International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 12, Issue 5 [May. 2023] PP: 479-482

Thyroid Disease Prediction Using Machine Learning

[1]Ms.Sowmiya.P S,[2]Dr.Savitha.J,[3]Ms.Soundarya.S

[1]II Msc Computer Science, Dr.N.G.P Arts and Science College,
Coimbatore-48,Tamil Nadu, India.
[2]Professor,Department of Information Technology
Dr.N.G.P Arts and Science College,
Coimbatore-48,Tamil Nadu, India.
[3]II Msc Computer Science, Dr.N.G.P Arts and Science College,
Coimbatore-48,Tamil Nadu, India.

ABSTRACT:
The goal of the paper is to create a machine learning model that can forecast thyroid disease more accurately. We employ supervised machine learning techniques to forecast thyroid disease and assess their performance in terms of accuracy. These techniques include Random Forest, Support Vector Machine (SVM), and Logistic Regression. We are using the UCI repository's dataset to make predictions about thyroid illness. For the purpose of determining whether the subject has thyroid disease or not, data from the UCI repository have been merged and cleaned.

KEYWORDS: *Logistic regression, Random Forest, SVM, and Thyroid Disease*

Date of Submission: 16-05-2023 Date of acceptance: 30-05-2023

I. INTRODUCTION:

One in ten people suffer from thyroid illness, which is a well-known condition that has an impact on human health. According to statistics, the number of individuals with hypothyroidism and thyroid disease in India is increasing. It is a condition where the thyroid gland fails to generate enough thyroid hormones to satisfy the body's requirements. Women with thyroid illness are frequently diagnosed between the ages of 18 and 35, when they are at their most vulnerable.

A tiny organ in the throat called the thyroid gland produces thyroid hormone.

Its contours resemble those of a butterfly, which has two broad wings that extend outward from the side of its throat and a tiny central portion.[8] Our bodies contain numerous glands that are in charge of producing and releasing the chemicals necessary for the body to carry out a variety of essential tasks.

Our complete body may suffer if the thyroid gland isn't functioning properly. Hyperthyroidism is a condition that develops when the body produces excessive thyroid hormone, and hypothyroidism is a condition that develops when the body produces very little thyroid hormone.[10] The thyroid glands generate T3 (triiodothyronine) and T4 as well as other hormones. (Thyroxine).

Together, T3 and T4 have a significant impact on almost all bodily cells. T4 is produced by the thyroid gland in greater amounts than T3, but when it enters the body's cells and tissues, it is transformed to T3.[13] The T4 hormone is therefore the most crucial hormone to check when checking for thyroid issues.[6] When it regulates the body's metabolism, temperature, and digestive system, the T3 hormone, which has three iodine molecules in its composition, is more metabolically active. Iodine is regarded as the thyroid gland's primary structural component. The pituitary gland is in charge of the thyroid gland.

Thyroid Stimulating Hormone, also known as TSH, is produced by the pituitary gland when the level of T3 and T4 hormones falls too low. TSH stimulates the thyroid glands to create more hormones. Research in the field of healthcare is growing as a result of the growth in data and the advancement of technology. The use of machine learning is necessary because it can be challenging to manage large quantities of patient data. Machine learning is a method that aids in early prediction and results in accurate disease detection.

By learning from training data, different machine learning algorithms help to uncover hidden patterns, train and create models, and make predictions. The machine is trained using well-labeled training data using a variety of supervised machine learning algorithms, and the machine then forecasts the output based on the data.[9]

II. LITERATURE SURVEY:

The healthcare industry has seen a lot of effort using various machine learning algorithms. To predict the disease, many individuals have employed a wide range of data mining techniques. Users who use machine learning may also benefit from personalised care thanks to its forecast analysis. The burden on physicians can be lessened by machine learning by assisting with disease diagnosis.

