Journal of Applied Science and Education, 2023,
Vol. 00, Iss. 00, S. No. 000, pp. 0-00
ISSN (Online): 2583-1372

Thyroid Disease Detection Using Machine Learning

Mohammad Faiz[1], Syed Wajahat Abbas Rizvi[2]

Amity School of Engineering and Technology Lucknow, Amity University Uttar Pradesh, India
mohammadfaiz9935@gmail.com[1], swarizvi@lko.amity.edu[2]

How to cite this paper: M. Faiz, and S. W. A. Rizvi, "Thyroid Disease Detection Using Machine Learning," *Journal of Applied Science and Education (JASE)*.

https://doi.org/10.54060/jase.v3i2.32

Received: 21/04/2023
Accepted: 07/06/2023
Online First: 29/07/2023

Abstract

The thyroid gland, which is in charge of controlling metabolism and other biological functions, is affected by thyroid illness, a frequent medical disorder. Successful thyroid disease management and therapy depends on early diagnosis and treatment. Recent years have seen the development of numerous machine learning methods and artificial intelligence (AI) algorithms to help with the early detection and diagnosis of thyroid disease. These methods entail evaluating a range of patient data, such as laboratory results, imaging studies, and clinical complaints. These algorithms can find patterns and correlations in vast volumes of patient data that might not be obvious to human experts. This may result in earlier identification and more precise diagnosis of thyroid illness, enhancing patient outcomes and lowering medical expenses. Additional study and development are necessary to improve these methods and incorporate them into clinical practice.

Keywords

Machine Learning, Thyroid Disease Detection, Decision Tree, Random Forest.

1. Introduction

Millions of individuals worldwide suffer from a thyroid condition, a prevalent glandular system ailment. For effective treatment and management, it is important to find and diagnose thyroid disease as soon as possible. But thyroid disease can be hard to identify because its symptoms aren't always clear and can be like those of other illnesses.[2] Laboratory tests and medical scans are often used to help figure out what's wrong with the thyroid, but these methods can take a long time and cost a lot.

Thyroid disease detection and diagnosis have been greatly aided by the advent of machine learning algorithms and artificial intelligence (AI) techniques in recent years.[15] These algorithms can look at a huge amount of patient data, like lab tests,

medical images, and clinical symptoms, to find patterns and connections that may not be obvious to human experts. Using machine learning algorithms to find thyroid disease has several benefits, such as finding the disease early, getting a correct diagnosis, and making personalized treatment plans.

The goal of this study is to give an overview of the current state of the art in using machine learning to find thyroid diseases.[5] It will talk about the problems with standard methods of diagnosis, the benefits of using machine learning algorithms, and the different techniques and algorithms used to find thyroid disease. The study will also look at the limits and future directions of machine learning in detecting thyroid disease, as well as how it might affect the way healthcare is given.[13] In order to predict Thyroid in individuals, we have tested a wide range of machine learning techniques. We evaluated the performance of eleven various models to develop a prediction model, including:

- K Nearest Neighbors
- Decision Tree
- Support Vector Machine
- Random Forest

2. Proposed Model

The ML-based prediction model that suggested at each level. The first thing to do is to look at research data. The second step is to process the data before it is used. In the third step, we will change all these numbers to 'nan' values. Then, in the fourth step, we deal with variables that don't have any values, and in the fifth step, we deal with nominal categorical variables. Then, we balance the data to make better predictions. In this step, the datasets that have already been cleaned up are sent to different machine-learning methods. In the final stage, involves analyzing the algorithmic results using a range of measures. The best-performing model out of all the ML algorithms used is saved and used at a later stage.

3. Methodology

First, we do exploratory data analysis on the downloaded data set, and then we do pre-process on it. As we advance further into the next phase of the work that is what we call as the Data pre-processing step, the relationships between the dataset's traits are looked at to find features that help to predict disease. Then, the information is split into two distinct categories train and test. Several machine learning methods and the training set are employed in the development of predictive ML models. The proposal's success is then judged based on several parameters. At last, the best ML model is put into use. Here is a quick look at how each part works:

3.1. Data Collection

The data was found on Kaggle and got from there. In the data set, there are 3221 events and 28 attributes. Below is a list of the attributes that make up the dataset.

