Ann Med Res

Current issue list available at AnnMedRes

Annals of Medical Research

journal page: www.annalsmedres.org

# The estimating of hypothyroidism with the bagged CART model based on clinical dataset and identify of risk factors

Bahri Evren[a], Seyma Yasar[b,*]

[a]Inonu University, Faculty of Medicine, Department of Endocrinology and Metabolism, Malatya, Türkiye
[b]Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

ARTICLE INFO

**Keywords:**
Bagged classification and regression trees (CART)
Machine learning
Hypothyroidism
Classification

Received: Oct 05, 2022
Accepted: Oct 13, 2022
Available Online: 22.10.2022

DOI:
10.5455/annalsmedres.2022.10.302

## Abstract

**Aim:** The purpose of this study is to use machine learning techniques, Bagged CART, to classify hypothyroidism, which typically results from insufficient thyroid hormone synthesis in the body or seldom affects target tissues, and to identify potential risk factors.

**Materials and Methods:** In this study, the open source data set obtained from the UCI database was used. The 10-fold cross-validation technique was used in the creation of the Bagged CART model from the Decision Tree Ensembles class to classify hypothyroid, and the performance criteria of this model were accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-Score, G-mean and Matthews Correlation Coefficient ( MCC) was given. Then, the significance of the variable was calculated through the model created and possible risk factors for hypothyroidism were determined.

**Results:** The accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-Score, G-mean and Matthews Correlation Coefficient (MCC) performance criteria for the model created for the classification of hypothyroidism were 99.9%, 99.2%, 98.3%, 100.0%, 100.0%, 99.9%, 99.2%, 99.9%, and 99.1%, respectively. According to the created XGBoost model, the three most important factors that could be associated with hypothyroidism were determined as TBG, TSH, T4U, TT4, age, FTI, Query hypothroid, on thyroxine, on antithyroid medication, thyroid surgery, sex, TBG measured, sick, T3 mesured, Query hyperthyroid, goitre.

**Conclusion:** In conclusion, considering the results of the machine learning model created in this study, the hypothyroidism classification performance was quite high and the significance of the variables and possible risk factors for hypothyroidism were determined. In the light of the findings, it is predicted that these risk factors may be useful in the clinic.



## Introduction

One of the most prevalent endocrine disorders, thyroid diseases are defined by symptoms that have a negative impact on the patients' quality of life and are typically caused by either excessive or insufficient hormone production. Besides, nodule or tumor growth might be a sign of thyroid illness. The under-secretion of thyroid hormones as a result of an underactive thyroid gland is known as hypothyroidism. Weakness, exhaustion, dryness and flaking of the skin, swelling of the tongue, and hoarseness in the voice may result from reduced hormone secretion into the blood since these hormones govern our metabolism [1]. Accurate diagnosis is crucial for the proper management of thyroid illnesses since the hormones generated by the thyroid gland have a significant impact on metabolic processes, tissue development and growth, the pace at which foods are used for energy, and the rate of live weight increase [2]. However, there are numerous varieties of thyroid disease, and certain details are missed, making it challenging for medical professionals to provide a precise diagnosis.

Over the years, one of the most promising and significant study topics has been the analysis of health data. Health-related machine learning applications are instruments created to decipher and analyze complex data and to extract valuable information. These tools are meant to help physicians and patients alike with patient diagnosis, treatment, and follow-up. Many studies have employed machine learning techniques, from drug discovery to anticipating the processes involved in disease diagnosis and therapy [3]. CART (Classification and Regression Trees) classification and regression trees were created by Breiman in 1984 (Gordon et al., 1984). The CART decision tree al-

*Corresponding author:
*Email address:* seyma.yasar@inonu.edu.tr (Seyma Yasar)

gorithm is created by dividing subsets of the dataset and using all attributes as predictions to create two subnodes for each repeated split, starting from the entire dataset. It uses the gini index to select the best node. The Gini index works by selecting a split at each node, such that for each root node there is more than just its parent node. Next, the estimators are compared and the estimator with the best result is selected for the next division. This situation is repeated until the stop rules are active [4]. One of the bagging techniques, Bagged CART effectively reduces prediction variance and greatly improves classification performance and overfitting [5].