"Using Machine Learning for Thyroid Detection" Published Online January 2021 in IJEAST, in that they used a variety of machine learning algorithms to predict the likelihood that an individual will have thyroid disease, including SVM (Support Vector Machine), decision trees, logistic regression, KNN (K-Nearest Neighbours), and ANN (Artificial Neural Network). They have also developed a web application to collect information from users and forecast the disease's type.

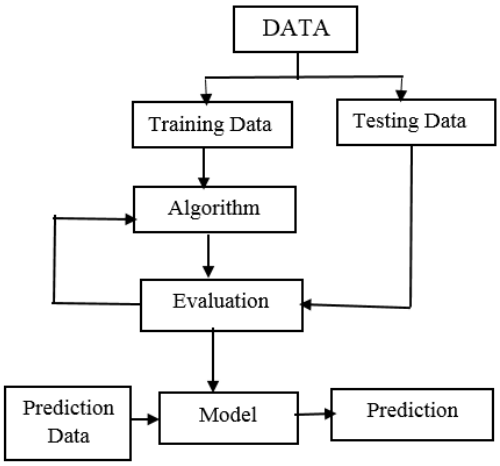
The dataset, which includes the labels hyper and hypo, was obtained from the UCI repository. The accuracy score obtained after training and constructing the model was 93.84% for the KNN algorithm, 95.38% for the SVM algorithm, 75.38% for the ANN algorithm, 92.3% for the decision tree algorithm, and 96.92% for the logistic regression algorithm. They chose the logistic regression algorithm as their prediction model because it allowed them to obtain the highest score.

They used machine learning algorithms like SVM (Support Vector Machine), KNN (K-Nearest Neighbours), and Decision Trees in their study, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," to predict thyroid disease using data from the UCI machine learning repository. They used data from the UCI repository, and using the KNN, SVM, ANN, and decision tree algorithms, they were able to obtain accuracy scores of 98.62%, 99.63%, 97.50%, and 75.76%, respectively. Using the SVM algorithm, they have obtained the highest result.

The dataset from the UC Irvin knowledge finding in databases archive was used in " Thyroid Disease Prediction Using Machine Learning Approaches" released in the National Academy of Science, India 2020. They have employed algorithms like KNN (K-Nearest Neighbours), logistic regression, and decision trees. They applied classification to the data using all of these methods.

With the KNN algorithm, the accuracy number was 96.875%; with the decision tree algorithm, 87.5%; and with the logistic regression classifier, 81.25%. Using the KNN algorithm, they were able to obtain the highest result. The primary goal of this research is to create a system that can determine whether or not a person has a thyroid condition. additionally to disease prediction using fewer factors. Moreover, to offer a practical answer to the healthcare issue. To forecast the disease, we will use a variety of machine learning algorithms, including KNN, Random Forest, and others.

III. RESEARCH METHODOLOGY



In order to forecast thyroid cancer For the purpose of analysing and predicting the illness, a thyroid dataset is needed. We will use a variety of supervised machine learning methods to analyse the dataset. The algorithm with the highest accuracy number will be selected to fetch the result based on the accuracy of various models.

The UCI repository is where the information is located. The dataset needs to be verified for empty or null values as well as extraneous values. The data is then made clean by having these types of numbers removed from it. The parameters that are necessary for the disease prediction are the only ones retained after the data has been cleaned; all other parameters are removed.

The cleaned data is then used as training and testing data, providing the programmes with input. In order to classify the data according to labels, the algorithms take features from datasets. The test data is then fed to the algorithm along with its accuracy score in order to determine whether the forecast is accurate or not, and to compare the accuracy of various models.

A. Attributes used for diagnosis of the thyroid disease:

The following is a list of the characteristics that are crucial for predicting the diagnosis. The following attributes have been used in nearly all of the research studies.

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. Analysis of the proposed algorithm's performance:

Random Forest: Both classification and regression issues can be resolved using this machine learning approach. It is based on ensemble learning, which aids in integrating various classifiers to tackle a challenging problem and enhance model performance. To increase the dataset's predicting accuracy, a number of decision trees are applied to different subsets of the data and an average is taken.

