IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

THYROID CANCER DETECTION USING MACHINE LEARNING TECHNIQUES

[1]R.Vanitha, [2]Dr. K. Perumal
[1]Research Scholar, [2] Professor,
Department of Computer Applications,
Madurai Kamaraj University,
Madurai.

Abstract: In the recent world of Digital and Technology, the level of disease rate also increases with time. According to the World Health Organization, the second most common endocrine disorder in the world is related to thyroid gland diseases which is next to diabetes. Hypothyroidism or hyperthyroidism is a major issue in India which rises due to non-functional thyroid hormones. This diagnosis of Thyroid Cancer is very monotonous and tough tasks at early steps with accuracy. The Accuracy prediction can be done through various Machine Learning, Algorithms. This paper proposes on Analysis of the thyroid dataset through Naïve Bayes, logistic regression, K-Neighbour Support Vector Machine, Decision Tree Classifier, Artificial Neural Network and AdaBoost classifier. Logistic regression and Support Vector Machine classifier, producing complete and better accuracy through WEKA Tool**.**

***Index Terms* - Thyroid Cancer, Accuracy, Prediction, Machine Learning, Algorithms.**

I. INTRODUCTION

Hypersecretion of thyroxine ie Hyperthyroidism affects about 2% of individuals, while hyposecretion of thyroxine in Hypothyroidism affects about 1% of individuals. Both these disorders may be mostly caused by thyroid gland dysfunction, at times it may be due to pituitary gland failure which is considered a secondary problem and when it is due to hypothalamic malfunction is considered a tertiary disorder.

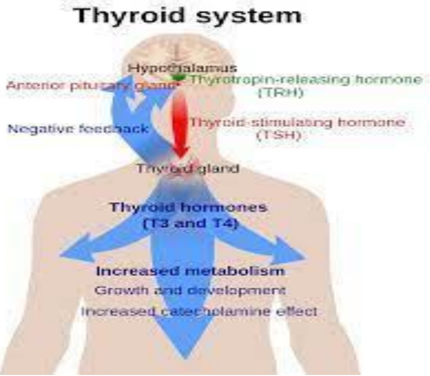

Thyroid system
Hypothalamus
Anterior pituitary gland
Thyrotropin-releasing hormone (TRH)
Negative feedback
Thyroid-stimulating hormone (TSH)
Thyroid gland
Thyroid hormones (T3 and T4)
Increased metabolism
Growth and development
Increased catecholamine effect

Figure 1. shows structure of Thyroid System

Machine Learning is a set of tools applied for the construction and assessment of algorithms that simplify prediction, pattern acknowledgment and classification.ML purpose of aggregation data, selection of the model, training the prototypical and testing the prototypical. In this study we use six machine learning classifiers which are Naïve Bayes, K-Neighbour, Support Vector Machine, Decision Tree Classifier, Artificial Neural Network and AdaBoost.

Machine learning engagements six main different algorithms those are explained below:

