5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, India

Interactive Thyroid Disease Prediction System Using Machine Learning Technique

Ankita Tyagi
Computer applications
Dit University
Dehradun, India
ankitatyagi26@gmail.com

Ritika Mehra
Computer applications
Dit University
Dehradun, India
hod.mca@dituniversity.edu.in

Aditya Saxena
Computer Science and engineering
Dit University
Dehradun, India
aditya.kishore@dituniversity.edu.in

***Abstract*—Thyroid disease is a major cause of formation in medical diagnosis and in theprediction, onset to which it is a difficult axiomin the medical research. Thyroid gland is one of the most important organs in our body. The secretions of thyroid hormones are culpable in controlling the metabolism.Hyperthyroidism and hypothyroidism are one of the two common diseases of the thyroid that releases thyroid hormones in regulating the rate of body's metabolism. Data cleansing techniques were applied to make the data primitive enough for performing analytics to show the risk of patients obtaining thyroid. The machine learning plays a decisive role in the process of disease prediction and this paper handles the analysis andclassificationmodels that are being used in the thyroid disease based on the information gathered from the dataset taken from UCI machine learning repository. It is important to ensure a decent knowledge base that can be entrenched and used as a hybrid model in solving complex learning task, such as in medical diagnosis and prognostic tasks. In this paper, we also proposed different machine learning techniques and diagnosis for the prevention of thyroid. Machine Learning Algorithms, support vector machine (SVM), K-NN, Decision Trees were used to predict the estimated risk on a patient's chance of obtaining thyroid disease.**

Keywords—Thyroid Disease, Prediction Model, Machine Learning Algorithms.

I. INTRODUCTION

The advancement of computational biology is used in the healthcare industry. It allowed collecting the stored patient data for the medical disease prediction. There are different intelligent prediction algorithms are available for the diagnosis of the disease at early stages. The Medical information system is rich of data sets, but there are no intelligent systems that can easily analysis the disease. Over the course of time, machine learning algorithmsplay a crucial role in solving the complex and nonlinear problems in developing a prediction model. In any disease prediction models are required to paramount the features that canbe selected from the different datasets which can easily be used as a classificationin healthy patient as precisely as possible. Otherwise, misclassification may result in a healthy patientthat endures unnecessary treatment. Hence, the factuality of predictingany disease in conjunction with the thyroid disease is of supreme cardinality.

The thyroid gland is an endocrine gland in the neck. It erects in the lessened part of the human neck, beneath the Adam's apple which aids in the secretion of thyroid hormones and that basically influences the rate of metabolism and protein synthesis. To control the metabolism in the body, thyroid hormones are useful in many ways, counting how briskly the heart beats and how quickly the calories are burnt.The composition of thyroid hormones by the thyroid gland helps in the domination of the body's metabolism. The thyroid glands are composed of two active thyroid hormones, levothyroxine (abbreviated T4) and triiodothyronine (abbreviated T3).To regulate the temperature of the body these hormones are imperative in the fabrication and also in the comprehensive construction and supervision [1][2]. Specifically,thyroxin (T4) and triiodothyronine (T3) are the two types of active hormones that are customarily composed by the thyroid glands [1][2][3]. These hormones are decisive in proteinmanagement; dissemination in the body temperature, along with the energy-bearing and transmission in every part of the body. For these two thyroid hormones i.e. (T3 and T4), iodine is considered as the main building chunk of the thyroid glands and are prostrated in a few specific problems, some of which are exceptionally prevalent. Insufficiency of thyroid hormones elements to hypothyroidism as well as an excessive thyroid hormones element to Hyperthyroidism. There are many origins related to hyperthyroidism and underactive thyroids. There are various kinds of medications. Thyroid surgery is liable to ionizing radiation, continual tenderness of the thyroid, deficiency of iodine and lack of enzyme to make thyroid hormones [4].

