PREDICTION AND PROVIDING MEDICATION FOR THYROID DISEASE USING MACHINE LEARNING TECHNIQUE

Dr. M. Ramasubramanian[1], Pothula. Tulasi Sri Sai Pratyusha[2], Venturi. Divya Maha Santoshi[3], Sirimalla. Sreeja[4]

Article History: **Received:** 08.02.2023 **Revised:** 23.03.2023 **Accepted:** 08.05.2023

Abstract

The advancement of medical diagnosis and prognosis is mostly driven by thyroid disorders, which are a complex concept in medical research. One of the major organs in our body is the thyroid gland. The release of thyroid hormones is in charge of controlling metabolism. The two most common thyroid conditions that cause thyroid hormones to be produced for the regulation of body metabolism are hyperthyroidism and hypothyroidism. In order to research and classify thyroid illness using data from hospital databases, machine learning is essential for disease prediction. To tackle dynamic learning problems like medical diagnosis and prediction tasks, a good knowledge base must be assured, constructed, and used as a hybrid model. The identification and inhibition of the thyroid are accomplished using fundamental machine learning techniques. The SVM is used to estimate the likelihood of a thyroid patient with reasonable accuracy. Our system must make recommendations if the patient is at risk for thyroid disease, including home remedies, safety measures, medicine, etc.

Keywords: Machine learning, Support vector machine.

[1]Professor, Department of Computer Science & Engineering, Sridevi Women's Engineering College, Hyderabad, Telangana, India.
[2,3,4]Final Year B. Tech, Department of Computer Science & Engineering, Sridevi Women's Engineering College, Hyderabad, Telangana, India

Email: [1]mailtoraams@gmail.com, [2]pratyupratyusha2001@gmail.com, venturidivyams@gmail.com, sreeja.sirimalla333@gmail.com

DOI: 10.31838/ecb/2023.12.s3.277

1. Introduction

Healthcare employs cutting-edge machine biology. For the purpose of predicting medical diseases, data collection was necessary. Various intelligent prediction algorithms are used to detect diseases in their early stages. The medical information system is proficient with data sets, however there are no intelligent tools accessible for rapid illness diagnosis. When building a prediction model, machine learning algorithms ultimately play a crucial role in resolving challenging non-linear problems. Any disease prediction model requires characteristics that can be chosen from the various data sets and used to describe a healthy patient as precisely as possible. Otherwise, misclassification might cause a good patient to get the wrong kind of care. The difficulty of predicting any thyroid-related problem with thyroid illness is also of the greatest cardinal number. In the stomach, there is an endocrine thyroid gland. It is situated beneath the apple of Adam in the lowest region of the human neck and aids in the release of thyroid hormones, which in turn impact metabolic rate and protein synthesis. These hormones take into account the rate at which the heart beats and the rate at which calories are burned to regulate body metabolism. The thyroid hormones' chemical makeup aids in regulating the body's metabolism. Two mature levothyroxine (abbreviated T4) and triiodothyronine thyroid hormones (abbreviated T3) are present in these glands. These thyroid hormones are necessary for the building, general upkeep, and control of body temperature. The only two thyroid hormones that are normally produced by thyroid glands are T4 and T3.

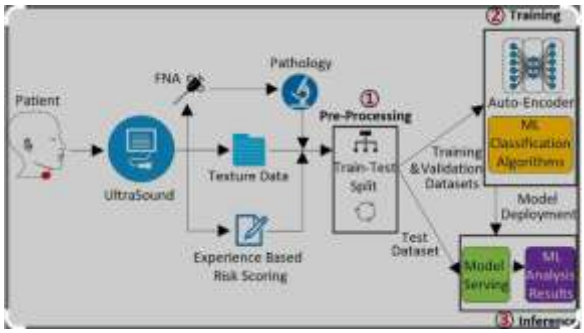

Fig.1: model figure

These hormones are essential for the regulation of proteins, their distribution at body temperature, their energy-bearing capacity, and their spread throughout the entire body. Iodine is a crucial component of T3 and T4 hormones and is only deficient in a small number of rare but extremely common issues with the thyroid glands. These hormones contribute insufficiently to hypothyroidism and excessively to hyperthyroidism. There are several causes of both hyperthyroidism and underactive thyroid. Numerous drugs are available. Ionising radiation, ongoing thyroid softening, an iodine deficit, and the loss of enzymes that create thyroid hormones make thyroid surgery weak.