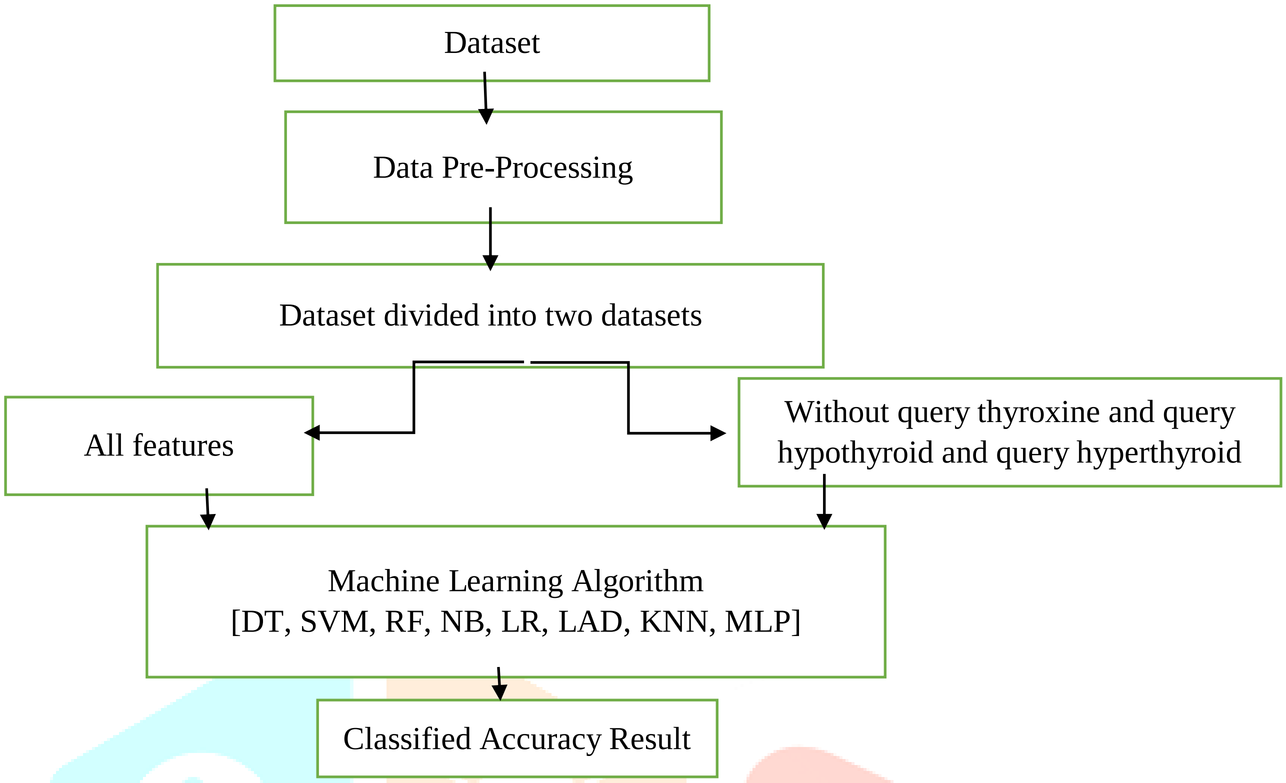

Dataset
Data Pre-Processing
Dataset divided into two datasets
All features
Without query thyroxine and query hypothyroid and query hyperthyroid
Machine Learning Algorithm
[DT, SVM, RF, NB, LR, LAD, KNN, MLP]
Classified Accuracy Result

Figure.2. Shows the data process using ML Algorithms

II. Machine Learning Techniques

1. Artificial Neural Network: Neural network delivers the comfortable and a logical method in training the complete, discrete as well as vector valued purposes and it is a parallel system based on nervous system for knowledge 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 an agreement way to solve definite problems. Backpropagation is the most commonly worn learning technique in ANN. It has three-layered architecture such as input layer, hidden layer and an output layer in the neural networks.

2. Support Vector Machine: Support vector machine is one of the most popular and measured as a various research algorithm that helps in accomplishing the study in an accurate way. SVM is designated to separate the arguments in the input variable space by their class, either 0 or class 1.A hyper plane or multiple planes are fashioned by the support vector machine classifier in high dimensional space. The training data samples are being divided as a positive and negative data sample by the hyperplane.