The program uses the forecast from each decision tree and predicts the outcome based on the majority of votes. Using several independent variables, the categorical dependent variable is predicted using the logistic regression algorithm. The classification issue is resolved with this approach. With this approach, we can anticipate values like 0 or 1 by fitting a "S"-shaped logistic function. The logistic regression's value ranges from 0 to 1. As a result, it creates a "S"-shaped curve known as the sigmoid or logistic function.

SVM: This algorithm's goal is to provide the best decision boundary—also known as a hyperplane—that can divide an n-dimensional space into classes so that additional data points can be quickly assigned to the appropriate category in the future. The SVM selects extreme points or vectors to build the hyperplane. Support vectors are data points or vectors that are close to hyperplanes and have an impact on where the hyperplane is located.

C. Dataset:

- The dataset was collected from the UCI repository. The final dataset was created by combining All hypo data with All hyper data. The dataset comprises 30 columns and 7544 rows.
- Age, sex, TSH, T3, TT4, T4U, FTI, and TBG are among the characteristics that are used.
- TSH is a thyroid stimulating hormone, T3 is a triiodothyronine hormone that influences nearly all psychological processes in the body, TT4 is a thyroxine hormone that assesses thyroid function and diagnoses thyroid disease, and FTI is a thyroxine index that is stable over time.

Model	**Accuracy**
Logistic Regression	94.13
Support Vector Machine	94.13
Random Forest	93.85

The table shows that the accuracy of the logistic regression and the support vector machine is 94.13%. In comparison to Random Forest, SVM and Logistic Regression have the highest accuracy. We also discovered that our model has the highest accuracy when comparing the Logistic Regression results to earlier work, and we are also applying several Machine Learning Models.

IV. CONCLUSION:

Consistent in people in good health. Higher FTI is a result of hyperthyroidism.

- The dataset contains numerical values for each thyroid hormone.
- The dependent variable, or predict class, which has labels like "hyperthyroid" or "negative," etc., is at the end.

• In addition to that, the dataset also includes some categorical values, such as sick, pregnant, I131 treatment (Iodine 131 treatment), thyroid surgery (to determine whether someone has undergone thyroid surgery), etc.

V. RESULT AND ANALYSIS:

We discovered that 97.33% of the population has hyperthyroidism, while only 2.67% do not after performing data cleaning and EDA on the dataset. Additionally, we now know that there are more female patients than male patients. Patients make up 84.0% of the total and men make up 16.0 %. We have used various machine learning methods to forecast thyroid illness more accurately. Here, the models are trained to recognise if a person has thyroid disease or not.

REFERENCES:

[1]. Thyroid Detection Using Machine Learning" Published Online in IJEAST in January 2021.
[2]. Ankita Tyagi, Ritika Mehra, and Aditya Saxena's paper, "Interactive Thyroid Disease Prediction System Using Machine Learning Technique," was presented at the 5th IEEE international conference on parallel, distributed, and grid computing which took place from December 2022.
[3]. Gyanendra Chaubey, Dhananjay Bisen, Siddharth Arjaria, and Vibhash Yadav's study, "Thyroid Disease Prediction Using Machine Learning Approaches," was published in the National Academy of Science's India journal.
[4]. https://www.javatpoint.com/machine-learning
[5]. Reference for the Dataset: UCI Repository, available at: https://archive.ics.uci.edu/ml/datasets/thyroid+disease
[6]. https://labtests.triiodothyronine-t3-tests.medlineplus.gov
[7]. Research on various data mining categorization methods published in the IRJET journal.
[8]. Shaik Razia and M.R. Narasinga Rao's evaluation of machine learning techniques for diagnosing thyroid disorders was published in the Indian Journal of Science and Technology.
[9]. Machine learning-based disease prediction.
[10]. Umar Sidiq, Dr. Syed Mutahar Aaqib, and Dr. Rafi Ahmed Khan's classification of thyroid disorders utilising data mining techniques.
[11]. By Michael Yeh, "Normal-thyroid-hormone-levels."