Literature Review

Using Artificial Neural Networks to Diagnose a Thyroid Condition:

Finding the right disease diagnosis before treating it is a major challenge in medical science. This paper presents the analysis of thyroid issues utilizing fake brain organizations (ANNs). Three ANN algorithms were utilized in the training of the feed-forward neural network; the Back spread calculation (BPA), the spiral premise capability (RBF) Organizations, and the learning vector quantization (LVQ) organizations. Using MATLAB, the networks are simulated and their performance is evaluated in terms of things like diagnosis accuracy and training time. The best model for diagnosing thyroid disorders can be found through performance comparison.

A clever mixture strategy in view of fake resistant acknowledgment framework (AIRS) with fluffy weighted pre-handling for thyroid illness determination:

Appropriate translation of the thyroid organ utilitarian information is a significant issue in the finding of thyroid sickness. The thyroid gland's primary function is to aid in metabolism regulation. This is provided by the thyroid hormone that the thyroid gland makes. The kind of thyroid disease is determined by whether too little thyroid hormone is produced (hypothyroidism) or too much hormone is produced (hyperthyroidism). A new but effective subfield of artificial intelligence are artificial immune systems (AISs).

Among the frameworks proposed in this field up until this point, the fake safe acknowledgment framework (AIRS), which was proposed by A. Watkins, has shown a compelling and charming execution on the issues it was applied.

Using this classification system and a novel hybrid machine learning approach, the purpose of this study is to diagnose thyroid disease. This diagnosis issue can be resolved through classifying by combining AIRS and a newly developed fuzzy weighted pre-processing. Using cross-validation, the method's robustness to sampling variations is investigated. We utilized the UCI machine learning respiratory thyroid disease dataset. The highest classification accuracy we have achieved thus far is 85%. Through 10-fold cross-validation, the accuracy of the classification was achieved.

An examination of neural networks in the diagnosis of thyroid function:

We examine the capability of fake brain networks in diagnosing thyroid illnesses. We demonstrate the connection between conventional Bayesian classifiers and neural networks.

As a result of their ability to provide accurate estimates of posterior probabilities, neural networks may perform better in classification than conventional statistical techniques like logistic regression. In addition, it is demonstrated that the neural network models are unaffected by changes in the sampling. It is shown that for clinical finding issues where the information are much of the time exceptionally unequal, brain organizations can be a promising order strategy for reasonable use.

An examination of how machine learning methods can be used to treat diseases:

Databases and repositories have grown at an exponential rate over the past few years as scientific knowledge has grown and a lot of data has been created. One of the rich data domains is the biomedical domain.

There is currently a large amount of biomedical data available, including information about clinical symptoms, biochemical data, and imaging device outputs. Due to the vast, dynamic, and complex knowledge in the biomedical domain, it is challenging to manually extract biomedical patterns from data and convert them into machine-understandable knowledge. Information mining is equipped for working on the nature of removing biomedical examples. An overview of how data mining can be used to manage diseases is presented in this study. The investigation of machine learning (MLT) methods, which are frequently utilized for the prediction, prognosis, and treatment of significant common diseases like cancer, hepatitis, and heart disease, is the primary focus.

Analyses and examples are provided for the Artificial Neural Network, K-Nearest Neighbor, Decision Tree, and Associative Classification methods. A general assessment of the state of disease management using MLT is provided by this survey. Depending on the disease, the solved problem, and the data and method used, the various applications' achieved accuracy ranged from 70% to 100%.

2. METHODOLOGY

A difficult axiom of medical science, thyroid disorder is a significant cause of medical diagnosis and estimation. Regulating metabolism is the responsibility of thyroid hormone secretions. Hyperthyroidism and hypothyroidism are the two predominant thyroid problems that discharge thyroid chemicals to control body digestion. Information cleaning strategies have been utilized to cause information sufficiently crude to do to investigation to show the probability of patients having thyroid.