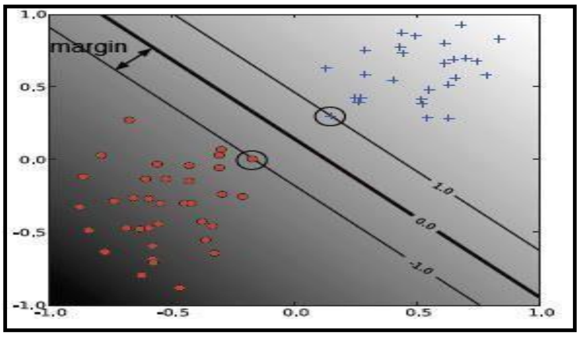


Figure 3. Example of a two-class linear SVM classifier

3. Decision Tree: The Tree just like a graph is used in the decision tree for classifying the data. A decision tree is concluded by its 3 nodes i.e., internal nodes, leaf nodes, and the root nodes. The internal node committed as the test on an attribute, the leaf node committed as the distribution of the class and the root node committed as the tree that has the top most node. Tree is too fast to learn and make decisions. There are two most general algorithms that are used in the as appearances of a decision tree for analytic and predictive models of thyroid diseases are C4.5 and ID3.

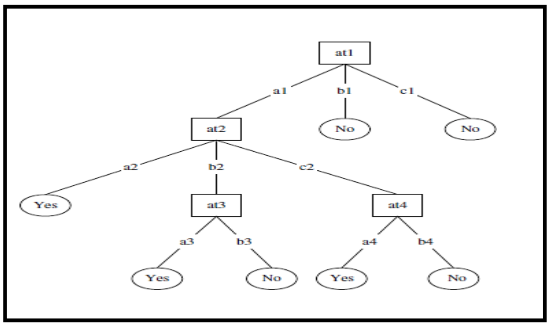


Figure 4. Shows the structure of Decision tree algorithm.

4. K- Nearest Neighbour: KNN is a simple and effective classifier. Predictions are creating a new data point by finding through the entire training set for the neighbours and summarizing the output variable for those instances. Euclidean distance, a number we can calculate straight based on the alterations between each participation variable.

5. Naive Bayes: Naïve bayes classifiers are a team concerning straightforward probabilistic and surprisingly powerful classifiers. This model has two types, one is the probability of each class and another one is conditional probability for each class given each x value. They are among the least complicated Bayesian delivery models. It shares a traditional government as the closeness about a specific thing in a category lamely after the proximity of fractional mean elements.

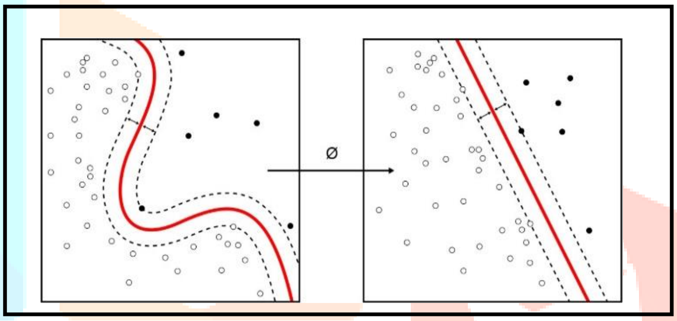
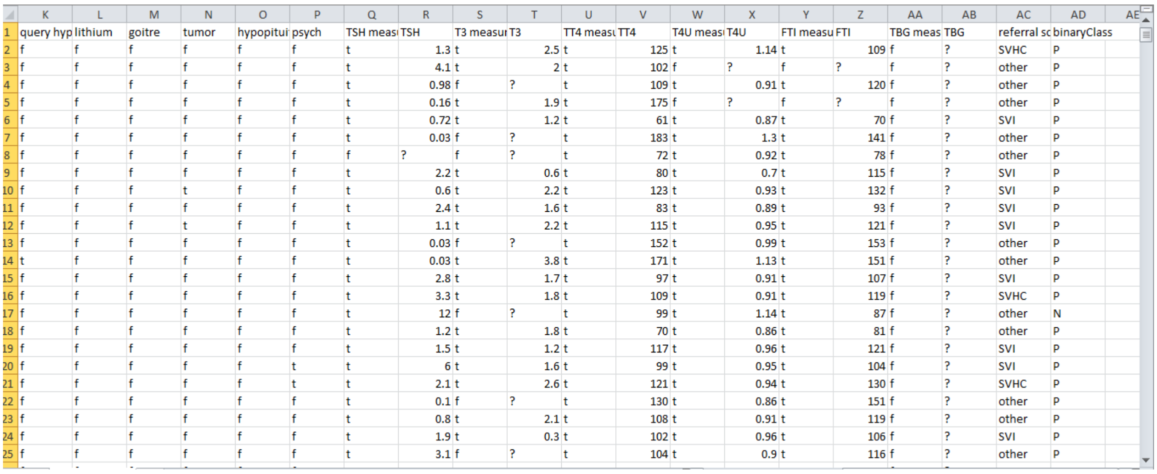