II. LITERATURE SURVEY

In the latter years, there has been a lot of work done to diagnose the discrete diseases in thyroid. Many authors have used various kinds of data mining technique. The authors proved to obtain an adequate approach and certainty to find out the diseases analogous to the thyroid by the work that includes various datasets and algorithms linked with the work that is to be done in the future perspective to accomplish effective and better results. The intent of the paper interprets various techniques of data mining mechanisms and the statistical attributes that is been popularized in the latter years for interpretation of thyroid diseases with the certainty by various authors to attain various prospects and for various approaches.There are various algorithms of machine learning counting random forest, decision tree, naïve Bayes, SVM and ANN that are extensively used in the frequent diseases and in the prognostic problems. There are few functions that are comprised of diseases related to heart disease[5], diabetes, Parkinson's, hypertension, the Ebola virus(EV) [19-20], diagnoses and forecasting, R-NA sequenced data analysis and allocation of biomedical imaging[23-25].Despite, the advancement of a machine learning-placed disease

prediction mechanismanda medical determination is a nontrivial task .There is essential issues i.e. acquisition of data, compilation and grouping that are worn to train the machine learning structures.In the actual activity issues, estimation of large data sets in biomedical over a deep continuation are desired, and are essentially non-existent [12].

In [26] systematic approach for earlier diagnosis of Thyroid disease using back propagation algorithm used in neural network. ANN delicateandestablisheson back propagation of an error that isbeingusedforpriordiseasepredictions. The impact of ANN is being trained with the empirical data and testing mechanisms that are borne out as a data that was not in use during the process of training.ANN concludes in a good compliance with the preliminary data and indicates the advanced neural network which uses as a substitute for prior disease predictions.

In the authors scrutinized and comparethe four classification models namely Naive Bayes, Decision Tree, Multilayer Perceptron and Radial Basis Function Network. The conclusion demonstrates a momentous accuracy for all the classification models. The Decision Tree model exceeds by the other classification models. In this work 29 attributes of datasetis conscripted and enforced as a Feature Selection technique i.e. Chi-Square, The datasets are being filtered by conducting the unsupervised coated filters on the attributes for conversion in the continuous values into nominal and hence reduce the 29 attributes to 10 attributes.

Machine learning (ML) is a division of artificial intelligence and is infiltrated in the dimensions of scientific research at growing steps. Machine learning facilitates algorithms to review from experience without notably being prioritized [6]. Machine learning has been induced by the input detonation that is connected with an expanding computational capability, and classical epidemiology are an advanced blended recent data science approach to strap the capabilities of the cultured data [7]. To consider vast arrangements of data, the particular toolsexplores in nearby clinically relevant liaison between input and output criterion. Factual analyses of surgical conclusions are eminently deceivable to amend surgical accords. Decisive aspects of surgical accords are description of the patient's comrade that aids from surgery in the arbitration. Machine learning enables computers to determine from preceding data to make meticulous predictions on current data. The informative facet makes very authoritative prediction algorithms that can copy the formerly exotic communication in vast, convoluted sets of data and acclimate to effective data aura [16].

The composite characteristics and the curative procedures that are being used in the thyroid disorders cater an ample clustering of intricate and assorted data and hence, a propitious framework for the formulation of machine learning models [15]. This proposes an ample probable for the utilization of machine learning models and braces a flourishing tendency towards rigorous medicines in which therapeutics are sewn to the particular patients. In the field of machine learning, an extensive divergence could be contrived amid supervised and unsupervised learning. Supervised learning algorithms determine from "labelled" training data to crop a model that accomplishes predictions on formerly imaginary data [8]. For unsupervised mechanism of learning, only unlabelled data are feasible and the algorithms peeks to asset the analogies and devices, unsupervised learning algorithms may catch the vast number of unlabelled genomics data as input and analyze formerly anonymous assemblage of data. These algorithms may somehow be dominant in previously formerly arrangements in complex data that are not primarily measurable by humans [9] and may be used to develop labels to finally train a supervised model. In conventional programming, a programmer manually creates a set of information – "the programs" – to develop a crave output from a given set of input variables. In machine learning, the inputs are equipped together with the crave output and computer algorithms are inquired to derive the "rules from the classified training data". The computerized learning process is an adequate way of interpreting vast abundance of data, designing concealed communications in composite sets of data, and allering to dynamic aura [10-13].

