Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Thyroid Prediction System using Machine Learning Techniques

Haria Viral *[1] More Suraksha[2] Patel Bijal[3] and Patil Harshali[4]

[1*]Dept. of Computer, Thakur College of Engineering & Technology, Mumbai, India,
Email: viralharia97@gmail.com, +91-9004421068.
[2]Dept. of Computer, Thakur College of Engineering & Technology, Mumbai, India,
Email: more.suraksha@gmail.com, +91-9969302221.
[3]Dept. of Computer, Thakur College of Engineering & Technology, Mumbai, India,
Email: bijal1236@gmail.com, +91-7738222001.
[4]Dept. of Computer, Thakur College of Engineering & Technology, Mumbai, India,
Email: harshali.patil@gmail.com, +91-9819076622.

ABSTRACT

Disease diagnosis is not an easy task, especially without proper equipment. This problem can be solved by using some machine learning techniques. Until today there have been various researches on heart diseases, or such diseases and have been predicted appropriately using such techniques. In this project we try to implement prediction of thyroid disease system as not much work has been done on thyroid. This project builds a system which helps predict a normal person about his possibility of getting thyroid in future. This system will also help doctors to better diagnose their patients and provide proper treatment in time. The algorithm is first trained using the data set available from UCI repository and then tested on the data set. Then the user enters his details and the algorithm starts running, according to the values entered by the user the algorithm predicts that the user will be having thyroid in future or not. This system will help doctors as well as individuals to have a possible disease diagnosed. And once a person predicts whether or not he can be diagnosed with thyroid disease in the future, our system will be giving suggestions like blogs of experts, doctors etc. recommending home remedies, homeopathic and ayurvedic medicine suggesting sites etc.

KEYWORDS: Machine Learning, Accuracy, Classification Models, Thyroid, Prediction System, LDA Algorithm, PCA Algorithm, Disease

***Corresponding author**

Viral Haria

Department of Computer,

Thakur College of Engineering & Technology,

Kandivali-400101, Mumbai, INDIA.

Email:viralharia97@gmail.com, Mob No - +91-9004421068

INTRODUCTION

Diagnosis of diseases is a very complex and difficult task, requiring a lot of experience and knowledge. Usually people physically go to doctors' clinics for regular checkups or for diagnosis of some symptoms of any disease. The thyroid gland, situated near the base of the throat, secrets a hormone thyroxin which controls many of the metabolic processes like blood pressure, body weight and temperature. The thyroid disease can be caused due to an under-secretion or an over-secretion of thyroxin. Thyroid disease is very common illness among people living in hilly regions. The symptoms of this disease are usually not easily detected because it varies depending on the type and hence proper medicines cannot be provided on time. Diagnosis can often be done with laboratory tests. There are numerous tests for different types of this disease and we aim to accurately predict each type. For this, we have made provision for the user to enter each result of almost any tests for our system to predict the likelihood of occurrence of the disease. Data mining and machine learning is the way of semi-automatically classifying and analyzing large datasets to find what amount of the data can be grouped into a single category and then using those categories to accurately predict from the new, growing datasets.

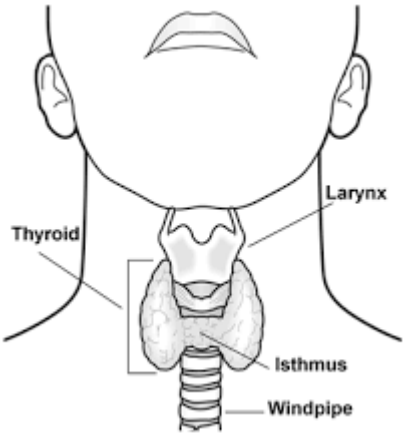

Figure1. Thyroid Gland

Many theoretical works have been proposed for the thyroid disease with different success rates. The importance of using these techniques of machine learning and ANNs is to improve the performance accuracy. There is a bulk of historical data contained in hospital databases which can be used to predict the occurrence of diseases but currently there are no systems in place to make predictions. The current systems can only make statistical calculations and can only trace the database word by word. The depth of knowledge and experience hidden cannot be explored without the use of machine learning and ANNs.

