ZKG INTERNATIONAL

ISSN: 2366-1313

# Machine learning based Thyroid Disease prediction and feature selection using Principal Component Analysis

Dayana Dasari[1] , M.Tech Scholar, Department of CSE, dayanadasari126@gmail.com

Dr K.Srinivas[2],Professor of CSE, email:srinivas.katikireddy@gmail.com,

Bonam Venkata Chalamayya Institute of Technology & Science, Batlapalem.

Dr S Srikanth[3] ,Professor of EEE, sambana.srikanth@gmail.com,

B V C Engineering College, Odalarevu

**Abstract**: *The Thyroid organ is a vascular organ and one of the fundamental organs of the human body. This organ sweats two chemicals that permit keeping up with the body's digestion. The two sorts of Thyroid sicknesses are Hyperthyroidism and Hypothyroidism. At the point when this condition shows up in the body, they end explicit chemicals, which disparities the body's digestion. A thyroid comparing Blood test is used to get this illness, yet it is as often as possible clouded, and commotion will join in. As a rule, the information cleaning techniques were utilized to make the information rudimentary adequately for the investigation to show the opportunity of patients gaining this sickness. AI plays a fundamentally deciding job in sickness forecast. In this paper, we have attempted to anticipate hypothyroid in the starter stage. To do as such, we have for the most part utilized three component determination strategies and different grouping techniques. Highlight determination strategies utilized by us are Recursive Feature Selection (RFE), Univariate Feature Selection(UFS), and Principal Component Analysis(PCA), alongside grouping calculations named Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression(LR) and Naive Bayes(NB).*

***Keywords***: *Thyroid disease prediction, machine learning algorithms, Feature Selection,Principal Component Analysis, Recursive Feature Selection, Univariate Feature Selection.*

## I. INTRODUCTION

Computational science Evolution is utilized in medical services. It considers the assortment of put away persistent insights for infection expectation. Foreseeing calculations might be accessible to anticipate illness in the beginning phases. Clinical information frameworks are wealthy in informational collections, yet scarcely

ISSN: 2366-1313

any sound frameworks can without much of a stretch survey the condition. After some time, robotized examination calculations started to assume a significant part in taking care of complex and non-straight issues in the improvement adaptation. In any sickness expectation, strategies are utilized to pass includes that can choose from remarkable informational indexes that can involve to order solid people as precisely as could really be expected. On the off chance that it doesn't, misclassification might bring about the sound impacted individual getting superfluous help[1].

As indicated by the flow circumstance, hypothyroidism is one of the main sicknesses of all and can possibly transform into a broad infection for some ladies. Subject matter authorities agree, 50 million individuals in Bangladesh experience the ill effects of thyroid sickness. Among these, ladies are multiple times bound to foster thyroid illness. Albeit a larger part of fifty million individuals have thyroid sickness, almost 30 million know nothing about this condition. The Bangladesh Endocrine Society (BES) perception showed that around 20-30% of young ladies have thyroid illness [2]. The thyroid organ is an organ situated in our edge in the neck. It is butterfly-molded and little. It secretes a few chemicals that blend in with the blood and travel all through the body to control various exercises. The thyroid chemical is answerable for keeping up with digestion and rest, development, sexual capability, and mind-set. Contingent upon the discharge of thyroid chemical we can feel drained or anxious and furthermore mayhave weight reduction. There are two fundamental thyroid hormones:Triiodothyronine (T3) and Thyroxin (T4). These two hormonesare predominantly liable for keeping up with the energy in our bodies.Thyroid Stimulating Hormone(TSH) is delivered by thepituitary organ that assists the thyroid organ with delivering T3 and T4.There are two normal thyroid illnesses 1) Hypothyroid 2)Hyperthyroid [3].

ISSN: 2366-1313

The thyroid organ is an endocrine organ situated in the human throat underneath Adam's apple that aides in the discharge of thyroid chemicals that influence digestion and protein blend. Thyroid chemicals are valuable in computing how quick the heart is thumping and how rapidly it consumes energy. The thyroid organ secretes two dynamic chemicals: levothyroxine (T4) and triiodothyronine (T3). These chemicals assist with directing internal heat level. These are additionally valuable assets for energy assimilation and transmission in any piece of the casing and are fundamental for protein the executives. Iodine is the main part of the thyroid organ. It is reflected in a couple of explicit inquiries. An absence of supply of these chemicals can prompt an overactive thyroid organ. There are many causes related with hyperthyroidism and hypothyroidism. There are various kinds of drug, for example, thyroid medical procedure, that conveys dangers of openness to ionizing radiation, diligent responsiveness of the thyroid organ, iodine inadequacy and absence of chemicals that make thyroid chemicals [4].