In this study, Bagged CART algorithm, which is one of the machine learning techniques, was used to help clinicians diagnose hypothyroidism. In an open-access hypothyroidism dataset, the classification of hypothyroidism and the risk factors associated with the disease will be determined according to the order of importance and used in diagnosis, with the machine learning method applied to the data obtained from healthy and diseased samples.

## Materials and Methods

*Dataset*

The present research was performed as a retrospective case-control study. An open access data set consisting of demographic and clinical data of 3163 individuals (Hypothyroid=150, Negative=3012) taken from the UCI database used in this study was used (6). In the data set used, 907 (29.4%) men, 2182 (70.6%) women and the mean age of men are 51.4±18.9 years and the mean age of women is 51.1±19.5 (Table 1).

*Bagged Classification and Regression Trees (Bagged CART)*

The bagging strategy can significantly increase the accuracy of the CART, which is regarded as an unstable model [7]. The bagged CART significantly enhances classification performance and reduces overfitting while also successfully reducing prediction variance. First, CART divides training sample units recursively using a predetermined number of variables. In order to determine which binary split in a predictive variable is least likely to depart from the projected response variable, it then analyses all of the predictive variables. When homogenous final nodes are formed in a hierarchical tree, the procedure usually continues for each result obtained from the initial split. When cross validation produces the lowest error, CART prunes the trees to prevent overfitting [8, 9].

*Statistical analysis*

Qualitative variables in the data set were summarized by number (percentage) and quantitative variables by median (min-max). The conformity of the variables to the normal distribution was examined using the Shapiro-Wilk test. Mann-Whitney U test and Chi-Square test were used to investigate the differences of the variables in terms of target variable. A value of $p<0.05$ was considered statistically significant. SPSS 26.0 was used for statistical analysis [10].

**Table 1.** Explanatory information about the data set.

| Variable | Variable Description |
|---|---|
| Age | Integer |
| Sex | Male(M),Female(F) |
| On thyroxine | False(f),true(t) |
| Query on thyroxine | False(f),true(t) |
| On antythyroid | False(f),true(t) |
| Sick | False(f),true(t) |
| Pregnant | False(f),true(t) |
| Thyroid surgery | False(f),true(t) |
| T131 treatment | False(f),true(t) |
| Query Hypothyroid | False(f),true(t) |
| Query Hyperthyroid | False(f),true(t) |
| Lithium | False(f),true(t) |
| Goiter | False(f),true(t) |
| Tumor | False(f),true(t) |
| Hypopitutory | False(f),true(t) |
| Psych | False(f),true(t) |
| Tsh measured | False(f),true(t) |
| TSH | Real |
| T3 measured | False(f),true(t) |
| T3 | Real |
| TT4 measured | False(f),true(t) |
| TT4 | Real |
| T4U measured | False(f),true(t) |
| T4U | Real |
| FTI Measured | False(f),true(t) |
| FTI | Real |
| TBG Measured | False(f),true(t) |
| TBG | Real |
| Referal source | SVHC, other, SVI, STMW, SVHD |
| Class | negative, hypothyroid |

*Modelling and performance metrics*

Bagged CART machine learning model was established with 5-fold cross validation on the obtained data set. In the n-fold cross-validation technique, which is a method used to evaluate the success of machine learning models, the method is run k-times. At each step, $1/k$ of the dataset, which has not been used for testing before, is used for testing, while the rest is used for training [11]. Performance evaluation of the established model was evaluated with accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-Score, G-mean and Matthews Correlation Coefficient (MCC) [12].

## Results

The data set used in the modeling consists of demographic and clinical information of 150 hypothyroid and 3012 negative patients. Descriptive statistics for other variables in terms of dependent/target variable (class) are given in Table 2.

