Middle Black Sea Journal of Health Science / Mid Blac Sea J Health Sci
August 2024; 10(3): 255-265

ORGINAL ARTICLE
DOI: 10.19127/mbsjohs.1498383

Determination of Possible Biomarkers for Predicting Well-Differentiated Thyroid Cancer Recurrence by Different Ensemble Machine Learning Methods

Şeyma Yaşar[1](ID)

[1]Inönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Received: 09 June 2024, Accepted: 25 August 2024, Published online: 31 August 2024

Abstract

Objective: Well-differentiated thyroid cancer (WDTC) is the most common thyroid malignancy and although it is curable, the risk of recurrence is high. In this study, classification algorithms based on clinicopathologic features of WDTC patients were used to determine the possible of recurrence in WDTC and to evaluate potential predictive factors, and possible biomarkers based on the optimal model were identified.

Method: In this study, open access data on 383 patients with WDTC, 108 with recurrence and 275 without recurrence, were used. In order to predict recurrence in WDTC patients, features were selected using recursive feature elimination variable selection method among features and classification was performed with two ensemble learning methods (Random Forest, Adaboost).

Results: Two different ensemble learning models used to classify recurrence in WDTC were Random Forest with an accuracy of 0.957, sensitivity of 0.889, specificity of 0.978, positive predictive value of 0.923, negative predictive value of 0.967, Matthews correlation coefficient of 0.878, G-mean of 0.945, F1-score of 0.906, and accuracy of 0.940, sensitivity of 0.889, specificity of 0.955, positive predictive value of 0.857, negative predictive value of 0.966, Matthews correlation coefficient of 0.833, G-mean of 0.910, F1-score of 0.873.

Conclusion: According to variable importance based on the Random Forest, the 5 possible clinical biomarkers for predicting WDTC recurrence are Response, Risk, Node, Tumor, and age. In the light of these findings, patient management and treatment planning can be organized.

Keyword: Well-Differentiated, Thyroid Cancer, Recurrence, Random Forest, Adaboost

Suggested Citation Yasar Ş. Determination of possible biomarkers for predicting well-differentiated thyroid cancer recurrence by different ensemble machine learning methods. Mid Blac Sea Journal of Health Sci, 2024;10(3):255-265.

Address for correspondence/reprints:

Şeyma Yaşar

Telephone number: +90 (422) 341 06 60/13 21

E-mail: seyma.yasar@inonu.edu.tr

INTRODUCTION

Thyroid cancer is a type of cancer that starts in the thyroid gland and usually grows slowly. The thyroid gland is a gland located in the lower part of the neck, in front of the larynx. The function of the thyroid gland is to produce hormones that regulate metabolism. Symptoms of thyroid cancer are usually noticed in a person with a thyroid nodule. These symptoms may include a lump or mass in the neck, difficulty swallowing, hoarseness, difficulty breathing, and sometimes pain in the neck or ear (1). Thyroid cancer is a gender-neutral cancer that has increased in recent years with the widespread use of ultrasonography and US-guided fine needle biopsies, and according to GLOBOCAN 2023 data, it has a rate of 1% among all other cancers. Although thyroid cancer can be seen in any gender, approximately 75% of thyroid cancer patients are women and it is the seventh most common cancer type in women (2). Well-differentiated thyroid cancer represents an important subtype among thyroid cancers. This type of thyroid cancer usually refers to a condition called "well-differentiated", in which cells in the thyroid gland behave and grow in a similar way to normal thyroid cells. Well-differentiated thyroid cancers can be divided into subtypes such as papillary thyroid cancer and follicular thyroid cancer. Well-differentiated thyroid cancer is usually a slow-growing cancer and usually has a better prognosis than other types of thyroid cancer. If detected in the early stages and treated appropriately, cure rates are quite high. This type of cancer is usually treated with surgery. Surgery to remove the thyroid gland and surrounding tissues (thyroidectomy) is a common treatment method. In some cases, radioactive iodine therapy or hormone therapy may be required after surgery. Although this cancer tends to have a better prognosis than other types of thyroid cancer, regular follow-up after treatment is important because there is always a chance that the cancer may recur (3, 4).