## II. PROBLEM STATEMENT

III. As per measurements, thyroid problems are on the ascent in India. Around 1 out of 10 Indian grown-ups experience the ill effects of thyroid issue. It has been assessed that around 42 million people groups experience the ill effects of thyroid sickness. Foreseeing thyroid turmoil by specialist is a drawn-out process which could prompt negative expectation, just experienced specialist can look at the case appropriately. To help specialists AI can help them in analysis of sickness and decreases their weight.

## VI. LITERATURE SURVEY

A few boards have been finished lately to analyze individual thyroid organ sicknesses. Many creators have utilized various kinds of reality tracking down strategies. The creators

ISSN: 2366-1313

have exhibited that they are reaping a decent and genuine strategy for distinguishing thyroid-like illnesses through other datasets and calculations connected with the work to be finished toward the edge of destiny to accomplish higher outcomes. The paper's subject makes sense of the different procedures of factual prospecting components and measurable elements that have been advocated lately for the translation of thyroid illnesses while guaranteeing that various conceivable outcomes and methods are utilized. A few equipment location calculations, Random backwoods, sickness Tree, Naive Bayes, SVM and ANN, can be generally utilized in like manner illnesses and prescient issues. There are a couple of consecutive abilities of sicknesses connected with coronary conduit illness [5], diabetes, Parkinson's infection, hypertension, Ebola infection (EV), conclusion and visualization, evaluation and planning of R-NA, and information can be created utilizing Biomedical imaging. In any case, fostering a framework to logically endlessly distinguish illness forecast systems is a straightforward undertaking. There are indispensable issues, for example wearable register set, gathering and gathering, to train the gadget to acquire information on structures. With respect to the genuine leisure activity issues, gauges for uber standard rates in biomedicine are desirable over profound progression and are basically non-existent.

In [6], an organized way to deal with early finding of thyroid sickness utilizing the backpropagation calculation is utilized in the neurological local area. The ANN is touchy, and when the blunder is spread, it observes that this is utilized for early expectations of illness. The impact of the ANN is shown utilizing checked exploratory realities and components, which can be affirmed as information that is not generally utilized during the preparation interaction. The creators analyzed and concentrated on the four characterization models: Naive Bayes, Decision Tree, Multilayer Perceptron

ISSN: 2366-1313

and Radial Basis Function Network. The completion shows incredible exactness for all class modes. The choice tree form abrogates the contrary sorts. In this work, 29 characteristics were gathered from the informational collection and forced as a component choice strategy, for example, Chi-Square. Datasets are sifted by applying solo line channels to change ascribes from persistent to ostensible qualities, in this way decreasing 29 attributes to ten. The ANN presumes that the essential information are in great arrangement and recommend the better neurological local area than be utilized rather for early illness expectations.

In [7], this work recommended the kind of thyroid illness that is among the main issues in the classification. Two hypothyroidism and hyperthyroidism liable for system digestion have been portrayed with various determination or extraction as an earlier interaction, including successive forward choice, consecutive back determination, and GA (Genetic Algorithm). SVM is utilized as a classifier to isolate thyroid illnesses. This study is principally founded on the memory records of the UCI framework study, and the second is the genuine measurements gathered by the Imam Khomeini Health Foundation with the assistance of the Intelligent System Laboratory of KN Toosi University.

VI. In [8] these pictures, they proposed a variant for characterizing these thyroid information utilizing first-request capability choice and a piece based characterization technique. In this methodology, MKSVM is utilized to separate thyroid pollution with high exactness of 98.65%. The curiosity and plan of this proposed model are utilized as an element choice to work on the exhibition of the characterization methodology utilizing progressed Gray Wolf enhancement.

AI (ML) is a part of man-made brainpower that is progressively entering the components of

ISSN: 2366-1313

clinical examination. Machine authority permits calculations to really take a look at the fun without focusing on something over the top. AI was incited by entering explosion related with cutting edge computational capabilities, and old style the study of disease transmission is a complex logical strategy for catching the gifts of refined information [9]. Not failing to remember the far reaching measurable game plans and cautious gear research in a clinically pertinent association among information and result norms. Authentic examination of careful ends is totally off-base to change careful standards. Basic parts of the careful plans are the depiction of the patient friend who helps with passing judgment on the careful treatment. Machine dominance permits PC frameworks to precisely foresee current data in light of verifiable measurements. The media side offers exceptionally solid prescient calculations that reproduce recently specific verbal trades in wide and complex arrangements of realities and become acclimated to areas of strength for the secrecy environment.