Figure.5 This structure shows the Naïve Bayes algorithm proceeding left with respect to support vector machine on right side for classification.

6. AdaBoost: Boosting is a cooperative technique that energies to produce a robust classifier from the number of low value classifiers. Boosting system used for constructing a model from the training data and making second data then accurate the errors from the first data. Data to be added until predict the seamless model from the trained data. It was established in 1997 by Freund and Schapire . ML algorithm emerging for reducing consequence in boosting inclined with instructor.

III. Description of Data

Dataset is committed from UCI AI storehouse. Database contains concerning sufferer's thyroid records. Every thyroid patient's document is composed of 15 characteristic files beneath. By analysing the above research work it is found that regularly used medical attributes to accomplish experimental effort for the diagnosis of thyroid cancer are given below in below table no.1.Among these attributes almost every researcher has selected attributes to perform work for thyroid cancer diagnosis. Characteristic execution remains Boolean (genuine/bogus) then steady esteemed are addicted beneath. Below figure 6 shows the data set Hypothyroid.csv.



	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	query hyp	lithium	goitre	tumor	hypopitui	psych	TSH meas	TSH	T3 measur	T3	TT4 meas	TT4	T4U meas	T4U	FTI measu	FTI	TBG meas	TBG	referral sc	binaryClass	
2	f	f	f	f	f	f	t	1.3	t	2.5	t	125	t	1.14	t	109	f	?	SVHC	P	
3	f	f	f	f	f	f	t	4.1	t	2	t	102	f	?	f	?	f	?	other	P	
4	f	f	f	f	f	f	t	0.98	f	?	t	109	t	0.91	t	120	f	?	other	P	
5	f	f	f	f	f	f	t	0.16	t	1.9	t	175	f	?	f	?	f	?	other	P	
6	f	f	f	f	f	f	t	0.72	t	1.2	t	61	t	0.87	t	70	f	?	SVI	P	
7	f	f	f	f	f	f	t	0.03	f	?	t	183	t	1.3	t	141	f	?	other	P	
8	f	f	f	f	f	f	f	?	f	?	t	72	t	0.92	t	78	f	?	other	P	
9	f	f	f	f	f	f	t	2.2	t	0.6	t	80	t	0.7	t	115	f	?	SVI	P	
10	f	f	f	t	f	f	t	0.6	t	2.2	t	123	t	0.93	t	132	f	?	SVI	P	
11	f	f	f	f	f	f	t	2.4	t	1.6	t	83	t	0.89	t	93	f	?	SVI	P	
12	f	f	f	t	f	f	t	1.1	t	2.2	t	115	t	0.95	t	121	f	?	SVI	P	
13	f	f	f	f	f	f	t	0.03	f	?	t	152	t	0.99	t	153	f	?	other	P	
14	t	f	f	f	f	f	t	0.03	t	3.8	t	171	t	1.13	t	151	f	?	other	P	
15	f	f	f	f	f	f	t	2.8	t	1.7	t	97	t	0.91	t	107	f	?	SVI	P	
16	f	f	f	f	f	f	t	3.3	t	1.8	t	109	t	0.91	t	119	f	?	SVHC	P	
17	f	f	f	f	f	f	t	12	f	?	t	99	t	1.14	t	87	f	?	other	N	
18	f	f	f	f	f	f	t	1.2	t	1.8	t	70	t	0.86	t	81	f	?	other	P	
19	f	f	f	f	f	f	t	1.5	t	1.2	t	117	t	0.96	t	121	f	?	SVI	P	
20	f	f	f	f	f	t	t	6	t	1.6	t	99	t	0.95	t	104	f	?	SVI	P	
21	f	f	f	f	f	t	t	2.1	t	2.6	t	121	t	0.94	t	130	f	?	SVHC	P	
22	f	f	f	f	f	f	t	0.1	f	?	t	130	t	0.86	t	151	f	?	other	P	
23	f	f	f	f	f	f	t	0.8	t	2.1	t	108	t	0.91	t	119	f	?	other	P	
24	f	f	f	f	f	f	t	1.9	t	0.3	t	102	t	0.96	t	106	f	?	SVI	P	
25	f	f	f	f	f	f	t	3.1	f	?	t	104	t	0.9	t	116	f	?	other	P	