Machine learning algorithms, which play a crucial role in disease classification, can help accurately identify and classify diseases by learning complex patterns and relationships from large amounts of data. Machine learning algorithms can classify/predict disease on new data using inputs such as a patient's symptoms, test results or imaging findings. Therefore, machine learning methods can predict which disease a patient has or which disease risk he/she is at with some results about patients. On the other hand, machine learning algorithms can help doctors diagnose diseases and create treatment plans. For example, deep learning algorithms can be used to detect cancer or identify disease symptoms using imaging techniques (e.g. MRI or CT scans) (5). By using ensemble learning methods, which are machine learning techniques that aim to create a stronger model by combining multiple learning algorithms, and weak learners that usually work together, a stronger performance can be achieved when they come together, even though each of them does not have high performance on its own. Ensemble learning methods are often used to reduce overfitting, increase accuracy, and create more generalizable models (6, 7).

In this study, using an open-source dataset of well-differentiated thyroid cancer patients with different clinicopathologic features, different ensemble learning models (Adaboost, Random Forest) that can predict recurrence in these patients were created and risk factors that may be associated with well-differentiated thyroid cancer recurrence were identified according to the optimal model (Random Forest) result.

METHODS

Dataset and Data Preprocessing

The dataset used in the study is an open source dataset containing 13 different clinicopathological features of well-differentiated thyroid cancer patients and published at https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence (8). Ethical approval for this study was received from Inonu University Health Sciences Non-Interventional Clinical Research Ethics Committee (approval number: 2024/5931). Detailed information about the 13 features in the dataset is given in Table 1.

Table 1. Explanations about 13 different features used in the study

Features	Features Type
Age	Represents the age of individuals in the dataset.
Gender	Indicates the gender of individuals (e.g., Male or Female).
Smoking	Possibly an attribute related to smoking behavior. The specific values or categories would need further exploration.
Smoking History	Indicates whether individuals have a history of smoking
Radiotherapy History	Indicates whether individuals have a history of radiotherapy treatment
Thyroid Function	Possibly indicates the status or function of the thyroid gland
Physical Examination	Describes the results of a physical examination, likely related to the thyroid
Adenopathy	Indicates the presence and location of adenopathy (enlarged lymph nodes)
Types of Thyroid Cancer (Pathology)	Describes the types of thyroid cancer based on pathology examinations, including specific subtypes like "Micropapillary Papillary," "Follicular," and "Hürthle cell."
Focality	Indicates whether the thyroid cancer is unifocal or multifocal
Risk	Represents the risk category associated with thyroid cancer
Tumor	Represents the T (Tumor) stage of thyroid cancer, indicating the size and extent of the primary tumor.
Lymph Nodes	Represents the N (Node) stage of thyroid cancer, indicating the involvement of nearby lymph nodes.
Cancer Metastasis	Represents the M (Metastasis) stage of thyroid cancer, indicating whether the cancer has spread to distant organs
Stage	Represents the overall stage of thyroid cancer based on the combination of T, N, and M stages
Treatment Response	Describes the response to treatment, including categories such as “Indeterminate”, “Excellent”, “Structural Incomplete” and “Biochemical Incomplete”.
Recurred	Indicates whether thyroid cancer has recurred.

Data Collection

Variable selection, a fundamental step for building more effective and generalizable models, is critical to improve the performance of the machine learning model, prevent overfitting, reduce computational costs, increase understandability, and improve data preprocessing. In the current study, the recursive feature elimination (RFE) variable selection method was applied. In the RFE method, it determines the subset of predictions required for an accurate model by eliminating the predictors backwards according to their order of importance. Predictors are ranked in order of importance, and the least important ones are removed in order (9). As the last stage of data preprocessing, the holdout method, one of the simplest methods of cross-validation, was applied and the data set was randomly divided into 70% training set and 30% testing set. The Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive predictive value, Negative predictive value, Matthews correlation coefficient (MCC), G-mean and F1-Score metrics in the performance evaluation of machine learning models. Formulas for performance metrics are given below.

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

Sensitivity=TP/(TP+FN)

Specificity=TN/(TN+FP)

Balanced accuracy= (Sensitivity+ Specificity)/2

Positive predictive value=TP/(TP+FP)

Negative predictive value=TN/(TN+FN)

Matthews correlation coefficient (MCC) = (TP*TN)-(FP*FN)/((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))1/2

G-mean= (Sensitivity* Specificity)1/2

F1-Score=(2*Precision*Recall)/(Precision+Recall)

Random Forest

Random Forest is an ensemble learning method and is a model created by combining multiple decision trees. Each decision tree is trained with randomly selected features and a random sample. Random Forest is widely used for classification and regression problems and generally provides high accuracy and generalization ability. It captures the variability and complexity in the data set well while reducing overfitting. It is similar to the bagging technique, but each tree is trained with random samples and features (10).