1. The thyroid condition is the primary factor in the development of medical diagnoses and predictions, and medical research is based on a complicated axiom.
2. The study and classification of thyroid disease based on data from hospital datasets, as well as the process of disease prediction, require machine learning.

In the forecast cycle, AI assumes a key part, and paper research and the characterizations of models utilized in thyroid illness depend on data from UCI AI storehouses. To address complex learning issues, such as medical diagnostics and statistical tasks, it is necessary to preserve a decent knowledge base that can be centered and utilized as a hybrid paradigm. Additionally, we suggested various methods for thyroid diagnosis and machine learning. An estimated probability of a patient having thyroid disease was calculated using Machine Learning Algorithms, Vector Support Machines, Decision Trees, and K-NN.

Advantages:

1. Avoids long-term risks of anti-thyroid and radioactive iodine.
2. Provides histology tissue, for childbearing instantly.

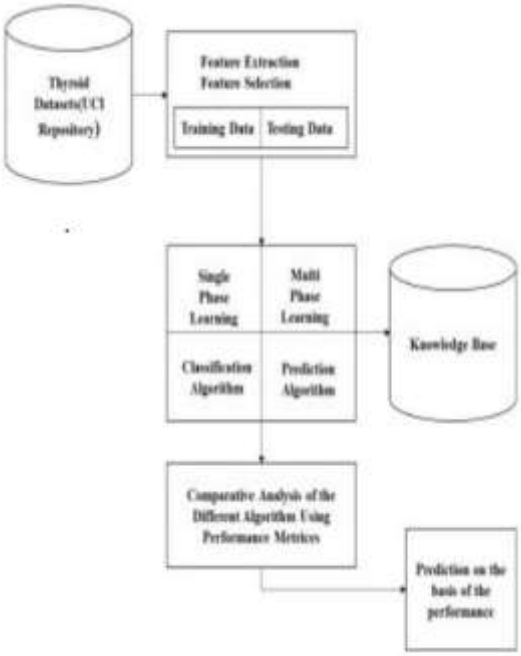

Fig.2: System architecture

Modules:
In this project, we have designed the following modules

Data Collection
The initial owners of the dataset carried out this step. Additionally, the dataset's composition comprehends the connection between various features. a graph of the entire dataset and the core features. The dataset is further divided into two-thirds for training and one-third for algorithm testing. Additionally, each class in the full dataset is roughly evenly distributed across the training and testing datasets in order to produce a representative sample. The varying proportions of the testing and training datasets utilized in the paper.

Data Preprocessing
The collected data may have missing values that could cause inconsistencies. The efficiency of the algorithm can be enhanced by preprocessing the data in order to achieve better outcomes.
The anomalies must be eliminated and furthermore factor change should be finished. To conquer these issues we utilize the guide capability.

Model Selection
Predicting and recognizing patterns and producing appropriate outcomes after comprehending them are the fundamentals of machine learning. ML calculations concentrate on designs in information and gain from them. Every attempt will teach and improve an ML model. It is essential to first divide the data into training and test sets in order to evaluate a model's efficacy.

Predict the outcomes
The performance of the designed system is guaranteed after it is put through its paces with a test set. The description and modeling of regularities or trends for things whose behavior changes over time is known as evolution analysis. Precision is one of the most frequently derived metrics from the confusion matrix. Accuracy. The ability to construct a predictive model using a standard SVM model is the most significant feature of these features.