Figure 6. Hyperthroid.csv

Table1: Attribute for the feature selection

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

IV. Results

The data sets for the thyroid cancer have been influenced from the UCI machine learning repository. The work is resided with two different stages. The major phase covers the subset selection that is executed by adjusting mutual information and prediction of the thyroid datasets done using ANN. Specifically in the understanding of diseases neural networks are successfully compulsory in the distinctive fields in the medical land. The inevitability of the examination for the datasets of the thyroid cancer are assigned as the designated presence by every feature range algorithm. Based on the research and outcomes gained we came up with the following answers that were produced in the preliminary part of this paper. We separately recognized the pathological and serological parameters of Thyroid Cancer with the help of hormone specialists. We established the projecting model for Thyroid Cancer diagnosis based upon the extreme parameters, serological parameters distinctly as well as by combining of both.

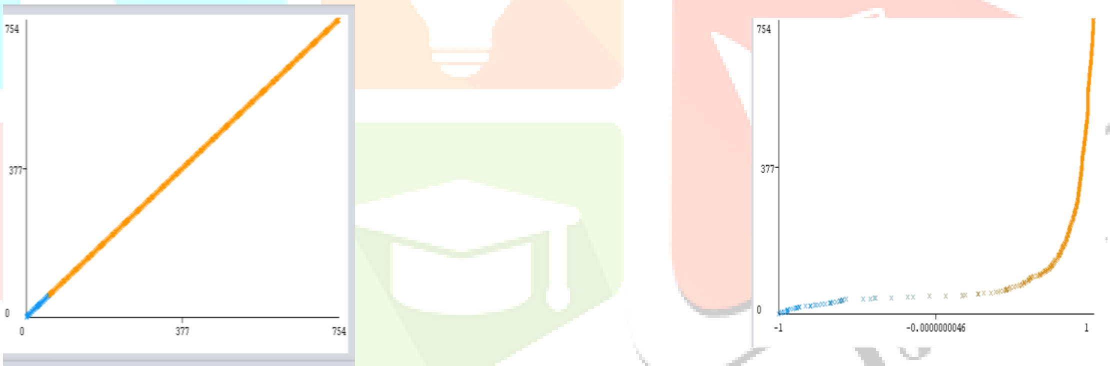


Figure 7 (a)

Figure 7 (b)

Figure 7 (a) and (b) shows the result of support vector machine and logistic regression for thyroid cancer

Table 2. Measurements for classification with all attribute of dataset

NO	Algorithms	Accuracy
1	Logistic Regression	91.73
2	Decision Tree	90.13
3	SVM	92.53
4	Naive Bayes	90.67

V. Conclusion

The Analysis of all the four algorithms ended up with better prediction than human intervention resulting in SVM with best prediction with the data set preceded. Machine Learning classification algorithms are used to identify the thyroid difficulties. Proposed technique assistance to reduce the noisy evidence of a patient. Machine Learning algorithms such as KNN, Naïve bayes, Support vector machine are measured for the study. The results of these classification approaches are constructed on accurateness and concert of the prototypical. The resulting classification of effective data supports finding the dealing to the thyroid patients with recovering charge and level the management. For the given data set the accuracy using SVM is 92.53, logical regression is 91.73 and Decision tree is 90.13 and the data set is greater, the computational cost of SVM will rise.

