International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Thyroid Disease Detection Using Machine Learning Classification Algorithms

M. Bangaru Lakshmi[1], V Nikhil Sai Karthik [2], Shreya Saride [3] , G. Rakesh Reddy[4], Kartik Khajuria[5]

[1,2,3,4,5]Department Of Computer Science and Engineering, GST, GITAM University, Visakhapatnam, AP, INDIA.

Abstract

Thyroid disease, also known as one of the most common diseases among the humans. It is one of the essential organs in our body responsible for controlling metabolism. In this project, we predict the Thyroid disease into four classes. The four classes namely being "Primary Hypothyroid", "secondary Hypothyroid"," Compensated Hypothyroid" and "Negative". We also aim at bringing out the best accuracy possible using the Classification algorithms such as KNN and Random Forest. The Dataset has been taken from Machine learning UCI repository. The primary aim of this project is to predict the risk of Hypo-thyroid on a person.

***Key Words*:** Thyroid, Hypothyroid, KNN, Random Forest, Primary, Secondary and Compensated Hypothyroid, Classification

1.INTRODUCTION

Thyroid disease is a common cause of medical diagnosis. The disease's prediction and forecasting is difficult in the medical industrial research. One of the most important organs in our body is the thyroid gland. Releases of thyroid hormone are in charge of controlling metabolism. There are generally two types of thyroid caused by the thyroid gland. Hyperthyroidism occurs when there is excessive release of thyroid hormone. The next one being the Hypothyroidism, which occurs where there is not much of a release of the thyroid hormone. These might have different symptoms but there is a good possibility of the overlaps among these two.

A. FTI Test:

It is the T4 value upon Thyroid binding capacity. it is a normalized distribution which is usually same in healthy individuals. Higher FTI value indicates Hyperthyroidism and a lower FTI value indicates Hypothyroidism.

B. TSH Test:

It is a blood test used to measure this hormone. If the TSH levels are too high or too low then it may cause thyroid. Thyroid generally makes hormones which usually are the brains of how we use our energy

We will use a dataset which has the patients data having respective hypothyroid classes to train our model. The dataset contains 30 attributes including one labelled set and about 3772 rows.

2. LITERATURE REVIEW

The Thyroid disease detection has been performed before in many of the projects and everyone have produced results which are better than the previous ones. There has been improvement in using the algorithms that are being used and the more relevance in the data for predicting the disease.

In the paper titled "Predictive analysis for Thyroid disease diagnosis using machine learning", they have built a predictive model which has the categories "Hyperthyroid", "Hypothyroid" and "Normal". They used the Decision tree, Naïve bayes and KNN in which the Decision tree has outperformed the other two algorithms with an accuracy rate of 99.7%.

In the paper "Detecting six different types of thyroid diseases using deep learning approaches" they have found that Men, Women and children are the one's that are developing the disease frequently. They have also found that women over 30 are the ones who are being prone to this disease more often than others. They tried to detect the disease at an early stage so it is easier to be treatable then at a later point of time. They have used KNN, Decision Tree and multilayer perceptrons in order to using deep learning method as well. They concluded that the MLP has performed better with an accuracy value of 95.73.

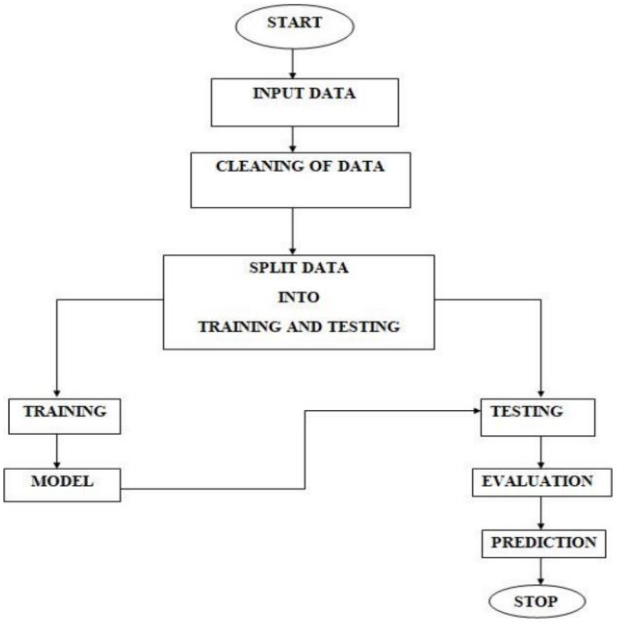
"A machine learning approach to predict thyroid at early stages of diagnosis" has also build models to detect the thyroid at an early stage so it'll be easier to cure them. They have used predictive modelling classification, Binary classification in which they used decision trees and Naïve bayes. The decision tree has been used to predict whether the patient has Thyroid and if yes, Then naïve bayes classifier has been used to detect the stage of the thyroid in the patient.