Adaboost

AdaBoost, first proposed by Freund and Schapire in 1997, is one of the most used boosting algorithms (11). AdaBoost algorithm, which is considered the first boosting algorithm, has solved many of the practical difficulties of previous boosting algorithms.

For this reason, it is preferred over other boosting methods due to its features such as high prediction speed, low memory usage, and easy applicability. The working steps of the AdaBoost algorithm are based on the logic of creating a weak classifier from each feature and obtaining an ensemble from these weak classifiers. The decision limits of weak classifiers are found by taking the weighted average of negative and positive examples for each feature. A new strong classifier is created with the help of weak classifiers with the lowest error rate. Thus, the features of weak classifiers that are not included in the strong classifier are deleted (12).

Statistical analysis

Quantitative variables in the dataset were summarized as mean±standard deviation and qualitative variables were summarized as number (percentage). Since the age variable did not meet the assumption of normal distribution according to recurrence categories, the difference between groups was examined with the Mann-Whitney U test. The relationship between categorical variables such as Gender, Smoking, Smoking History, Radiotherapy History and the recurrence of thyroid cancer was examined with Fisher Exact, Continuity Correction and Pearson Chi-Square tests, as appropriate. A value of $p<0.05$ was considered statistically significant. IBM SPSS 26.0 and “Adabag” package for the R programming language were used to perform the analyses (13,14).

RESULTS

Descriptive statistics for variables in the data set by thyroid cancer recurrence categories are presented in Table 2.

According to the results of Table 2, there is a statistically significant difference between the recurrence of thyroid cancer and all other variables in the data set, except for the variable indicating the status/function of the thyroid gland. After RFE variable selection was applied to all other data variables except thyroid cancer recurrence, which was the target variable in the modeling, 10 variables were included in the model. The metrics regarding the training and testing performances of machine learning classification models (Random Forest, Adaboost) created with these variables are given in Table 3.

Considering the data in Table 3, Random Forest is the machine learning algorithm that makes the best classification with all performance metrics. The performance metrics for classifying thyroid cancer recurrence with the Random Forest machine learning algorithm are 0.957, 0.933, 0.889, 0.978, 0.923, 0.967, 0.878, 0.945, 0.906 for Accuracy, Balanced accuracy, Sensitivity, Specificity, PPV, NPV, MCC, G-mean, F1-Score, respectively. The importance of variables that play a role in the recurrence of thyroid cancer through the Random Forest algorithm is given in Figure 1. According to Random Forest machine learning, the five most important factors that play a role in the recurrence of thyroid cancer were found to be Response, Risk, Node, Tumor, and age.

Table 2. Descriptive statistics for variables in the data set by thyroid cancer recurrence categories

		Recurred		p-value
		No	Yes	
		Count	Count	
Gender	Female	246a (89.45%)	66b (61.11%)	**<0.001**
	Male	29a (10.55%)	42b (38.89%)	
Smoking	No	259a (94.18%)	75b (69.44%)	**<0.001**
	Yes	16a (5.82%)	33b (30.56%)	
Hx Smoking	No	261a (94.91%)	94b (87.04%)	**0.014**
	Yes	14a (5.09%)	14b (12.96%)	
Hx Radiothreapy	No	274a (99.64%)	102b (94.44%)	**0.002**
	Yes	1a (0.36%)	6b (5.56%)	
Thyroid Function	Clinical Hyperthyroidism	27a (9.82%)	5a (4.63%)	0.244
	Euthyroid	234a (85.09%)	98a (90.74%)	
	Subclinical Hyperthyroidism	14a (5.09%)	5a (4.63%)	
Physical Examination	Diffuse goiter	7a (2.55%)	0 (0.00%)	**0.011**
	Multinodular goiter	88a (32.00%)	52b (48.15%)	
	Normal	5a (1.82%)	2a (1.85%)	
	Single nodular goiter-left	63a (22.91%)	26a (24.07%)	
	Single nodular goiter-right	112a (40.73%)	28b (25.93%)	
Adenopathy	Bilateral	5a (1.82%)	27b (25.00%)	**<0.001**
	Extensive	0 (0.00%)	7a (6.48%)	
	Left	5a (1.82%)	12b (11.11%)	
	No	265a (96.36%)	62b (57.41%)	
Pathology	Follicular	16a (5.82%)	12a (11.11%)	**<0.001**
	Hurthel cell	14a (5.09%)	6a (5.56%)	
	Micropapillary	48a (17.45%)	0 (0.00%)	
	Papillary	197a (71.64%)	90b (83.33%)	
Focality	Multi-Focal	66a (24.00%)	70b (64.81%)	**<0.001**
	Uni-Focal	209a (76.00%)	38b (35.19%)	
Risk	High	0 (0.00%)	32a (29.63%)	**<0.001**
	Intermediate	38a (13.82%)	64b (59.26%)	
	Low	237a (86.18%)	12b (11.11%)	
Tumor	T1a	48a (17.45%)	1b (0.93%)	**<0.001**
	T1b	38a (13.82%)	5b (4.63%)	
	T2	131a (47.64%)	20b (18.52%)	
	T3a	55a (20.00%)	41b (37.96%)	
	T3b	2a (0.73%)	14b (12.96%)	
	T4a	1a (0.36%)	19b (17.59%)	
	T4b	0 (0.00%)	8a (7.41%)	
Node	N0	241a (87.64%)	27b (25.00%)	**<0.001**
	N1a	12a (4.36%)	10a (9.26%)	
	N1b	22a (8.00%)	71b (65.74%)	
Metastasis	M0	275 (100.00%)	90a (83.33%)	**<0.001**
	M1	0 (0.00%)	18a (16.67%)	
Stage	I	268a (97.45%)	65b (60.19%)	**<0.001**
	II	7a (2.55%)	25b (23.15%)	
	III	0 (0.00%)	4a (3.70%)	
	IVA	0 (0.00%)	3a (2.78%)	
	IVB	0 (0.00%)	11a (10.19%)	
Response	Biochemical Incomplete	12a (4.36%)	11b (10.19%)	**<0.001**
	Excellent	207a (75.27%)	1b (0.93%)	
	Indeterminate	54a (19.64%)	7b (6.48%)	
	Structural Incomplete	2a (0.73%)	89b (82.41%)	