In the learning mechanism, algorithms endeavor to asset the excellent aggregation of input variables (features) and weights are included to these features in the model, by that diminishing the disparity between the anticipated and substantial results.Machine learning is used in training the system over vast databases, where the enforced machine learning techniques are recycled to develop abstraction devices or frame a model and use the accomplished devices or frame a model and use the accomplished devices or models in making predictions in the future for anonymous cases [11-14].

III. ARCHITECTURE OF THYRIOD PREDICTION SYSTEM

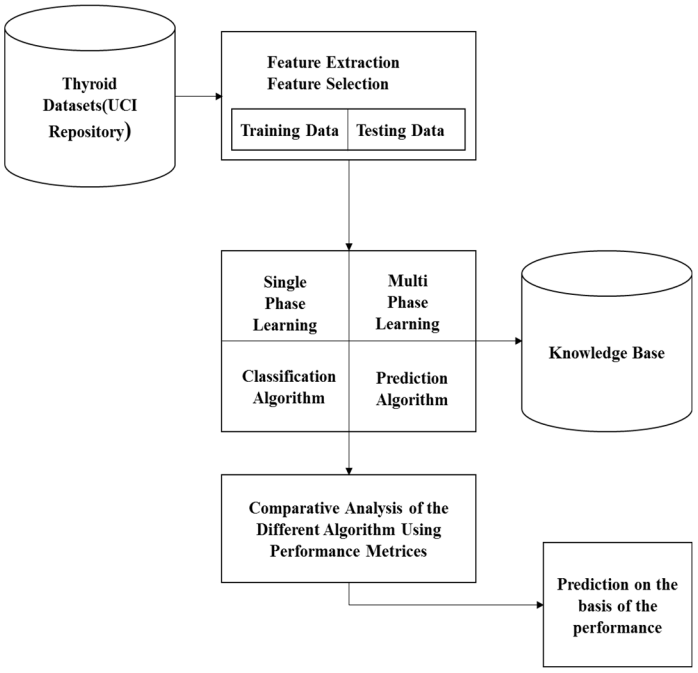


Fig1. Thyroid Prediction System

IV. METHODOLOGY

Supervised learning is an information miningundertaking of inferring a function from namedtraining information. The training informationcomprised of an arrangement of preparingillustrations. In managed adapting, every case is acouple comprising of an information input object (commonly a vector) and the desired output value(additionally called the supervisory flag).A supervised learningcalculation investigates thetraining information and produces an indirectfunction, which can be utilized formappingnewillustrations [17]. An idealimprovement will take intoaccount the calculation to effectively decide theclass names for unseen cases. This requires thetaking in calculation to sum up from the traininginformation to hidden circumstances in a "sensible"manner.

A. Attributes Usedto diagnose thyroid Diseases:

By analyzing the above research work it is found that frequently used medical attributes to perform experimental work for the diagnosis of thyroid diseases are given below in below table no.1.Among these attributes almost every researcher has selected attributes to perform work for thyroid disease diagnosis.

TABLE I. ATTRIBUTE FOR THE FEATURE SELSECTION

Attributes	Description
Age	In years
Sex	Male or female
TSH	Thyroid-Stimulating Hormone
T3	Triiodothyronine
TBG	Thyroid binding globulin
T4U	Thyroxin utilization rate
TT4	Total Thyroxin
FTI	Free Thyroxin Index

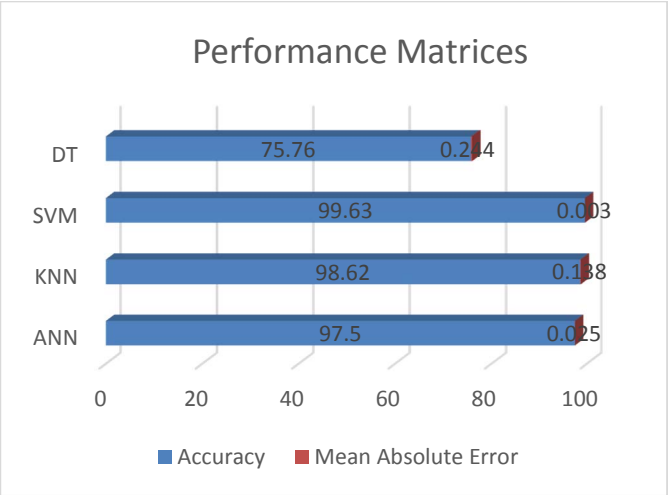
B. Performance Study of the proposed Algorithms:

