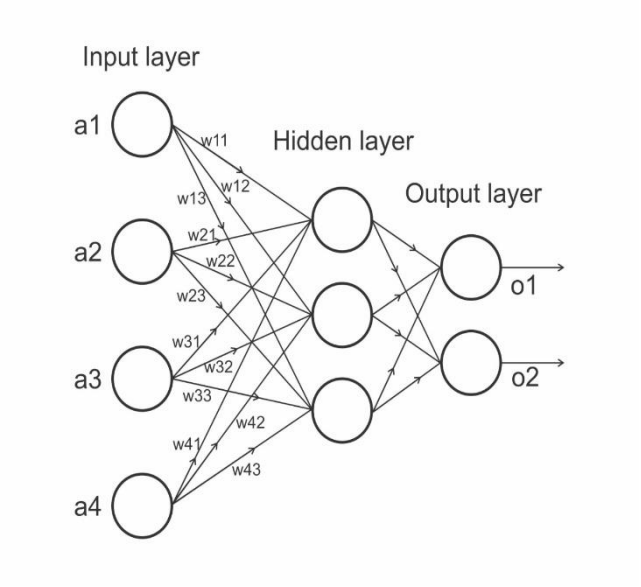

Fig.3.Schematic diagram of multi-layer perceptron

C. Decision tree

This classifier is based on classification algorithm while work on numerical as well as categorical data. It is required for creating tree shaped graph while analyzing the data. The analysis of decision trees is based on three nodes and those are root node, interior node and leaf node.

The idea behind the decision algorithm including the best attribute using information gain, gain ratio. It makes that attribute a decision tree and breaks into sub-datasets. Further, it starts building the tree and process repetition recursively [25][27].

Information gain,

$$Information(M) = -\sum_{i=1}^{n} \pi \log_2 \pi$$

$$Information_A(M) = \sum_{j=1}^{m} \frac{M_J}{M} XInformation(M_j)$$

Gain ratio,

$$Split_A(M) = -\sum_{j=1}^{m} \frac{M_J}{M} \log_2 \frac{M_J}{M}$$

$$GainRatio(N) = \frac{Gain(N)}{Split_A(M)}$$

VI. Simulation Result and Analysis

K-NN, MLP, DT are the three ml models which are simulated using thyroid disease dataset.70% and 30% of the data are used for training and testing the models. Each of these model's performance have been estimated in terms of accuracy and AUC as given in table 2. Table 2 clearly shows that the multilayer perceptron model 95.72percentage accurately predict the thyroid disease. Decision tree of 90.96, and K-NN of accuracy 90.96 are obtained for forecasting thyroid illness. Similarly, the diagnostic rate of MLP has been obtained as 94.23 followed by decision tree with an AUC value of 87.94 followed by K-NN with an AUC of 57.11 respectively.

Parameters used in the algorithms

For the simulation we have taken k=4, minkowski distance metric are used for implementation of K-NN model. Hidden layer size 40, learning rate =0.001, adam optimizer, relu activation function, maximum iteration=200 are used for implementation of MLP model. Entropy criterion, maximum depth =10,10-fold cross validation has used for implementation of decision tree algorithm.

Table 2. Accuracy and AUC score of ml models

ML models	Accuracy	AUC
K-NN	90.96	57.11
Decision tree	94.18	87.94
MLP	95.72	94.23

Accuracy and AUC comparison of k-NN, decision tree and MLP classifiers have been shown in Fig 5 and 6. It clearly shows that MLP performs better in predicting thyroid disease as compared to K-NN and decision tree.

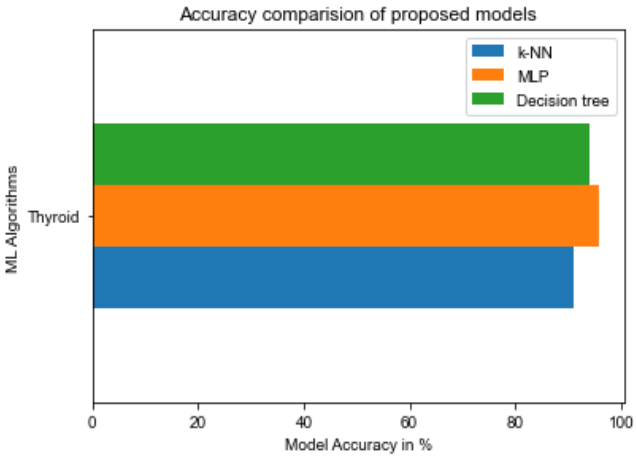

Figure 5. Accuracy Comparison of ml models

The figure 5 shows that MLP achieved maximum accuracy of 95.72 percent, subsequently decision tree and K-NN, which have 94.18 percent and 90.96percent accuracy, respectively in prediction of thyroid disease.