Table 2. The performance metrics of Random Forest and Adaboost machine learning models with training and testing datasets

	Random Forest		Adaboost	
	Training	Testing	Training	Testing
	Value (95% CI)	Value (95% CI)	Value (95% CI)	Value (95% CI)
Accuracy	0.996(0.989-1.00)	0.957(0.920-0.994)	0.996(0.989-1.00)	0.940(0.896-0.983)
Balanced accuracy	0.994(0.984-1.00)	0.933(0.888-0.979)	0.997(0.991-1.00)	0.922(0.873-0.971)
Sensitivity	0.988(0.933-1.00)	0.889(0.708-0.976)	1.00(0.955-1.00)	0.889(0.708-0.976)
Specificity	1.00(0.98-1.00)	0.978(0.921-0.997)	0.995(0.97-1.00)	0.955(0.889-0.988)
PPV	1.00(0.955-1.00)	0.923(0.749-0.991)	0.988(0.934-1.00)	0.857(0.673-0.96)
NPV	0.995(0.971-1.00)	0.967(0.906-0.993)	1.00(0.98-1.00)	0.966(0.904-0.993)
MCC	0.991(0.98-1.00)	0.878(0.818-0.938)	0.991(0.98-1.00)	0.833(0.766-0.901)
G-mean	0.997(0.991-1.00)	0.945(0.903-0.986)	0.994(0.985-1.00)	0.91(0.858-0.962)
F1-Score	0.994(0.984-1.00)	0.906(0.852-0.959)	0.994(0.985-1.00)	0.873(0.812-0.933)