PROBLEM STATEMENT

Disease diagnosis involves analyzing symptoms and detecting whether a disease persists in a body, but analyzing symptoms itself is a complex task. Providing disease diagnosis at early stages with higher accuracy is an important task.[1] Data mining plays an important role in medical field for disease diagnosis. Thyroid disease is very common disease in human. Nowadays most of the people suffering from thyroid disease are women as compared men. These diseases have many side effects such as gain or loss of weight, stress and so on to our human body. If this disease is detected in earlier stage, then doctors can give proper treatment to the patients. Collecting all the past data, analyzing it with the help of two algorithms and compare the end results.

ALGORITHMS

A huge dataset composed of many instances, consisting of data from reports of hypo and hyper thyroid patients. Of this dataset, 80 percent of the data points will be used for the training of the dataset and 20 percent of the data points will be used for the testing of the dataset. The accuracy and the running time of these models was evaluated. Also, finally, the effect of component analysis on the performance of these models will be evaluated.

Predefined Algorithms

LDA Algorithm: LDA is basically a classification algorithm which is based upon the concept of finding a linear combination of two or more variables that can most accurately distinguish between two classes. The LDA algorithm makes prediction by estimating the probability of a new set of input belong to which category or class. To do so, the probability of the new set of input belonging to each class is calculated and the class with maximum probability is predicted to be the output.

$$S = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{within}}} = \frac{(\vec{w}\cdot\vec{\mu}_1 - \vec{w}\cdot\vec{\mu}_0)^2}{\vec{w}^T\Sigma_1\vec{w} + \vec{w}^T\Sigma_0\vec{w}} = \frac{(\vec{w}\cdot(\vec{\mu}_1 - \vec{\mu}_0))^2}{\vec{w}^T(\Sigma_0 + \Sigma_1)\vec{w}} \quad (1)^{16}$$

This is Fisher's linear discriminate formula.

LDA Algorithm overcomes the limitations of classifications like logistic regression. LDA addresses problems like two-class problems, instability of well separated classes and unstable with other examples. This algorithm also makes certain assumptions like the data is Gaussian and each variable when plotted will generate a bell like curve. Another assumption is that, each attribute of the algorithm has the same variance and the mean and average are such that each variable varies same amount among them.

PCA Algorithm: PCA, an algorithm mostly used for making predictive models, is used when there are many variables and we need to find correlation between them. It is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.[2] Both the algorithms - LDA and PCA perform almost the same thing but LDA algorithm focuses on maximizing the reparability among classes. PCA basically focuses on reducing the dimensionality of the data set and retaining the dataset variations to the maximum extent. It helps us obtain a lower-dimensional picture which is a projection of the object in a most informative point of view.

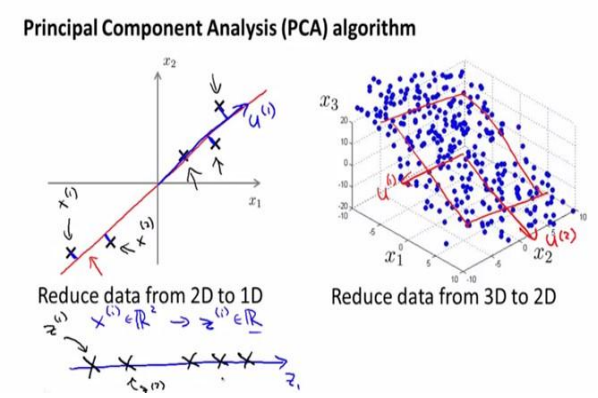

Figure2. Principal Component Analysis[11]

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \quad \text{where} \quad \bar{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i \qquad (2)[11]$$

The output of PCA algorithm is in the form of principal components (PC), which are either equal to the number of the input variables or are less than the input variables when we want to reduce the dimension of our dataset.

The properties that principal component possess include - the principal components are orthogonal in nature and the ordering of PCs is such that the 1st PC has the most variations and the last PC has the least. PCA algorithm is used for image compression.

Classification Analysis

The Logistic regression is a linear classification model whereas the KNN, SVM, Naïve Bayes, Decision Tree are nonlinear. Nonlinear classifiers are expected to display better results since

the various activities exhibit non-linearity while remarking the data. The hyper parameter for every classifier ought to be set to an esteem that expands the precision of the model over the test data. The Principal Component Analysis is a statistical procedure used to convert the observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Based on these results equivalent predictions is made by the Machine Learning model.