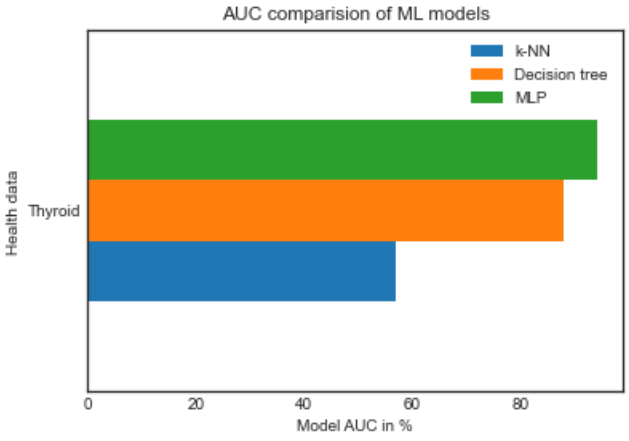

Figure 6.AUC Comparison of ml models

MLP achieves the greatest AUC of 94.23 percent, followed by decision tree and K-NN with AUC values of 87.94 percent and 94.23 percent as shown in figure 6.

Table 3. Performance comparison of proposed work with the existing work

Authors	ML models	Accuracy	AUC
Kaur et al. [10]	KNN Decision tree MLP	61.45 66.57 66.95	65.8 74.0 80.7
Turanoglu-Bekar et al. [20]	NBTREE LADTREE REPTREE BFTREE	75.00 66.25 62.50 65.00	NA
Chaubey et al. [28]	Decision tree	87.5	NA
Aversano et al. [29]	ET	84	NA
Proposed work	**KNN** **Decision tree** **MLP**	**90.96** **94.18** **95.72**	**57.11** **87.94** **94.23**

Performance of existing work have been compared with the proposed work in table 3. Kaur et al. predict thyroid disease with an accuracy value of 61.45,66.57,66.95 using KNN, Decision tree, MLP machine learning classifiers respectively. They have obtained the AUC values of 65.8,74.0,80.7 respectively. Turanoglu-Bekar et al. predict thyroid disease using NBTREE, LADTREE, REPTREE, BFTREE ml classifiers with accuracy of 75.00,66.25,62.50,65.00 respectively. Aversano et al predict thyroid disease using machine learning classifiers with an accuracy of 84%. Chaubey et al. predict thyroid disease using decision tree with an accuracy value of 87.5. Aversano et al. predict thyroid disease using extra tree and they have obtained an accuracy of 84%. We have compared our work with existing work and we found that proposed work performs better in prediction of thyroid disease in terms of accuracy and AUC.

Limitation of ML classifiers

i)K-NN

K value should be properly selected

For large sample size it is inefficient due to large run time

ii)Decision tee

It is more prone to outliers

With large sample size tree structure became complex

iii)MLP

Number of parameters is high with the increase of number of layers which makes the model complex.

VII. Conclusion

Machine learning algorithms are now employed in medical decision-making to eliminate human errors and assist professionals in examining biomedical data in greater detail. In this work we have compared three machine learning algorithms namely KNN, Decision tree, MLP for prediction of thyroid disease and obtained accuracy of 90.96,94.18,95.72 and diagnosis rate of 57.11,87.94,94.23 respectively. We found that MLP performs better in classification of thyroid disease as compared to other two algorithms. We have also compared our work with other existing work and found our proposed work performs well in prediction of thyroid disease as compared to other work. The used machine learning model can also be used for classification of other chronic disease such as heart disease, diabetes, cancer, covid-19 etc. with higher accuracy and diagnosis rate. Using deep learning method accuracy of prediction of thyroid disease can also be improved.

References

[1] Miller, K.D., et al.: Cancer treatment and survivorship statistics, 2016. CA Cancer J. Clin. **66**(4), 271–289 (2016)

[2] K. Polat, S. Sahan and S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis," *Expert Systems with Applications*, Vol. 32, 2007, pp. 1141-1147.

[3] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM," in *3rd International Conference on Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009*.

[4] L. Ozyılmaz and T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods," in: Proceedings of ICONIP'02 9th international conference on neural information processing (Singapore: Orchid Country Club, 2002) pp. 2033–2036.

[5] http://www.foxnews.com/health/2012/02/10/ hypo thyroidism - versuhyperthyroidism.html (accessed dec 2015)

[6]http://www.emedicinehealth.com/thyroid_faqs/articl e_em.htm

[7] Oz Yilmaz, Lale, and Tulay Yildirim. "Diagnosis of thyroid disease using artificial neural network methods." Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on. Vol. 4. IEEE, 2002.