- S.no - Serial number.
- Age – The patient’s age that is given in year old format.
- Sex – Whether the person is male, or female is depicted by this attribute.
- On Thyroxine - It indicates whether the patient is set on the intake of thyroxine or not.
- Query on Thyroxine - Its signifies whether the patient is having query on thyroxine.
- On Antithyroid Medication - It states whether the patient is on Antithyroid Medication or not.
- Sick – This attribute informs regarding the health of the patient, i.e., healthy or unhealthy.
- Pregnant – This depicts whether the patient is in her gestation period.

- Thyroid Surgery - Signifies that the patient has done any surgery in the past or not.
- I-131 Treatment - Hyperthyroidism and thyroid cancer can be treated with I-131 radiotherapy. And in the data set it states whether the patient has done I-131 treatment in the past or not.
- Query Hypothyroid - In data set it states whether the patient is having hypothyroidism or not.
- Query Hyperthyroid - In the data set it states whether the patient is having hyperthyroidism or not.
- Goitre - A goitre is an inflammation or bulge that develops in the upper part of the throat when the thyroid becomes excessively enlarged. It states whether the patient is having Goitre or not.
- Tumor - An aberrant accumulation of tissues which originates whenever cells multiply and divide excessively or fail to perish the way they ought to. In the data set it states whether the patient is having tumour or not.
- Hypopituitary - Hypopituitarism is when you don't have enough of one or more of the hormones produced by the pituitary gland. Lack of these hormones can affect many normal body processes, like growth, blood pressure, and reproduction.[7] In data set it signifies whether the patient is having hypopituitary or not.
- TSH - TSH is the abbreviation for "thyroid stimulating hormone." A blood test called a TSH test is used to measure this hormone. If your TSH number is too high or too low, it could mean that you have a thyroid problem.
- T3 - Triiodothyronine, or T3, is a hormone made by the thyroid. It is an important part of how the body controls metabolism, which is a group of processes that control how fast cells and tissues work.[10] To find out how much T3 is in your blood, a lab test can be done.
- T4 - The T4 test is done to check how well the thyroid is working. As part of a T4 test, two blood tests may be done: total T4, which measures the total amount of thyroxine in the blood, including how much is attached to blood proteins that help move the hormone through the body; and free T4, which measures how much of the hormone is not attached to blood proteins.[7]
- Category - it states which type of thyroid is having patient.

3.2. Data Analysis & Pre-processing

Before the datasets are loaded into the machine learning model, a variety of techniques are applied to enhance its efficacy.[5] Data normalization, encoding, handling missing values, and other pre-processing techniques are a few of them.

3.3. Handling the Missing Data

How to deal with lost data: Missing data are entries or numbers for one or more variables in a given dataset that were not collected or were not there. Missing numbers are a common problem in many real-world datasets.
When there are missing numbers, learning algorithms can get messed up,[3] or the accuracy of the model can go down. To make the model work better, the average value of each attribute was used to handle missing numbers.

Handling Missing Values

```
[14]: data['Age'].fillna((data['Age'].median()), inplace = True)
      data['TSH'].fillna((data['TSH'].median()), inplace = True)
      data['T3'].fillna((data['T3'].median()), inplace = True)
      data['TT4'].fillna((data['TT4'].median()), inplace = True)
      data['T4U'].fillna((data['T4U'].median()), inplace = True)
      data['FTI'].fillna((data['FTI'].median()), inplace = True)
```

Figure 1. Handling the missing data

3.4. Outliers Removal

An "outlier" is a piece of information or an item that is very different from the other, "normal" pieces.[10] They could be caused by mistakes in measuring or doing the job. Outlier mining is the name for the research that is done to find outliers. There are many ways to find outliers, and the deletion process for the Panda's data frame is the same as for the Panda's data frame itself.[18] The same method can be used to find outliers in lists and series-type items when analyzing data for real-world projects. The IQR (Inter-quartile Range) method is used to get rid of these outliers in this job.

3.5. Label Encoding

Working with datasets that have more than one label in one or more columns is a common job in machine learning. Label encoding is the name for this process. You can talk about or write down these identities.[6] The labels on the training tools are often written in English so that people can understand them. Machines unlike us humans cannot fathom the labels as set by us. The necessity of the conversion from human readable phrase to computer understandable numerals is referred to as label encoding. It is a vital phase in the supervised learning methodologies that comes before handling the structured dataset.

3.6. Prediction Model Construction

To build the prediction model, 80% of the preprocessed information was used for training, and the other 20% was used for testing.[14] Machine learning methods such as K Nearest Neighbors, Decision Tree, Support Vector Machine and Random Forest are used to build the prediction model.