VI. Future Enhancements

Different datasets with large instances can be applied on various Machine Learning Algorithms, so that the optimization in prediction will be at high. In the future, the predication and detection techniques can be further enhanced by adding more examples to the dataset foremost to more robust results. Furthermore, the model can be upgraded by spreading on various structures collection algorithms to grow the performance of Thyroid Cancer prediction. It would be improved to use real life datasets from dissimilar fields of science to exhaustively test the algorithms and comparison their performances.

REFERENCES

[1]. Lewiński A, Sewerynek E, Karbownik M: 2006. Aging processes and the thyroid gland. In Aging and Age-Related Diseases: The Basics. Edited by: Karasek M. New York: Nova Science Publishers, Inc;131–172.Google Scholar.

[2]. Faggiano A, Del Prete M, Marciello F, 2018. Marotta V, Ramundo V, Colao A: Thyroid diseases in elderly. Minerva Endocrinol 2011, 36: 211–231.PubMedGoogle Scholar International Journal of Computer Sciences and Engineering Vol.6(1), Jan 2018, E-ISSN: 2347-2693 ©, IJCSE All Rights Reserved 331.

[3]. Papaleontiou M, Haymart MR: 2012. Approach to and treatment of thyroid disorders in the Elderly. Med Clin North Am, 96: 297–310. 10.1016/j.mcna.2012.01.013.

[4]. Bahn, R., Burch, H, Cooper, D, et al. 2011. Hyperthyroidism and Other Causes of Thyrotoxicosis: Management Guidelines of the American Thyroid Association and American Association of Clinical Endocrinologists. Endocrine Practice. Vol 17 No. 3.

[5]. Braverman, L, Cooper D. 2012. Werner & Ingbar's the Thyroid, 10th Edition. WLL/Wolters Kluwer.

[6]. Dr. Rishitha Banu et. Al., 016 "Predicting thyroid disease using data mining Technique" International Journal of Modern Trends in Engineering and Research, pg: 666-670.

[7]. Senthilkumar et al., 2015 "Classification of Multi-dimensional Thyroid Dataset Using Data Mining Techniques: Comparison Study" Advances in Natural and Applied Sciences, 9(6), Pages: 24-28.

[8] Rasitha Banu, Baviya . 2015."A study on Thyroid disease using Data Mining Technique", IJTRA Journal.

[9] Banu, et al. 2016. "Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique", Communications on Applied Electronics (CAE) –Volume 4– No12.

[10] Ebru turanoglu-beka R et al., 2016. "Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms", Oxford Journal of Intelligent Decision and Data Science, PP: 13-28.

[11] C. Fan, F. Xiao, Z. Li, J. Wang. 2018. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy Build. 159, 296–308.

[12] W. Kleiminger, C. Beckel, T. Staake, S. Santini. 2013.Occupancy Detection from Electricity Consumption Data. In Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings, Rome, Italy, pp. 1–8.

[13] D. Mora, G. Fajilla, M. Austin, D. Simone. 2019. Occupancy patterns obtained by heuristic approaches: Cluster analysis and logical flowcharts. A case study in a university office. Energy Build. 186, 147– 168

[14] V. Cerqueira, L. Torgo, M. Mozetic. 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. Mach. Learn., 109, 1997–2028.

[15] Dreiseitl, Stephan, and Lucila Ohno-Machado. 2002. "Logistic regression and artificial neural networkclassification models: a methodology review." Journal of biomedical informatics 35.5-6 ,352-359.