Thyroid Disease Classification using Machine Learning Algorithms

R. P. Ram Kumar[1*], *M.* Sri Lakshmi[2], *B. S.* Ashwak[3], *K.* Rajeshwari[4], S Md Zaid[5]

[1]Department of AIMLE, GRIET, Hyderabad, Telangana State, India
[2]Associate Professor, Department of CSE, GPCET, Kurnool, India
[3]Student, Department of CSE, GPCET, Kurnool, India
[4]Student, Department of CSBS, GRIET, Hyderabad, Telangana State, India
[4]Department of AIMLE, GRIET, Hyderabad, Telangana State, India
[5]Student, G Pullaiah College of Engineering and Technology, Kurnool, India

Abstract: Thyroid gland is one of the body's most important glands because it regulates the metabolism of the human body. It controls how the body works by releasing specific hormones into the blood. The two different hormone disorders are hypothyroidism and hyperthyroidism. When these disorders occur, the thyroid gland releases a particular hormone into the blood that regulates the metabolism of the body. Iodine deficiency, autoimmune conditions, and inflammation can contribute to thyroid issues. The disease is diagnosed using a blood test, but there is frequently some noise and disturbance. Techniques for cleaning data can be used to make it simple enough to perform analytics that show the patient's risk of developing thyroid disease. This paper deals with the analysis and classification models used in thyroid disease based on the information gathered from the dataset taken from the UCI machine learning repository. Machine learning plays a crucial role in the detection of thyroid disease. This paper suggests various machine-learning methods for thyroid detection and diagnosis for thyroid prevention.

1 Introduction

Computational biology developments in health care had made it possible to compile patient data that had been saved to predict medical diseases. Numerous intelligent prediction algorithms are available for the early diagnosis of the disease. There is no such intelligent system that can analyze the information quickly, with multiple data sets in the medical information system. For In developing a prediction model, several kinds of nonlinear problems must be solved, and here machine learning plays a very vital role. The prediction models should work on the features that can be selected from various datasets and accurately and quickly recognize a healthy patient. Otherwise, misclassification might result in a healthy patient receiving unnecessary care. As a result, predicting any disease in conjunction with thyroid disease has the highest cardinality. Thyroid gland is an endocrine gland that is present

[*] Corresponding author: ramkumar1695@grietcollege.com

in the neck. Thyroid gland secretes hormones that, in turn, affect the rate of metabolism and protein synthesis, it develops beneath Adam's apple in the shortened region of human neck. Thyroid hormones help regulate the body's metabolism in many ways, including monitoring how quickly the heart beats and calories are burned. The thyroid gland helps dominate metabolism by producing thyroid hormone. Levothyroxine (also known as T4) and triiodothyronine (also known as T3) are the two active hormones secreted by the thyroid gland. These hormones are crucial for fabrication, as well as for comprehensive construction and supervision, to control body temperature. Specifically, thyroxine and triiodothyronine (T3) are the two types of active hormones customarily composed by the thyroid glands. These hormones are decisive in protein management, dissemination in the body temperature, along with energy-bearing and transmission in every part of the body. Iodine is the main building block of these two hormones. i.e., T3 and T4). It is prostrated in a few specific problems, some exceptionally prevalent. Insufficient thyroid hormone results in hypothyroidism, and excess thyroid hormone results in hyperthyroidism. There are many origins related to hyperthyroidism and underactive thyroids. There are various kinds of medications. Thyroid surgery is liable to ionizing radiation, continual thyroid tenderness, a deficiency of iodine, and a lack of enzymes to make thyroid hormones.