In the light of the information in Table 2, the difference between groups for Age, TSH, TT4, T4U, FTI, TBG, Thyroid surgery, Query hypothyroid, TSH measured, T3 measured, TT4 measured, T4U measured, FTI measured, TBG measured variables is statistically significant. The

**Table 2.** Descriptive statistics for output variables in terms of dependent/target variable (class).

| | | Class | | |
|---|---|---|---|---|
| | | Negative<br>Median (Min-Max) | Hypothyroid<br>Median (Min-Max) | p-value |
| Age | | 54(1-98) | 59(5-88) | 0.047 |
| TSH | | 0.6(0-200) | 36(0-530) | <0.001 |
| TT4 | | 1.8(0-8.11) | 0.9(0-6.7) | <0.001 |
| T4U | | 106(4-450) | 33.5(2-230) | <0.001 |
| FTI | | 0.96(0-2.21) | 1.03(0.61-1.79) | <0.001 |
| TBG | | 109(0-881) | 34.5(0-133) | <0.001 |
| | | Count (Percentage) | | |
| Sex | Male<br>Female | 870 (29.58%)<br>2071 (70.42%) | 37 (25.00%)<br>111 (75.00%) | 0.232 |
| On Thyroxine | False<br>True | 2565 (85.16%)<br>447 (14.84%) | 136 (90.67%)<br>14 (9.33%) | 0.081 |
| Query on thyroxine | False<br>True | 2957 (98.17%)<br>55 (1.83%) | 150 (100.00%)<br>0 (0.00%) | 0.111 |
| On antithyroid medication | False<br>True | 2971 (98.64%)<br>41 (1.36%) | 149 (99.33%)<br>1 (0.67%) | 0.721 |
| Thyroid surgery | False<br>True | 2918 (96.88%)<br>94 (3.12%) | 140 (93.33%)<br>10 (6.67%) | 0.030 |
| Query hypothyroid | False<br>True | 2791 (92.66%)<br>221 (7.34%) | 130 (86.67%)<br>20 (13.33%) | 0.011 |
| Query hyperthyroid | False<br>True | 2776 (92.16%)<br>236 (7.84%) | 143 (95.33%)<br>7 (4.67%) | 0.207 |
| Pregnant | False<br>True | 2950 (97,94%)<br>62 (2.06%) | 149 (99,33%)<br>1 (0.67%) | 0.368 |
| Sick | False<br>True | 2915 (96.78%)<br>97 (3.22%) | 148 (98.67%)<br>2 (1.33%) | 0.329 |
| Tumor | False<br>True | 2972 (98.67%)<br>40 (1.33%) | 150 (100.00%)<br>0 (0.00%) | 0.260 |
| Lithium | False<br>True | 3010 (99.93%)<br>2 (0.07%) | 150 (100.00%)<br>0 (0.00%) | 1.000 |
| Goitre | False<br>True | 2919 (96.91%)<br>93 (3.09%) | 144 (96.00%)<br>6 (4.00%) | 0.471 |
| TSH measured | False<br>True | 467 (15.50%)<br>2545 (84.50%) | 1 (0.67%)<br>149 (99.33%) | <0.001 |
| T3 measured | False<br>True | 681 (22,61%)<br>2331 (77.39%) | 14 (9,33%)<br>136 (90.67%) | <0.001 |
| TT4 measured | False<br>True | 249 (8.27%)<br>2763 (91.73%) | 0 (0.00%)<br>150 (100.00%) | <0.001 |
| T4U measured | False<br>True | 248 (8.23%)<br>2764 (91,77%) | 0 (0.00%)<br>150 (100.00%) | <0.001 |
| FTI measured | False<br>True | 247 (8.20%)<br>2765 (91.80%) | 0 (0.00%)<br>150 (100.00%) | <0.001 |
| TBG measured | False<br>True | 2755 (91.47%)<br>257 (8.53%) | 147 (98.00%)<br>3 (2.00%) | 0.002 |

**Table 3.** Performance metrics of the Bagged CART model.

| Metrics | Value | 95% Confidence Interval |
|---|---|---|
| Accuracy | 99.9% | 99.7-100 |
| Balanced Accuracy | 99.2% | 98.8-99.6 |
| Sensitivity | 98.3% | 94.2-99.8 |
| Specificity | 100.0% | 99.8-100 |
| Positive Predictive Value | 100.0% | 96.9-100 |
| Negative Predictive Value | 99.9% | 99.6-100 |
| F1-Score | 99.2% | 98.7-99.6 |
| MCC | 99.9% | 98.7-99.5 |
| G-Mean | 99.1% | 99.8-100 |