3.7. Model Evaluation, Comparison and Saving

Model review, comparison and saving: At this point, several factors have been used to compare how accurate each model is.[9] The best algorithm model, which is the one that is the most accurate, is kept and used to make web applications.

4. Machine Learning Algorithms used for Prediction

We looked at numerous articles and previous works that used machine learning techniques, and we chose each of the following algorithms for model training since they appear to be among the most precise and effective ones: Random Forest, Decision Tree, KNN and SVM. In this section, we'll go over all of the different machine learning methods used in the prediction model.

4.1. Random Forest

A type of ensemble learning technique called "Random Forest" uses decision trees to carry out regression and classification assignments. This method involves creating numerous decision trees during the learning period and calculating the group's yield by aggregating the performance of each tree. This method is reliable and effective, especially in situations where non-linear interactions are necessary. This approach does have some drawbacks, though, like the challenges involved in comprehending the findings, the possibility of excessive fitting, and the requirement to choose the right amount of decision trees to incorporate in the framework. How it works is as follows:

Data Preparation: The data set that we have acquired and that is to be fed to the model as input is to be initially broken down into two subsets of training subset accompanied by the testing subset. As the name suggests, the utilization of the training one is for the preparation and learning purposes of the predictive model. On the other hand, the testing one becomes handy for validation purposes to check the reliability of the model.

Choosing Features: For each decision tree, Random Forest picks a collection of features from the dataset at random. This process helps to reduce the link between individual trees and makes the model less likely to "overfit."

Bootstrapping: Random Forest makes multiple bootstrap samples from the training set by picking data at random with replacement. This process makes sure that each decision tree gets a different set of training data, which helps make a variety of models.

Construction of the Decision Tree: For each bootstrap sample, a decision tree is built using a chosen criterion (such as entropy or the Gini index) to find the best way to split at each point.[18] For each group of features, the process of making a decision tree is done again.

Voting: In this step, the results from each decision tree are added together. In the case of classification, the final answer is based on what most of the trees say. The average of all results is used for regression.

Performance Evaluation: The test set is used to figure out how well the Random Forest model works.[20] The real values in the validation set are used to check how well the model works.

Importance of use of Random Forest

a. Robustness: Algorithm is quite capable of dealing with noisy data and data that is missing.
b. Accuracy: Random Forest can be very accurate by putting together the plans of many decision trees.
c. Scalability: Random Forest can work with big datasets that have a lot of traits.
d. Importance of Features: Random Forest gives a way to measure how important a feature is. This can be used to choose the most important features for the model.
e. The employment of this algorithm is expanded through the varsity of fields, like banking, medicine, and image processing.[12] Many data scientists use it as their go-to method because it is accurate and reliable.

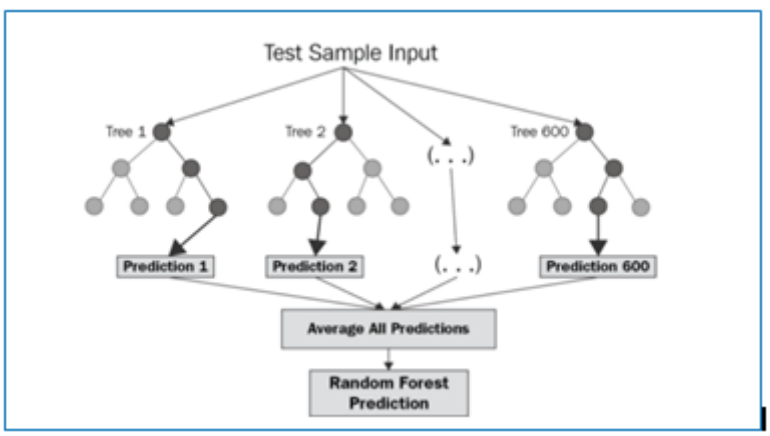

Figure 2. Random Forest

4.2. Decision Tree

A modelling method for classification or regression tasks is the decision tree. Recursively dividing a dataset into smaller, more focused subsets produces an integrated tree structure with decision and leaf nodes using this method. Decision nodes show an option between two or more potential values for a test attribute, such as "Sunny," "Dreary," and "Stormy" for the variable "Outlook." The outcome or prediction is represented by leaf nodes, for example, "Hours Played." The most significant predictor is thought to be the base node, which is the highest decision node. Decision trees are more than efficient to deal with

numeric as well as categoric information.