## VII. PROPOSED WORK

In mastering the system, there may be a statement that when entering a spam price, the best way to get rid of spam is to return. Using system study rules to predict something when a data set contains noisy records, which is not critical, effectively hampers the performance of algorithms to achieve the highest accuracy. To get the best accuracy in the algorithm, we need to feed those features which can be uniquely critical, and this is achieved with the help of using the feature selection method. In the first step, we collected hypothyroidism records from the registered diagnostic centre and cleaned up the data. In the second step, we performed feature selection in our data set to detect the important attributes. The method of feature selection is RFE, UFS and PCA. The third step is all based around using the

ISSN: 2366-1313

feature options, which adjust based on the performance of each rule set. We examined our data set based on these class algorithms - Supporting Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB). The framework is demonstrated in Figure 1.

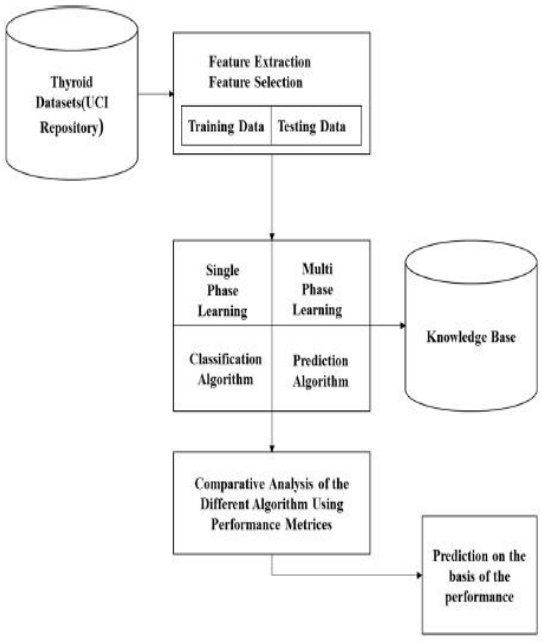


**Figure.1** System architecture

Support Vector Machine:Support vector machine is considered as a varying exploration calculation that aides in playing out the examination in an exact way.Support vector machine is a methodology that is started with an idea of an expert of separatinghyper planeto help in the circulation for testing ofdata. A hyper plane or numerous planes are made bythe support vector machine classifier in highdimensional space. The preparation information tests arebeing isolated as a positive and negative datasamples by the hyper plane.

Choice Tree: Tree-like diagram is utilized in decisiontree classifier. A choice tree is ordered by its 3nodes for example inner hubs, leaf hubs, and the rootnodes. The inner hub suggests as the test on anattribute, the leaf hub indicates as the dissemination ofthe class and the root hub implies as the tree thathas the top most hub. The two mostextensivealgorithms that are utilized in the assemblances of a choice tree for demonstrative andprognostic model of thyroid sicknesses are C4.5 andID3. Specialists use Decision Tree broadly inhealthcare field especially to analyze variousthyroid illnesses

Include Selection Technique

The course of component choice is to naturally choose those elements which are fundamentally essential to help in anticipated the result or factors we are

ISSN: 2366-1313

keen on. There are a few information that lies in our dataset that essentially decline the precision of our model. Furthermore, to dispose of these undesirable information highlight determination strategy assumes a significant part. The advantage of component determination is-

Recursive Feature Elimination (RFE):

Recursive component disposal (RFE) is an element choice procedure that works by form coordinating and killing delicate highlights and positions an individual's capacities utilizing coef_ and feature_importances credits. The significance of abilities is conveyed through an 'in-structure' approach, and less significant capacities are eliminated until they accomplish the ideal usefulness. We ran RFE on a few calculations to test the proper capacities of a specific calculation and found three significant highlights in view of the calculation. The planned exactness of involving RFE for each standard set is SVM (99.35%), Decision Tree (99.35%), Random Forest (99.35%), Logistic Regression (99.35%), and Naive Bayes (94, 23%).

Univariate Feature Selection

UFS is one more element choice strategy that utilizes SelectKBest with a chi-square test (Score_func = chi2) to find the best scoring highlights. Factual test The Chi-Square test (Score_func = chi2) measures the power of the gig relationship separately as indicated by the reaction variable. This approach uncovers three significant highlights from our informational collection. The assessed exactness of involving UFS for each gathering is SVM (98.71%), Decision Tree (99.35%), Random Forest (99.35%), Logistic Regression (99.35%), and Naive Bayes (96 .79%).

Head Component Analysis (PCA)

PCA is an information decrease strategy, a vital property determination that changes over high layered information into low layered information to pick the main property that can catch the full data about informational index. The interpreted_relative_related characteristic orders significant abilities

ISSN: 2366-1313

with not failing to remember the element that causes the most elevated change in a PCA call as the principal prevailing part, the quality that causes the second difference as the second primary issue, etc. The normal exactness of involving PCA for each standard set is SVM (89.74%), choice tree (87.17%), irregular timberland (88.46%), calculated relapse (89.74%), and Naive Bayes (89.74%).