1. Artificial Neural Network: Neural network provides accustomed and a pragmatic approach in training the absolute, discrete as well as vector valued functions and is a parallel system based on nervous system for learning real -valued, discrete-valued and vector-valued functions and is a parallel system based on human that have numerous corresponding alter elements basically known to be as the neurons,working in a consensus way to solve definite problems.Backpropagation is the most frequently worn learning techniques in ANN. It is a three-layered architecture that is placed in the algorithms in the neural networks. It is comprised of 3 layered architecture i.e. input layer, hidden layer and an output layer. The foremost layer that is the input layer fueled the inputs into this layer, the second layer i.e. a hidden layer- accords the output from the input layer and lastly an output layer, beams the network's prediction. This miniature network helps to classify thenew data.
2. Support Vector Machine:Support vector machine is considered as an assorted research algorithm that helps in performing the analysis in a precise way. Support vector machine is an approach that is commenced with a concept of an ace of separating hyper planeto aid in the distribution for sampling of data. A hyper plane or multiple planes are created by the support vector machine classifier in high dimensional space. The training data samples are being separated as a positive and negative data samples by the hyper plane.
3. Decision Tree: Tree-like graph is used in decision tree classifier. A decision tree is classified by its 3 nodes i.e. internal nodes, leaf nodes, and the root nodes. The internal node connotes as the test on an attribute, the leaf node connotes as the distribution of the class and the root node connotes as the tree that has the top most node. The two most extensivealgorithms that are used in the as semblances of a decision tree for diagnostic and prognostic model of thyroid diseases are C4.5 and ID3. Researchers use Decision Tree widely in healthcare field particularly to diagnose various thyroid diseases [18].
4. K- Nearest Neighbor: When given a training tuple K-Nearest Neighbor simply stores it and waits until it is given a test tuple. Hence it is a "lazy learner" as it stores the training tuples or the "instances", they are also known as "Instance- Based Learners" [21-22]. Thus, k is a positive integer and decides how many neighbors influence the classification. "Closeness" delineates as a distance metric such as "Euclidean Distance" or "Manhattan Distance".

V. RESULT AND DISCUSSION

The data sets for the thyroid diseases have been possessed from the UCI machine learning repository. The work is dwelled with 2 different stages. The foremost phase comprised of the subset selection that is executed byadapting mutual information and prediction of the thyroid datasets done usingANN. Specifically in the interpretation of diseases neural networks are successfully enforced in the distinctive fields in the medical realm.The certainty of the analysis for the datasets of the thyroid diseases are assigned as the elected appearance by every feature selection algorithm.

Algorithm Used	Accuracy	Mean Absolute Error
ANN	97.50	0.025
KNN	98.62	0.138
SVM	99.63	0.003
DT	75.76	0.244



Indeed, the expansion of our unified representative is a constructive mechanism to predictthyroid disease based on the limited dataset that isavailable with us. The model can further beenhanced to any desired level by increasing thenumber of inputs and outputs and dynamic data canbe generated as more data can be fed to it. In anutshell, not only we have developed a prototypeintegratedframe work to diagnose the thyroiddisease but also act as a decision maker fordiagnosing the thyroid disease.

VI. Conclusion

The intent of our work to be done further is to cater the research of idiosyncratic techniques of machine learning that can be mobilized in the diagnosis of thyroid diseases. There are numerous approachable analyses that are delineated and are being used in the latter years of adequate and competent thyroid disease diagnosis. The analysis shows that different technologies are used in all the papers showing different accuracies. In most research papers it is shown that neural network outperforms over other techniques. On the other hand, this is also given that support vector machine and decision tree has alsoperformed well. There is no doubt that researchersworldwide have attained a lot of success to diagnose thyroid diseases, but it is suggested to decrease the number of parameters used by the patients for diagnosis of thyroid diseases. More attributes mean a patient has to undergo a greater number of clinical tests which is both cost effective as well time consuming. Thus, there is a need to develop such type of algorithms and thyroid disease predictive models which require minimum number of parameters of a person to diagnose thyroid disease and saves both money and time of the patient.