[8] Mousavi. S, Zanjireh.M, Oghbaie.M" Applying computational classification methods to diagnose Congenital Hypothyroidism: A comparative study" Informatics in Medicine Unlocked 18 (2020) 100281

[9] S. Sathya Priya, Dr. D. Anitha" Survey on Thyroid Diagnosis using Data Mining Techniques" International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Special Issue 1, January 2017.

[10] Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Multimedia Tools and Applications. 2019; 78:19905–19916.

[11] Dogantekin, Esin, Akif Dogantekin, and Derya Avci. "An automatic diagnosis system based on thyroid gland: ADSTG." Expert Systems with Applications 37.9 (2010): 6368-6372.

[12] Yeh, Wei-Chang. "Novel swarm optimization for mining classification rules on thyroid gland data." Information Sciences 197 (2012): 65-76.

[13] Azar, Ahmad Taher, Shaimaa Ahmed El-Said, and Aboul Ella Hassanien. "Fuzzy and hard clustering analysis for thyroid disease." Computer methods and programs in biomedicine 111.1 (2013): 1-16.

[14] Stegmayer, Georgina, Matias Gerard, and Diego H. Milone. "Data mining over biological datasets: An integrated approach based on computational intelligence." Computational Intelligence Magazine, IEEE 7.4 (2012): 22-34.

[15]UrRehman, Chyi-Yeu Lin,Zohaib Mushtaq &Shun-Feng Su"Performance Analysis of Machine Learning Algorithms for Thyroid Disease" *Arabian Journal for Science and Engineering* volume 46, pages 9437–9449(2021)

[16] K. Geetha & Capt. S. Santhosh Baboo" An Empirical Model for Thyroid Disease Classification using Evolutionary Multivariate Bayesian Prediction Method" Global Journal of Computer Science and Technology: E Network, Web & Security Volume 16 Issue 1 Version 1.0 Year 2016ISSN: 0975-4350

[17]. Prasad, V.; Rao, T.S.; Babu, M.S.P.: Thyroid disease diagnosis hybrid architecture composing rough data sets theory and machine learning algorithms. Soft Compute. **20**(3), 1179–1189(2016)

[18]. Mushtaq, Z.; Yaqub, A.; Sani, S.; Khalid, A.: Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. J. Chin. Inst. Eng. **43**(1), 1–13 (2020)

[19] Tomar, D.; Agarwal, S.: A survey on data mining approaches for healthcare. Int. J. Bio-Sci. Bio-Technol. **5**(5), 241–266 (2013)

[20]. Jahantigh, F.F.: Kidney diseases diagnosis by using fuzzy logic. In: 2015 International Conference on Industrial Engineering and Operations Management, 2015 (IEOM2015), pp. 2369–2375. IEEE (2015)

[21] Obermeyer Z, Emanuel EJ. Predicting the future— big data, machine learning, and clinical medicine. N Engl J Med. 2016; 375:12161219.

[22] Breiman L. Statistical Modeling: the two cultures. Stat Sci. 2001; 16:199-231.
[23] Sonawane, Jayshril S.; Patil, D. R. (2014). *[IEEE 2014 International Conference on Information Communication and Embedded Systems (ICICES) - Chennai, India (2014.2.27-2014.2.28)] International Conference on Information Communication and Embedded Systems (ICICES2014) - Prediction of heart disease using multilayer perceptron neural network., (), 1–6.* doi:10.1109/icices.2014.7033860
[24] https://link.springer.com/article/10.1007/s42979-020-00365-y
[25]https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees
[26] R. Rajkumar, K. Ananda Kumar, A. Bharathi, Coronary artery disease (CAD) prediction and classification—a survey, ARPN J. Eng. Appl. Sci. 11 (9) (2006) 5749–5754.
[27] Bekar, E.T.; Ulutagay, G.; Kantarcı, S.: Classification of thyroid disease by using data mining models: a comparison of decision tree algorithms. Oxf. J. Intell. Decis. Data Sci. **2016**(2), 13–28
[28] Chaubey, G., Bisen, D., Arjaria, S., & Yadav, V. (2020). *Thyroid Disease Prediction Using Machine Learning Approaches. National Academy Science Letters, 44(3), 233–238*
[29]L.Aversano.Bernadi,M.Cimitile,M.Lammarino,P.M.Macchia,I.C.Nettore,C.Verdone,"Thyroid disease treatment prediction using machine learning approaches" procodia computer science,192,2021,1031-1040
[30] Yadav, Dhyan Chandra; Pal, Saurabh (2020). *Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration,* doi:10.1007/s42454-020-00006-y