The paper "Thyroid prediction using machine learning algorithms" have used SVM and random forest for their predictions. They found out that SVM and logistic Regression have outperformed random forest with an accuracy score of 94.13. They have found out that number of female patients are more than number of male patients. There are 84% of female patients and 16% of the male patients in the dataset they have used.

3. METHODOLOGY

In this research we classify the prediction into four classes namely "Primary Hypothyroid", "secondary Hypothyroid", "Compensated Hypothyroid" and "Negative". For this project, the TDD dataset has been taken from the UCI repository. Classification techniques like Random forest and k- nearest neighbors (KNN) are used in the project. The field of this project is Machine Learning algorithms for an even better training and testing of the data to produce much efficient results. In the project at first the Thyroid disease dataset is collected and uploaded. Later we read the data from the dataset. Then we perform Data Visualization i.e., we first analyze the number of disease and non-disease data. Data Visualization is a representation of the information in the form of pie charts, bar graphs, diagrams etc. Then we pre-process the data i.e., we check for any null values or any duplicate values that exist. Data pre-processing is a technique which is used to transform the data into a useful and an efficient format. After that we split the data for Training and Testing and build different classification models like KNN, random forest and select the model with best accuracy. Then Finally, we predict the output.

The workflow looks like this.



We also performed an operation to divide the data into smaller clusters and then performing the training and predictions on them and then putting them together to gain a better accuracy than performing the training and predictions on the whole data at once.

We used the following machine learning classification algorithms.

K- Nearest Neighbours (KNN):
KNN algorithm is one of the machine learning algorithms which comes into the category of Supervised Machine learning. It stores the data from the dataset and when a new data point comes in , it classifies the new data point into the one of the classes based on the most similar features lying in them. It is also known as a lazy learner algorithm as it does not learn at the time of training phase. Instead when a new data point arrives, it classifies them into one of the classes with most similar points or attributes.

Random forest:
It a supervised learning algorithm which uses the ensemble technique. It is used to solve both classification and regression problems. The random forest uses a lots of decision trees on subsets of the data to provide the better accuracy. It takes the average of all the outputs from all the decision trees and then based on the majority of the vote, the algorithm predicts the final output.

4. CONCLUSION

Model	Accuracy
Random forest	94.5%
KNN	92.7%

According to the data above, the results show that the Random forest has outperformed KNN algorithm. The F1, precision and recall were higher in Random forest than that for KNN.

REFERENCES

1. Thyroid Prediction Using Machine Learning Algorithms Sumit Mhaiskar Bharati Vidyapeeth's Institute of Management and Information Technology, Sector-8, CBD Belapur, Navi Mumbai, Maharashtra-400614. International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

2. Chandan & Vasan, Chetan & MS, Chethan & S, Devikarani. (2021). THYROIDDETECTION USING MACHINE LEARNING. International Journal of EngineeringApplied Sciences and Technology. 5. 10.33564/IJEAST.2021.v05i09.028. 2021, Vol. 5, Issue 9, ISSN No. 2455-2143, Pages 173-177

3. Tahir Alyas, Muhammad Hamid, Khalid Alissa, Tauqeer Faiz, Nadia Tabassum, Aqeel Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine LearningApproach", BioMed Research International, vol. 2022, Article ID 9809932, 10pages, 2022.

4. https://www.mayoclinic.org/diseases-conditions/goiter/symptoms- causes/syc-20351829