Implementation
Support Vector Machine(Svm)
Classification and regression problems can be solved using the "Support Vector Machine" (SVM) method of supervised machine learning. However, classification issues are the most common applications. With this algorithm, the value of each feature is the value of a specific coordinate, and each data point is represented as a point in an n-dimensional space (where n is the number of features you have). After that, we perform classification by locating the hyperplane that effectively differentiates the two classes (for an illustration, see the image below). The SVM algorithm is actually implemented with a kernel. It is beyond the scope of this introduction to SVM to discuss how to learn the hyperplane in linear SVM by converting the problem using some linear algebra. The way that the internal result of any two provided information might be utilized to reevaluate the direct SVM as opposed to the actual perceptions is a critical leap forward. The inner product of any two vectors is the sum of the multiplications of each pair of input values. For

example, 2*5 + 3*6 or 28 is the inward result of the vectors [2, 3] and [5, 6]. The accompanying condition might be utilized to gauge another info utilizing the spot item between the information (x) and each help vector (xi): B0 minus sum(ai*(x, xi)) yields f(x).
This equation calculates the inner products of each support vector in the training set with a new input vector (x). The training data must be used by the learning algorithm to estimate the coefficients B0 and ai for each input.

Random Forest Algorithm
The Random Forest algorithm is a technique for supervised classification. It's clear from the name that the goal is to make a forest in the wrong way. The results of a forest will be more accurate the more trees there are; On the other hand, the results of a forest will be less accurate the fewer trees there are. To be clear, using the information gain or gain index technique to build the decision is not the same as building the forest. The decision tree is a tool for making decisions easier. The potential repercussions are depicted on a graph that resembles a tree. A set of rules will be produced by the decision tree when a training dataset containing targets and features is fed to it. These rules can be used to make predictions.
We can use Random Forest to assist us in extracting classes from our dataset if we divide it into three categories. A collection of decision trees is a random forest; A decision tree will generate a set of rules that can be used to make predictions if you feed it a training dataset with characteristics and labels

.**Naive Bayes Classifier:**
Naive Bayes is a classification technique concept that states that all characteristics are independent and unconnected. It specifies that the state of one feature in a class does not impact the status of another. It is a strong method used for classification since it is based on conditional probability. It works well with data that has unbalanced values and missing values. The Bayes Theorem is used by Naive Bayes, a machine learning classifier. Using the Bayes theorem to calculate posterior probability P(C|X) can be calculated from P(C), P(X), and P(X|C).
Therefore ,P(C|X)=(P(X|C)
P(C))/P(X)
Where, P(C|X)=target class's posterior probability.
P(X|C)=predictor class's probability.
P(C)=class C's probability of being true.
P(X)=predictor's prior probability.

Decision Tree Classifier:
A Decision Tree is a classification problem-solving supervised machine-learning method. The primary goal of using a Decision Tree in this study effort is to forecast the target class using a decision rule derived from past data. It predicts and classifies using nodes and internodes. Root nodes categorise instances based on various characteristics. The root node may contain two or more branches, whereas the leaf node represents categorization. The Decision tree selects each node at each step by assessing the maximum information gain across all qualities. The Decision Tree technique's evaluated performance

Boosting:
Freund and Schapire introduced the ensemble modeling strategy known as "boosting" in 1997. Since then, Boosting has grown in popularity as a method for dealing with problems with binary classification. These calculations help expectation power by changing countless frail students into solid students.
The fundamental premise of boosting techniques is that, after developing a model based on the training dataset, we develop a second model to correct any errors in the first one. This procedure is repeated until the dataset can be accurately forecasted and errors are reduced. Let's look at an example to better understand this. Suppose you used the Titanic dataset to create a decision tree algorithm with an 80 percent accuracy. After that, a different method is used to check the accuracy, which turns out to be 75% for KNN and 70% for linear regression.
We discovered that the accuracy varied when a different model was constructed using the same dataset. In any case, imagine a scenario in which we consolidate these calculations to show up at the last forecast. We will get results that are more precise by averaging the data from these models. We can improve the accuracy of our forecasts by doing this. Similar to this, boosting algorithms achieve the desired result by combining numerous models (weak learners). In this post, we'll learn the math behind a few different ways to boost. The majority of boosting algorithms are divided into one of three groups: Calculation AdaBoost Slope drop is a method. method for extreme descent from a gradient.

Gradient Boosting Algorithm:
Gradient boosting classifiers are a type of machine learning approach that combines multiple weak learning models to create a powerful prediction model. Gradient boosting typically use decision trees. Gradient boosting models are gaining popularity and have recently been successful in a number of Kaggle data science competitions due to their ability in categorising huge datasets

. **Logistic Regression:**

Logistic regression is a type of predictive analysis. We use logistic regression to describe data and explain the relationship between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables.