Figure 3. Decision Tree

4.3. Support Vector Machine

Being one of the most prevalent and appreciated supervised learning algorithms, SVM may be used to solve both classification and regression issues, however it is most frequently employed to solve issues related to classification. The construction of Hyperplanes is SVM's main goal. By orchestrating the partition of n-dimensional space into classes, the boundary decision aids in classification.

Finding the region of the hyperplane which optimally divides observations from various categories in a particular attribute space is the basic goal of SVM. SVMs can use kernel capabilities to alter what was originally collected towards a space of greater dimensions whereby a hyperplane can be identified to accurately categorize the data when its linear hyperplane is unable to sufficiently differentiate the data.

SVMs' strong generalizability on brand-new, untested information represents one of its primary benefits. To do this, choose the portion of the hyperplane that optimizes the distance across every group's nearest point of information and the hyperplane itself. Tolerance guarantees that the machine learning algorithm operates well on information that has never been seen before and is resilient to disruptive information points that arise. SVMs, which are being utilized successfully in a variety of areas, including biological information technology, text categorization, and picture recognition. Nevertheless, training SVMs on enormous data sets can be computationally prohibitive and choosing the right kernel-level functions might be tricky. SVMs are better than other machine learning methods in a number of ways. They work well with high-dimensional data because they can find the best hyperplane even in many-dimensional areas. They are also strong against overfitting, especially when the C-parameter is set correctly. Also, they are easy to program because they only need a small part of the training data to find the best hyperplane. SVMs do have some problems, though. They can be affected by the kernel function and its hyperparameters, which can change how well the model works.[16] When working with big datasets or kernels that are hard to understand, they can also be hard to compute.

4.4. K Nearest Neighbors

KNN, additionally referred to as the K Nearest Neighbour, constitutes one of the several supervised learning strategies in use. It is also among the most straightforward and understandable algorithms. The fundamental premise of this strategy is to look for commonalities amongst the current set of data and the particular case currently being worked on. The newly created instance is assigned to the cluster with which it has the greatest degree of similarity. One stage in KNN is to calculate the Euclidean connecting the newly acquired data point compared to previously collected points. Then, between these K nearest

data points, the category with the maximum number of neighbors will be used to classify the current instance. For instance, the cluster of data points belonging to the cat category will be the closest to the current instance if a photograph of an animal (say, a cat) is to be categorized as either a cat or a dog.

The KNN algorithm is an uncorrelated, instance-based method. This indicates that it doesn't assume any particular data dispersion as well as that it keeps the complete set of training information in memory rather than building a model. KNN has a simple configuration process and performs well on petite and medium-sized collections. Nonetheless, it can be highly computationally costly and vulnerable to the scalability curse when confronted with enormous data sets or feature spaces that are extremely dense [4].

5. Results and Discussion

Now we will look at the results that we obtained using various types of machine learning algorithms. Master data pre-processing and data cleaning really moved '?' As it is there as a nan value. After that we have and did missing values by using median in place of missing values after this, we handled nominal data then we have balance out the data which was imbalanced earlier using imblearn module.

After this we split the data into 2 parts which is testing is off 20% and the training data is 80%. We have compared 4 type of algorithms which are decision tree and its accuracy is 91.5% on training data set and 89.3% on testing data set, in support vector machine algorithm the accuracy is 61.11% on training dataset and 60.3% on test data set, in KNN algorithm we have obtained the accuracy of 87% on training data set and 83.9% on testing dataset, now in last algorithm which is random forest the accuracy is 91.5% on training dataset and 89.8% on testing data set.

After completing all we were developments noting that render forest algorithms accuracy was best so we will go with random forest classifier for building a web application. The F1 score for random forest classifier was 90%. We have also done hyper parameter tuning with accuracy goes 90.03%, if we compare both hyper parameter tuning, and normal machine learning algorithm so best result is from random forest classifier so we will go with random forest classifier for web application.