## XIV. RESULTS AND DISCUSSIONS

We see that the RFE capability determination strategy assists us accomplish higher exactness with any remaining classifiers. Our outcomes show that RFE permits us to foresee stage 1 hypothyroidism utilizing a constant dataset. Gathering information in this epidemiological situation is truly challenging.

As an eventual outcome, we've gathered the 519 best realities. So on the off chance that we contemplate what is happening and the limitations, we will not have the option to take a gander at a bigger informational index. We have a situation of realities about the works of art. In our review, we found that no work has been finished on the thyroid organ situated in Bangladesh up until this point. So later on, we want to work with a bigger informational index. Ideally, individuals bigger than our nation will show side interest pictures about this sickness to assist us with finding an improved solution and foresee intra-stage illnesses with higher exactness. We trust with the goal to assist individuals of the United States of America with driving a sound society.

**Table.1 Result examination for different calculations**

| Serial Number | Algorithm | Feature Selection Technique (RFE) Accuracy% | Feature Selection Technique (UFS) Accuracy% | Feature Selection Technique (PCA) Accuracy% |
|---|---|---|---|---|
| 1 | SVM | 99.35% | 98.71% | 89.74% |
| 2 | Decision Tree | 99.35% | 99.35% | 87.17% |
| 3 | Random Forest | 99.35% | 99.35% | 88.46% |
| 4 | Logistic Regression | 99.35% | 99.35% | 89.74% |
| 5 | Naïve Bayes | 94.23% | 96.79% | 89.74% |

ISSN: 2366-1313

## VIII. CONCLUSION

Gathering data about the ongoing pandemic situation might challenge. In this paper, we can say that choosing the RFE highlight assists us get better precision with every one of the various classifiers. Our outcomes exhibited that RFE makes it more agreeable to foresee stage 1 hypothyroidism utilizing an ongoing dataset. As an end-product, we gathered the 519 information. Thusly, because of the situation and restrictions, we were unable to take a gander at a bigger informational index. Our examination found that Bangladesh-based thyroid pictures had not been gotten previously. We have a genuine predicament to work with. So later on, we should work with a bigger informational collection, and ideally, there will be extra individuals from our accomplices. S. would be keen on boards about this infection to assist us with tracking down an improved response and foresee the first-request sickness with better precision. We desire to assist with peopling in our country to keep a sound society.

## REFERENCES

[1] 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.

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

[3] L. Clinical epidemiology in the era of big data: newopportunities, familiar challenges. Clin Epidemiol. 2017; 9:245-250.

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

[5] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "ImprovingHeart Disease Prediction Using Feature Selection Approaches", 16thInternational Bhurban Conference on Applied Sciences & Technology(IBCAST), pp. 619-623, 18 March, 2019.

[6] P. Duggal, and S. Shukla, "Prediction Of Thyroid Disorders UsingAdvanced Machine Learning Techniques", 10th International Conferenceon Cloud Computing, Data Science & Engineering (Confluence), pp.670-675, 09 April 2020.

[7] Dhaka Tribune(2018), 50 million people suffer from thyroid disease inBangladesh. Available: https://www.dhakatribune.com/feature/healthwellness/2018/05/25/experts-50-million-people-suffer-from-thyroiddisease-in-bangladesh

[8] A. K. Aswathi, and A. Antony, "An Intelligent System for ThyroidDisease Classification and Diagnosis", 2nd International Conference onInventive Communication and Computational

Technologies (ICICCT2018), pp. 1261-1264, 27 September, 2018.

[9] A. Begum, and A. Parkavi, "Prediction of thyroid Disease Using DataMining Techniques", 5th International Conference on AdvancedComputing &Communication Systems (ICACCS), pp. 342-345, 06 June,2019.

[10] Prasadu Peddi (2018), "A STUDY FOR BIG DATA USING DISSEMINATED FUZZY DECISION TREES", ISSN: 2366-1313, Vol 3, issue 2, pp:46-57.

[11] Prasadu Peddi (2019), Data Pull out and facts unearthing in biological Databases, International Journal of Techno-Engineering, Vol. 11, issue 1, pp: 25-32.

[12] A. Tyagi, R. Mehra, and A. Saxena, "Interactive Thyroid DiseasePrediction System Using Machine Learning Technique", 5th IEEEInternational Conference on Parallel, Distributed and GridComputing(PDGC-2018), pp 689-693, 27 June, 2019