3. EXPERIMENTAL RESULTS

Fig.3: Home screen

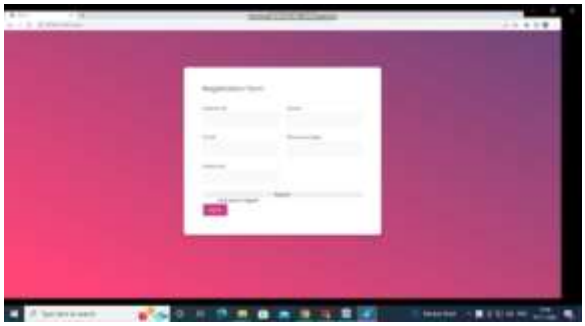

Fig.4: User registration

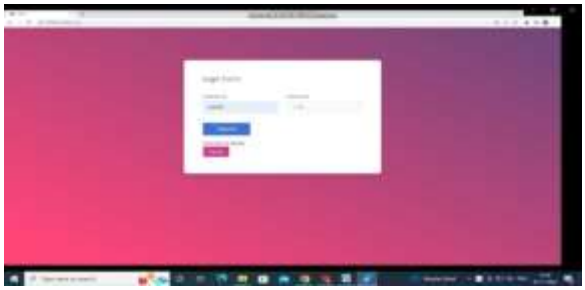

Fig.5: User login

Fig.6: Main page

Fig.7: User input

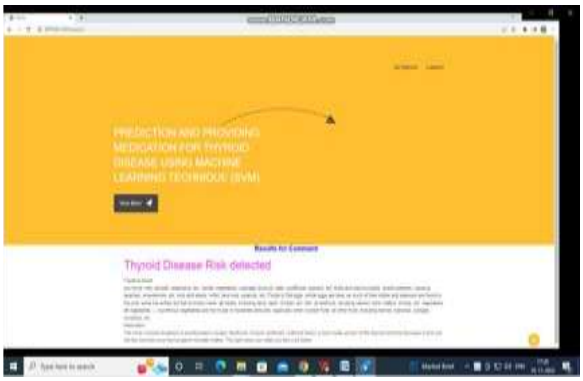

Fig.8: Prediction result

4. CONCLUSION

The study effort examines the novel machine-learning techniques that can be used to diagnose thyroid disorders. Recent years have seen the development and application of various practicable methods for the accurate and expert diagnosis of thyroid illness. The research shows that the various technologies utilised in the two publications exhibit varying degrees of accuracy. The majority of scholarly studies show that the neural network performs better than alternative tactics. On the other side, this is also because the decision tree and assist vector machine have performed effectively. There is no doubt that the ability of professionals to diagnose thyroid illnesses has significantly improved, but it is advised that the number of criteria patients employ to do so be kept to a minimum. A patient must conduct a wider variety of time-consuming and cost-effective health examinations as a result of more features.

Future Scope

To diagnose thyroid illness, one must meet a minimum number of criteria, thus it is necessary to build algorithms and prediction models of thyroid disease. This will save patients' time and money. There are certain approaches that we believe can be developed further and used in further studies.

5. REFERENCES

L. Ozyılmaz and T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods," in Proceedings of ICONIP'02 9th international conference on neural information processing (Singapore: Orchid Country Club, 2002) pp. 2033–2036.

K. Polat, S. Sahan and S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis," Expert Systems with Applications, vol. 32, 2007, pp. 1141-1147.

F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM," in 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009.

G. Zhang, L.V. Berardi, "An investigation of neural networks in thyroid function diagnosis," Health Care Management Science, 1998, pp. 29-37. Available: http://www.endocrineweb.com/ thyroid.html, (Accessed: 7 August 2007).

V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, 2012.

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

Breiman L. Statistical Modeling: the two cultures. Stat Sci. 2001; 16:199-231.

Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. 2017; 9:245-250

Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015; 521: 452-459.

Azimi P, Mohammad I HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. J Neuro l Neurosurg Psychiatry. 2015; 86:251-256.