2 Literature survey

Ankitha Tyagi and Mehra employed a variety of classification algorithms in this work, including the k-Nearer-Neighbor algorithm, decision trees, support vector machines, and artificial neural networks. Classification and prediction based on the data set obtained from the UCI Repository were carried out, and accuracy was obtained based on the output produced. In order to find the best technique with the highest accuracy, the accuracy of the algorithms used has been analyzed and comparisons have been made. Sunshine Godara [7] presented an approach to analyze the thyroid dataset; they used SVM machine learning and logistic regression. Based on Precision, Recall, F measure, ROC, and RMS error, a comparison between these two algorithms was made. In the end, the best classifier was logistic regression. Yang Feng Wang [5] illustrated a approach using thyroid ultrasound images. A benign or malignant thyroid nodule can be identified using image analysis, radiomics, and deep learning-based methods. The effectiveness of these two strategies is compared. The classification accuracy, sensitivity, and specificity of the radiomics-based method are 66.81%, 51.19%, and 75.77%, respectively, while the evaluation indices for the deep learning-based method trained on the test samples are 74.69%, 63.10%, and 80.20%, respectively. The most effective methods ended up being deep learning.

3 Problem definition

Based on the current statistics, thyroid problems or disorders are on the rise in India. As a result of a large-scale survey conducted in India in 2021, about 13% of respondents above 60 years of age suffer from thyroid disorders. Approximately 1 in 10 adults suffer from a thyroid disorder. Predicting thyroid disease by doctors is a tedious task. The blood test reports may contain noise and disturbances, leading to pessimistic predictions, and only experienced doctors can predict them correctly. In these situations, machine learning plays a vital role in predicting the burden of these errors and helping to easily predict the disease.

3.1 Dataset Description

The dataset was collected from the UCI Machine Learning Repository website [14]. There are 3772 instances in the dataset. In this dataset, 3481 instances belong to the category "negative," 194 instances belong to the "compensated hypothyroid" category, 95 instances belong to the "primary hypothyroid" category, and 2 instances belong to the "secondary hypothyroid" category. There are a total of 30 attributes. In the research, we considered only 12 attributes. ‘F’ represents false, while ‘T’ represents true.

Table 1. Hypothyroid dataset description.

S No	Attribute Name	Value Type
1	class	Negative, Compensated_thyroid, primary_thyroid, secondary thyroid
2	On thyroxine	F, T
3	Pregnant	F, T
4	TSH measured	F, T
5	TSH	Continuous
6	Goitre	F, T
7	T3	Continuous
8	TT4 measured	F, T
9	TT4	Continuous
10	Query hypothyroid	F,T
11	Thyroid Surgery	F,T
12	FT1	Continuous

3.2 Methodology

3.2.1 Preprocessing

Preprocessing is a technique in which the volume of the data is reduced. Data preprocessing is part of data preparation. There are many preprocessing techniques. In our work, we have used the dimensionality reduction technique to select a subset of attributes from the original data.

3.2.2 Classification

Classification is a data mining technique that is used to group instances that belong to the same category. It is supervised learning that groups data based on training data. Examples of data mining classification techniques are decision trees and neural networks.

3.2.3 Decision Tree

A decision tree is one of the supervised techniques that can be used for both classification and regression. It's a graphical representation for getting all possible solutions to a problem/solution based on the given conditions. The decisions are based on the dataset. In our work, we have used two decision tree classifiers J48 and Decision Stumps. Algorithms for J48 and Decision tree are discussed in the following section.

3.3 Algorithms

3.3.1 J48 Algorithm

The J48 algorithm is machine learning algorithm used to examine the data categorically and continuously. The J48 algorithm was developed by Ross Quinlan and is based on the iterative Dichotomizer 3 algorithm. To divide a root node into a subset of two partitions until a leaf node (the target node) appears in the tree, J48 employs the divide-and-conquer algorithm. The following steps are used to build the tree structure given a set T of all possible instances.

- Step 1: If T has fewer instances or all instances in T belong to the same group class, T is leaf-labeled with the tree's most prevalent class.
- Step 2: If step 1 is unsuccessful, choose a test based on a single attribute with two or more possible results. Then divide T into corresponding T1, T2, T3..., according to the result for each respective case, and the same may be applied in a recursive manner to each sub-node. Next, think of this test as the root node of the tree, with one branch for each test outcome.
- Step 3: The algorithm J48 uses two heuristic criteria to rank the information gain and default gain ratios.