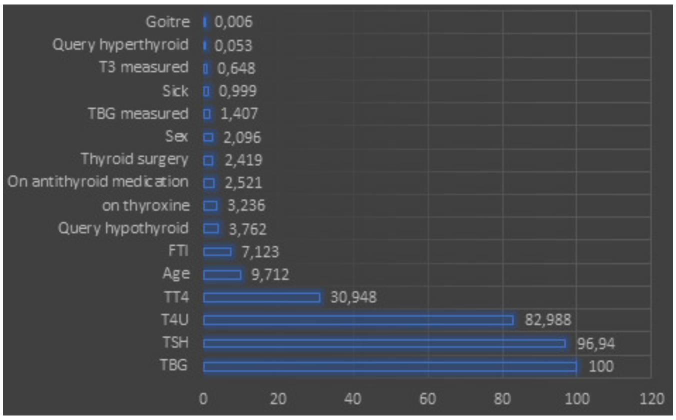


**Figure 1.** The importance values for possible risk factors.

performance criteria for the created Bagged CART model and the 95% confidence interval for these criteria are given in Table 3.

The importance values of variables related to possible risk factors for hypothyroidism determined by the model are shown in Figure 1.

## Discussion

A frequent thyroid condition called hypothyroidism is characterized by tissue-level thyroid hormone shortage (rarely ineffectiveness) and worsens with metabolic slowing [13]. The patient should not undergo any further tests in order to reduce their level of fatigue and frazzle them during the disease's diagnosis and treatment, which has a big impact on people's quality of life. For this reason, a diagnosis/diagnosis model will be developed in this study to aid physicians in both diagnosing hypothyroidism and minimizing the patient's exposure to tests and examinations along the course of this disease. Studies are conducted to improve the quality of human life in addition to the regular application of machine learning techniques in the field of health. The final point demonstrates how computer-aided diagnostic and follow-up systems employed in illness diagnosis are acquiring more and more importance day by day [14].

With the development of technology, large data sets in the field of health can be made understandable and interpretable with machine learning methods. Putting the medical definition is a very critical process for specialists, which requires accurate patient data, accurate medical literature knowledge and clinical experience. The application of machine learning methods directly to the datasets with the class imbalance problem has very important effects on the classification performance [15].

In the literature, there are many studies on the classification of hypothyroidism with machine learning methods. In one study, Random Forest (RF), Support Vector Machine (SVM), and K-Neaest Neighbors (KNN) classifiers were compared using a thyroid disease dataset consisting of a total of 7200 samples with 3 classes as hyperthyroidism, hypothyroidism and normal subject taken from the UCI machine. The overall classification accuracy of RF, SVM and K-NN was found to be 98.50%, 97.02% and 95.81%, respectively. The result shows that the RF classifier performance is better than other machine learning methods for diagnosing thyroid disease using the UCI dataset [16].

In another study, a 3-class hypothyroid disease dataset was used, with the output variable being compensated hypothyroidism, primary hypothyroidism, and negative. A set of machine learning algorithms such as support vector machine (SVM), Nave Bayes, decision trees and ensemble have been applied to predict the disease. The decision tree algorithm showed the highest performance with 97.6 percent accuracy [17].

There are other studies conducted with the data set used in this study. One of them, Akgül and Göksu et al. [18], using the same data, determined whether the new samples had hypothyroidism. In order to eliminate the uneven distribution, different sampling approaches were applied in data collection and hypothyroidism models were created using Logistic Regression, K Nearest Neighbor and Support Vector Machine classifiers. In this context, the logistic regression model gave the best results for this study, which was trained using oversampling approaches on the data set. The precision for this model was 97.8%, the F-Score was 82.26 percent, the area under the curve was 93.2 percent, and the Matthews correlation coefficient was 81.8 percent. Another study aimed to classify hypothyroidism by applying the Extreme Learning Machine model, one of the artificial neural network models, on the same data set. The Extreme Learning Machine model, one of the artificial neural network models, was used to classify hypothyroidism. The accuracy obtained from the model was calculated as 0.922, balanced accuracy 0.523, sensitivity 1, positive predictive value 0.922, negative predictive value 1, and F1-score 0.959 [19].