Figure 4. Algorithm comparison

Figure 5. Hyperparameter Tuning

6. Conclusion

In conclusion, machine learning has shown promise in finding thyroid diseases. Algorithms like KNN, SVM, Decision Tree and Random Forest Classifier are used to correctly classify thyroid diseases from clinical data. These algorithms could be used to help doctors make decisions and diagnose conditions, making it easier and more accurate to find thyroid illnesses. But the success of these machine learning algorithms depends on a number of things, such as the quality and amount of data, the features chosen, and how well the model works across different groups and settings. More study is needed to make these models more accurate and reliable and to find out if they could be used in clinical settings. Overall, machine learning techniques can help find thyroid diseases early, which can have a big effect on how patients do and their health. With more work and testing, these algorithms could be used in clinical settings to find and diagnose thyroid cancer faster and more accurately.

References

[1]. Y.-H. Yang, C.-H. Tsai, H.-T. Hsu, and Y.-H. Chen, "*Application of machine learning algorithms in thyroid disease diagnosis*," Computer Methods and Programs in Biomedicine, vol. 139, pp. 53-63, 2017.

[2]. N. L. Brancati, S. Giani, S. Ferrari, A. Bazzocchi, G. Battista, and R. Maroldi, "*Thyroid disease classification using machine learning: a systematic review*," Journal of Digital Imaging, vol. 34, pp. 431-439, 2021.

[3]. D. Dey, P. Mitra, and S. Chakraborty, "*Prediction of thyroid disease using machine learning algorithms*," in 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), pp. 570-573, 2017.

[4]. S.K. Sharma, "*Performance Analysis of Reactive and Proactive Routing Protocols for Mobile Ad-hoc –Networks*", International Journal of Scientific Research in Network Security and Communication, Vol.**1,** No.**5**, pp.**1-4**, **2013**.

[5]. A. M. Salem, "*Classification of thyroid disease using machine learning techniques*," in 2017 IEEE 2nd International Conference on Control, Instrumentation, and Automation (ICCIA), pp. 267-271, 2017.

[6]. L. Chen, J. Li, J. Li, and J. Li, "*A comparative study of machine learning algorithms for thyroid disease diagnosis*," in 2018 37th Chinese Control Conference (CCC), pp. 2643-2648, 2018.

[7]. J. Shen, W. Yang, Y. Zhang, X. Wang, and Z. Liu, "*Intelligent diagnosis of thyroid disease using machine learning techniques,"* Journal of Medical Systems, vol. 42, p. 104, 2018.

[8]. N. T. Nguyen, M. Qiu, H. Tran, T. D. Nguyen, and T. C. Tran, "*Machine learning approaches for thyroid disease diagnosis: a review*," in 2019 International Conference on Communications, Management, and Telecommunications (ComManTel), pp. 329-334, 2019.

[9]. Y. Liu, Z. Zhang, Y. Wang, Y. Wang, and Q. Chen, "*Comparison of machine learning algorithms for thyroid disease diagnosis*," in 2019 International Conference on Electrical Engineering, Control and Robotics (EECR), pp. 85-89, 2019.

[10]. H. Zhang, L. Yang, and J. Wang, "*A thyroid disease diagnosis model based on machine learning*," in 2019 4th International Conference on Intelligent Transportation Engineering (ICITE), pp. 417-421, 2019.

[11]. L. He, J. Yang, Y. Zhang, and Y. He, "A machine learning-based approach to thyroid disease diagnosis," in 2020 IEEE International Conference on Information and Automation (ICIA), pp. 155-160, 2020.

[12]. S. Shrivastava, R. Sharma, and S. K. Sahoo, "*Thyroid disease diagnosis using machine learning algorithms*," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1-6, 2020.

[13]. L. Wang, J. Zhang, Y. Sun, and Y. Fu, "*A thyroid disease diagnosis method based on machine learning*," in 2020 5th International Conference on Robotics and Automation Sciences (ICRAS), pp. 571-575, 2020.

[14]. A. Kumar, A. Kumar, and S. Kumar, "*A hybrid deep learning approach for thyroid disease detection*," in 2020 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 169-173, 2020.

[15]. S. Mohd, S. R. Jaafar, and S. H. Ahmad, "*Thyroid disease detection using machine learning and optimization techniques: a review*," SN Applied Sciences, vol. 3, no. 4, p. 325, 2021.

[16]. Y. Wang, C. Liu, and J. Zhang, "*Thyroid disease detection based on machine learning and convolutional neural network*," in 2022 International Conference on Intelligent Robotics and Intelligent Systems (IRIS), pp. 159-164, 2022.

[17]. R. K. Gupta and P. Goyal, "*Thyroid disease detection using machine learning techniques: a comprehensive review*," in 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 216-220, 2022.