3.3.2 Decision stump

The decision stump is a machine learning model that contains a one level decision tree. That is, it is a decision tree with one internal node, which is immediately connected to the next node, or the terminal tree. The decision stump makes the prediction based on the value of just one feature.

4 WEKA tool

Weka is a data mining and machine learning tool, that has collection of algorithms. It contains tools for data preparation, classification, clustering, and visualization. Weka ia named after a flightless bird with an inquisitive nature that is found only in New Zealand. Weka is a open source software issued under the GNU General Public License. Weka is portable, since it is fully developed using the Java programming language and thus runs on almost every modern computer. The use of Weka results in the quick development of several machine learning and data mining models. Weka can be downloaded from the website.

5 Performance measure of algorithms

The performance measure of an algorithm says how effectively an algorithm can function. Performance measures are used to analyze how “on track” a project or a program is to achieve the desired outcome. In our work, a dataset is provided to the J48 algorithm. The performance measures of the algorithms are performed using a confusion matrix.

5.1 Confusion Matrix

A confusion matrix is a two-dimensional table that is used to calculate the performance of classification algorithm. A confusion matrix defines the performance of the classification algorithm.

Table 2. Confusion matrix.

	Actually Positive (1)	**Actually Negative (0)**
Predicted Positive (1)	True Positives (TPs)	False Positive (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

In confusion matrix, correctly classified instances are calculated as the sum of diagonals True Positive (TP) and True Negative (TN). Incorrectly classified instances are calculated using False Positive (FP) and False Negatives (FN).

5.2 Error rate

The error rate is calculated as the number of incorrect predictions that are made divided by the total number of datasets. The best case for error is 0.0, and the worst case is 1.0. [15]
Error rate = (FP+FN)/(TP+FP+FN+TN)

5.3 Accuracy

Accuracy is defined as the number of correct predictions divided by the total number of datasets. The best case of accuracy is denoted by 1.0, and the worst case of accuracy is denoted by 0.0. [15].
Accuracy = (TN+TP)/(TP+TN+FP+FN)

6 Result analysis

Overall, there are 3772 instances in the hypothyroid dataset. All the instances are classified under negative, compensated hypothyroid, primary hypothyroid, or secondary hypothyroid. The following matrix represents the confusion matrix of the decision stump algorithm.

Table 3. Confusion matrix for Decision stump.

Target	**Negative**	**Compensate Hypothyroid**	**Primary Hypothyroid**	**Secondary Hypothyroid**
Negative	3404	77	0	0
Compensate Hypothyroid	0	194	0	0
Primary Hypothyroid	0	95	0	0
Secondary Hypothyroid	2	0	0	0

The decision stump classifier identified 3598 instances correctly and 174 instances incorrectly. The following matrix represents the confusion matrix for J48 algorithm

Table 4. Confusion matrix for J48 algorithm.

Target	Negative	Compensated Hypothyroid	Primary Hypothyroid	Secondary Hypothyroid
Negative	3476	3	2	0
Compensate Hypothyroid	0	192	6	0
Primary Hypothyroid	2	5	88	0
Secondary Hypothyroid	2	0	0	0

J48 algorithm identified 3756 instances correctly and 16 instances incorrectly. Accuracy of the algorithms is shown in below table

Table 5. Accuracy of algorithms.

Classifier	Accuracy
Decision Stump	95.38%
J48	99.57%

$$\text{Misclassification Error Rate} = 1\text{-Accuracy} \quad (1)$$

Formula (1) is used to find the misclassification error rate.