Logistic Regression

The logistic regression would probably be the one of the most poorly showing model for the accuracy. This is because it a linear classifier and thus cannot easily adapt to drastic changes.

K-Nearest Neighbor

This classifier chooses the k nearest neighbors and predicts the newer data point using the Euclidean distance. The newer data point is assigned to the category with the most neighbors. [If we have 2 classes then we need to pick any odd value of k. K should not be a multiple of the number of classes.]

SVM (Support Vector Machine)

An SVM model is a representation of the instances as points in space, plotted so that the instances of the distinct categories are divided by a clear slit that is as wide as possible. Newer data points are then mapped into the same space and prediction is made whether to which category the data points belongs to, based on which side of the gap they fall.

Naïve Bayes

Naive Bayes is a family of probabilistic algorithms that take the use of probability theory and Bayes' Theorem to predict the category of a given subject. They are probabilistic, i.e. that they compute the probability of each category for a given subject, and then output the category with the highest one. It predicts the subject's datasets with the help of the Bayes theorem that uses the predefined values of the two sets. For example, it feeds whether an employee drives to work or walks to work and predict if a new employee is added whether it will walk or drive to work.

$$P(A/B) = (P(B/A)\,P(A))/P(B) \qquad (3)$$

Decision Tree

Decision trees are a kind of Supervised Machine Learning where the information data is ceaselessly fragmented according to a specific parameter. The decision tree and be clarified by two entities, decision nodes and decision leaves. The leaves can be stated as the final outcome and the nodes as where the data is fragmented.

PROPOSED METHODOLOGY

The algorithm is first trained with the help of UCI repository values. The algorithm learns that what inputs could give a positive output and what inputs would result in a negative output. The web portal will allow the user to enter their details such as age, gender, thyroxine details, antithyroid medication details, thyroid surgery, pregnancy, sickness, hyperthyroid query, hypothyroid query, tumor, psych, tsh value measured, t3 value measured, tt4 value measured, t4u value measured, fti value measure and tbg value measured. After the user enters all these details, the algorithm runs and predict the result.