Acknowledgment

The author thanks the DIT University, Dehradun for providing the research grant to support this research work. The corresponding author wishes to thanks Prof K.K. Raina and Prof S.K. Gupta for the great cooperation and motivation for this research.

References

[1] 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.

[2] K. Polat, S. Sahan and S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted preprocessing 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 andBiomedicalEngineering, 2009. ICBBE 2009.*

[4] G. Zhang, L.V. Berardi, "An investigation of neural networks in thyroid function diagnosis," *Health Care Management Scien*ce, 1998, pp. 29-37. Available: http://www.endocrineweb.com/ thyroid.html,(Accessed: 7 August 2007).

[5] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, 2012.

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

[7] Breiman L. StatisticalModeling: the two cultures.Stat Sci. 2001;16:199-231..

[8] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. 2017; 9:245-250

[9] Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015;521: 452-459.

[10] Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. J Neurol Neurosurg Psychiatry. 2015; 86:251-256.

[11] Deo RC. Machine learning in medicine. Circulation .2015;132: 19201930.

[12] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, D.S. Lee, Using methods from the data-miningand machine-learning literature for disease classification and prediction: a case studyexamining classification of heart failure subtypes, J. Clin. Epidemiol. 66 (4) (2013)398–407.

[13] A.K. Pandey, P. Pandey, K.L. Jaiswal, A heart disease prediction model using DecisionTree, IUP J Comput. Sci. 7 (3) (2013) 43.

[14] S. Ismaeel, A. Miri, D. Chourishi, in: Using the Extreme Learning Machine (ELM) techniquefor heart disease diagnosis, IEEE Canada International Humanitarian TechnologyConference, 2015, pp. 1–3.

[15] L. Verma, S. Srivastava, P.C. Negi, A hybrid data mining model to predict coronaryartery disease cases using noninvasive clinical data, J. Med. Syst. 40 (7) (2016) 1–7.

[16] R. Rajkumar, K. Anandakumar, A. Bharathi, Coronary artery disease (CAD) predictionand classification—a survey, ARPN J. Eng. Appl. Sci. 11 (9) (2006) 5749–5754.

[17] Y.T. Lo, H. Fujita, T.W. Pai, Prediction of coronary artery disease based on ensemblelearning approaches and co-expressed observations, J. Mech. Med. Biol. 16 (01) (2016) 1640010.

[18] B. Farran, A.M. Channanath, K. Behbehani, T.A. Thanaraj, Predictive models to assessrisk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms andvalidation using national health data from Kuwait—a cohort study, BMJ Open 3 (5)(2013):e002457.

[19] M. Heydari, M. Teimouri, Z. Heshmati, S.M. Alavinia, Comparison of various classificationalgorithms in the diagnosis of type 2 diabetes in Iran, Int. J. Diabetes Dev. Countries36 (2) (2015) 167–173.

[20] 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.

[21] A. Jabeen, N. Ahmad, K. Raza, Machine learning-based state-of-the-art methods for theclassification of RNA-Seq data, in: N. Dey, A. Ashour, S. Borra (Eds.), Classification inBioApps. Lecture Notes in Computational Vision and Biomechanics, vol. 26, Springer,Cham, 2018.

[22] S.S. Ahmed, N. Dey, A.S. Ashour, D. Sifaki-Pistolla, D. Ba˘las-Timar, V.E. Balas, J.M. Tavares, Effect of fuzzy partitioning in Crohn's disease classification: a neuro-fuzzybasedapproach, Medical Biol. Eng. Compute. 55 (1) (2017) 101–115.

[23] Enas M.F. El Houby, A survey on applying machine learning techniques for management of diseases, Journal of Applied Biomedicine, January 2018

[24] Crohn's disease classification: a neuro-fuzzybasedapproach, Medical Biol. Eng. Compute. 55 (1) (2017) 101–115.

[25] Prerana, Parveen Sehgal, Khushboo Taneja"Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network" International Journal of Research in Management, Science & Technology Vol. 3, No. 2, April 2015.

[26] 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.