In this study, machine learning was applied to an open source hyproid dataset and the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-Score, G-mean and Matthews Correlation Coefficient (MCC) values obtained from the model were 0.987, 0.987, 0.980, 0.994, 0.994, 0.980, 0.987, 0.975 and 0.987, respectively. Considering these metrics, the model created is quite successful in classifying hypothyroidism. In addition, the most important 3 factors that may be associated with hypothyroidism were determined as T4U, TBG and TSH.

## Conclusion

In conclusion, considering the results of the machine learning model created in this study, it has a very good classification performance of hypothyroidism. On the other hand, possible risk factors for hypothyroidism were determined with the significance of the variables obtained as a result of the model. Therefore, in the light of the findings, it is predicted that these risk factors may be beneficial in the clinic.

*Ethics approval*

Since the data is open source, there is no need to obtain an ethics committee.

## References

1. Topaloğlu O, ÇAKIR B. Tedaviye Dirençli Hipotiroidi. Ankara Medical Journal. 2018;18(2):235-41.
2. Yılancıoğlu K. Vocal Cord Measures Based Artificial Neural Network Approach for Prediction of Parkinson's Disease Status. SDU Journal of Health Science Institute/SDÜ Saglik Bilimleri Enstitüsü Dergisi. 2017(1).
3. Lai K, Twine N, O'brien A, Guo Y, Bauer D. Artificial intelligence and machine learning in bioinformatics. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. 2018;1:3.
4. Steinberg D. CART: classification and regression trees. The top ten algorithms in data mining: Chapman and Hall/CRC; 2009. p. 193-216.
5. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees: Routledge; 2017.
6. Blake C. UCI repository of machine learning databases. http://www ics uci edu/~ mlearn/MLRepository html. 1998.
7. Murphree DH, Arabmakki E, Ngufor C, Storlie CB, McCoy RG. Stacked classifiers for individualized prediction of glycemic control following initiation of metformin therapy in type 2 diabetes. Computers in biology and medicine. 2018;103:109-15.
8. Duan H, Deng Z, Deng F, Wang D. Assessment of groundwater potential based on multicriteria decision making model and decision tree algorithms. Mathematical Problems in Engineering. 2016;2016.
9. Hamze-Ziabari S, Bakhshpoori T. Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5´ and CART algorithms. Applied Soft Computing. 2018;68:147-61.
10. George D, Mallery P. IBM SPSS statistics 26 step by step: A simple guide and reference: Routledge; 2019.
11. Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S, editors. The 'K'in K-fold cross validation. 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN); 2012: i6doc. com publ.
12. Yasar S, Yagin FH, Arslan AK, Colak C, Yologlu S. Interactive web-based software for evaluating diagnostic tests and roc curve analyses in health sciences. 2021.
13. Gaitonde DY, Rowley KD, Sweeney LB. Hypothyroidism: an update. South African Family Practice. 2012;54(5):384-90.
14. Umut K, YILMAZ A, Dikmen Y. Sağlık alanında kullanılan derin öğrenme yöntemleri. Avrupa Bilim ve Teknoloji Dergisi. 2019(16):792-808.
15. Zhou Z-H. Machine learning: Springer Nature; 2021.
16. Shahid AH, Singh MP, Raj RK, Suman R, Jawaid D, Alam M, editors. A Study on Label TSH, T3, T4U, TT4, FTI in Hyperthyroidism and Hypothyroidism using Machine Learning Techniques. 2019 International Conference on Communication and Electronics Systems (ICCES); 2019: IEEE.
17. Almahshi HM, Almasri EA, Alquran H, Mustafa WA, Alkhayyat A, editors. Hypothyroidism Prediction and Detection Using Machine Learning. 2022 5th International Conference on Engineering Technology and its Applications (IICETA); 2022: IEEE.
18. Akgül G, Çelik AA, Aydın ZE, Öztürk ZK. Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı. Bilişim Teknolojileri Dergisi. 2020;13(3):255-68.
19. Çiçek İB, Küçükakçalı Z. Classification of hypothyroid disease with extreme learning machine model. The Journal of Cognitive Systems. 2020;5(2):64-8.