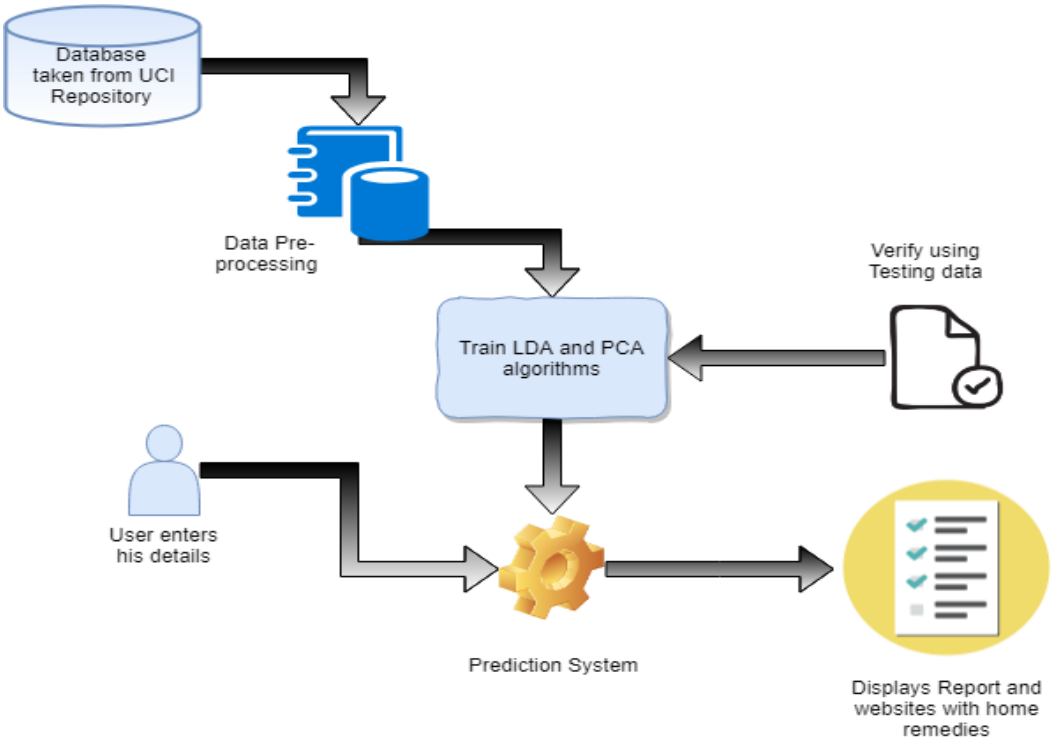

Figure3. Proposed System Architectur

CONCLUSION AND FUTURE SCOPE

Thyroid disease is one of the major diseases and prediction of it is very difficult task to complete without using a computer technology. Disease diagnosis plays a vital role and it is necessary for any busy clinician. Also, prediction of Thyroid disease at early stage will help the doctors to provide medication at right time.

Currently there is no such system to predict Thyroid, people have to visit doctors personally. This system will help users predict thyroid disease at early stage.

In this paper, we have discussed two machine earning algorithms: PCA and LDA and we will be comparing them using the classification models like Logistic regression, K-nearest neighbor, SVM, Naive Bayes, Decision tree and Random forest. The best algorithm will be used in the prediction system for maximum accuracy.

These algorithms can be further applied to other diseases also to help predict it.

REFERENCES

1. Prerana, Parveen S, Khushboo T, Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network, International Journal of Research in Management, Science & Technology (E-ISSN: 2321-3264); 2015; 3(2): 75-79.
2. "Principal Component Analysis" [online], 2015 [2001 January 13], Available from: URL: https://en.wikipedia.org/wiki/Principal_component_analysis
3. "Principal Component Analysis Tutorial" [online], 2018 [2011 August 23], Available from: URL: https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial
4. "Linear Discriminant Analysis" [online], 2015 [2001 January 13], Available from: URL: https://en.wikipedia.org/wiki/Linear_discriminant_analysis
5. Shaik R, M. R. Narasinga R, Machine Learning Techniques for Thyroid Disease Diagnosis - A Review, Indian Journal of Science and Technology;2016; 9(28): 1-9.
6. G. Rasitha B, Predicting Thyroid Disease using Linear Discriminate Analysis (LDA) Data Mining Technique, Communications on Applied Electronics (CAE) (ISSN:23944714), 2016; 4(12): 4-6.
7. "Glossary of common Machine Learning, Statistics and Data Science terms" [online], 2018 [2013 April 11], Available from: URL: https://www.analyticsvidhya.com/glossary-of-common-statistics-and-machine-learning-terms/
8. "Endocrine System" [online], 2017 [2004 July 15], Available from: URL: https://quizlet.com/94210487/endocrine-system-flash-cards/
9. Dr. G. Rasitha, M.Baviya, Predicting Thyroid Disease Using Data mining Technique, International Journal of Modern Trends in Engineering and Research (IJMTER); 2015; 2(3): 666-670.
10. S. Umadevi, Dr. K. S. Jeen M, Applying Classification Algorithms to Predict Thyroid Disease, International Journal of Engineering Science and Computing;2017; 7(10): 15118-15120.

11. Anupama S, Prabhdeep K, Diagnosis of thyroid disorders using artificial neural networks, IEEE International Advance computing Conference (IACC 2009) – Patiala, India; 2009: 1016-1020.
12. Kousarrizi, Nazari MR, Seiti F, Teshnehlab M., An experimental comparative study on thyroid disease diagnosis based on feature subset selection and classification, International Journal of Electrical & Computer Sciences; 2012; 12(1):13–19.
13. "UCI Machine Learning Repository of machine learning database" [online], 2018 [1985 September 30],Available from: URL: http://www.ics.uci.edu/.
14. F. S. Gharehchopogh, M. Molany and F. D. Mokri, Using Artificial Neural Network In Diagnosis Of Thyroid Disease: A Case Study, International Journal on Computational Sciences & Applications (IJCSA);2013;3(4): 49-61.
15. "Hyperthyroidism Symptoms, Causes, Treatments, and Diet" [online],2017 [1995 October 18], Available from: URL: https://www.medicinenet.com/hyperthyroidism/article.htm
16. "Economic Feasibility Study" [online], 2018,Available from: URL: https://ofm.wa.gov/sites/default/files/public/legacy/policy/40.40.htm
17. "Hypothyroidism" [online],2015 [2001 January 13],Available from: URL: https://en.wikipedia.org/wiki/Hypothyroidism.