7 Discussions

Various classification methods are available in WEKA. These thyroid diseases, as well as some other clinical diagnosis problems, are diagnosed using these classification techniques. Numerous researchers used various techniques to identify thyroid disease, according to studies, and they were successful in producing classifiers that were highly accurate for the dataset they used, which was taken from the UCI machine learning repository. [K. Saravana Kumar et al 2014] proposed KNN and SVM classification algorithms on thyroid disease diagnosis. They showed that the prediction accuracy of SVM is 94.4336%. However, KNN accuracy is 96.343%. [Pandy et al. (2015)] proposed a classification algorithm using random forest and C4.5 that has a 99.47% prediction accuracy. The UCI machine learning repository's website was used to download the hypothyroid dataset that we used in our research. The hypothyroid dataset consists of 3772 instances, of which 3481 instances fall under the category of negative hypothyroidism, 194 instances fall under the category of compensated hypothyroidism, 95 instances fall under the category of primary hypothyroidism, and 2 instances fall under the category of secondary hypothyroidism. 29 characteristics are present in total. Only 12 attributes were selected for our research project so that we could classify the data. To enhance the performance of the classifier, a subset of the original data's attributes was chosen using dimensionality reduction. 29 attributes are present in the original dataset. To increase accuracy, we chose a subset of 12 out of 29 attributes using the ranking method. To diagnose thyroid disease, we proposed the two classification algorithms J48 and decision stump. With the help of the confusion matrix, classifier performance is assessed and 3756 instances were correctly identified by the J48 classifier, while 16 instances were incorrectly identified. The correctly identified instances in the decision stump classifier, however, are 3598, while the incorrectly identified instances are 174. In this experiment, the classifier J48 outperformed the decision stump with an accuracy of 99.58% and a minimum error rate of 424.

Table 6. Error rate of algorithms.

Classifier	Error Rate
Decision stump	4.61
J48	0.424

In health care industry, the major problem faced is diagnosis of the disease. In the process of making decisions, many data mining techniques are employed. In our work, we used dimensionality reduction to choose the subset of attributes from the original data and we applied J48and decision stump data mining classification techniques which are used to classify the hypothyroid disease. The confusion matrix is used to assess the accuracy and error rate of classifiers. The J48 algorithm yields 99.58%, which is more accurate than decision stump tree accuracy and has a much lower error rate. In a follow-up study, the same methodology is applied to other disease datasets, including those related to heart disease, breast cancer, lung cancer, and other conditions.

References

1. A. T. Azar, A. E. Hassanien, and T. Kim, AI, arXiv:1403.0522, (2012)
2. K. Salman, E. Sonuc, J. Phys, **1963**, 012140, (2021)
3. A. C. C. Heuck, “World Health Organization,” 2000, https://www.who.int/
4. A. Tyagi and R. Mehra, *Interactive Thyroid Disease Prediction System using Machine Learning Techniques*, in the Proceedings of the 5th IEEE International Conference on Parallel, Distributed Grid Computing (PDGC-2018), 20-22 Dec, 2018, Solan, India, (2018)
5. Y. F. Wang, *Comparison Study of Radiomics and Deep-Learning Based on Methods for Thyroid Nodules Classification Using Ultrasound Images*, **8**, IEEE Access, (2020)
6. R. Chandan, M. S. Chethan, C. Vasan, H. S. Devikarani, Int. J. Eng. App. Sci. Tech. 5, 9, (2021)
7. S. Godara, Intl. J. Elect. Engg. **10**, 2, (2018)
8. P. Pavani, P. P. Sadu Naik, Int. J. Inno. Res. Tech. **9**, 3, (2022)
9. Sunila, R. Singh, and Sanjeeev Kumar, Ind. J. Sci. and Tech. **9**, (2016)
10. S. Godara and R. Singh, Ind. J. Sci. and Tech. **910**, (2016)
11. Z. Obermeyer, E. Manuel, N. Engl. **375**, (2016)
12. A.K. Pandey, P. Pandey, and K.L. Jaiswal, IUP J. Comp. Sci., **7**, 3, (2013)
13. S. Ismaeel, A. Miri, and D. Chourishi, *Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis*, in Proceedings of the IEEE Canada International Humanitarian Technology Conference, (2015)
14. Thyroid dataset, “UCI Machine Learning Repository”, https://archive.ics.uci.edu/ml/datasets/thyroid+disease
15. Evaluation metrics, “Basic measurements derived from confusion matrix”, https://classeval.wordpress.com/introduction/basic-evaluation-measures/