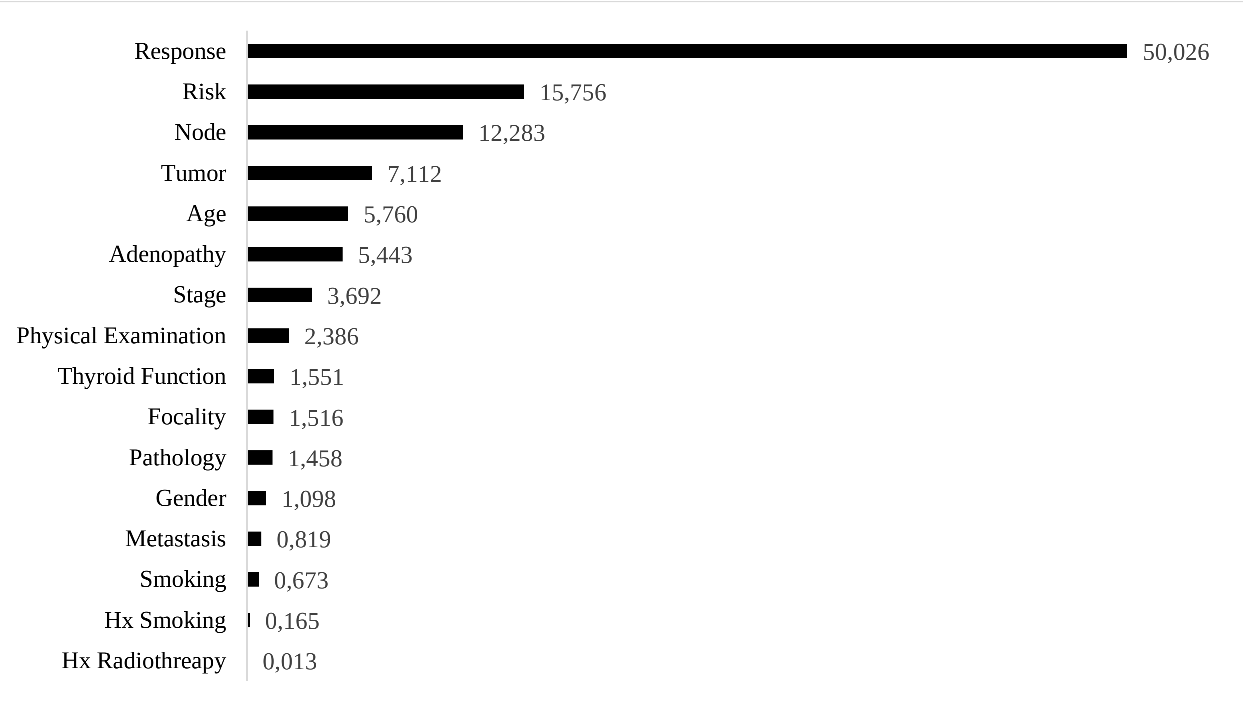

Figure 1. The importance of variables that play a role in the recurrence of thyroid cancer according to Random Forest machine learning

DISCUSSION

Thyroid cancer progresses with minimal symptoms that are difficult to diagnose. This may prevent patients from accessing early diagnosis and treatment and may lead to advanced stages. The primary treatment method for patients with well-differentiated thyroid carcinoma, which constitutes more than 90% of thyroid cancers, is surgery. Recurrent disease may present as a biochemical disease, nodal metastasis, residual disease, and distant metastasis without structural evidence (15,16).

When well-differentiated thyroid cancer is treated with surgery, successful results are usually achieved. However, the risk of relapse still exists. One reason for this is microscopic residues. The tumor may not have been completely removed during surgery or microscopic residues may have been left behind. These residues may grow over time, increasing the risk of recurrence. The another reason may be lymph node involvement. Well-differentiated thyroid cancer can sometimes spread to surrounding lymph nodes. These lymph nodes may not have been completely cleared during surgery, which may increase the risk of recurrence. The another is that the aggressiveness, size, and cellular characteristics of the tumor may affect the risk of recurrence. Especially large tumors and high-grade tumors have a higher risk of recurrence. The another reason is the response to post-treatment iodine therapy. Iodine therapy targets and can help destroy thyroid cancer cells. However, in some cases, cells may be resistant to this treatment and the risk of recurrence may increase. Therefore, there are many factors that create the risk of recurrence in this well-differentiated thyroid cancer. In this study, the recurrence risk was classified using different machine learning methods through data including pathological findings as well as clinical characteristics of well-differentiated thyroid cancer patients treated with surgery, and the most important risk factors were determined as a result of the optimal model according to the performance metrics obtained. The considering the performance ratings of two different (Random Forest, Adaboost) machine learning methods used in the study to classify well-differentiated thyroid cancer recurrence, the Random Forest algorithm is the best classifying model with 0.957, 0.933, 0.889, 0.978, 0.923, 0.967, 0.878, 0.945, 0.906 for accuracy, balanced accuracy, sensitivity, specificity, PPV, NPV, MCC, G-mean, F1-Score, respectively. These values reveal that the Random Forest model is highly successful in classifying well-differentiated thyroid cancer. In another study conducted with the same data set, thyroid cancer recurrence was classified using different machine learning models (17). In the study in question, SVM, K-nearest neighbors, Decision Tree, Random Forest and artificial neural networks algorithms were used for classification using the full data set. While the Random Forest algorithm has the highest sensitivity value (99.66%), the Decision Tree algorithm has the highest specificity value (100%). In another study using the same dataset, 23 different machine learning methods were used to classify well-differentiated thyroid cancer recurrence in two different scenarios, with and without feature selection (18). Random Forest machine learning method was one of the classification models created to predict thyroid cancer recurrence without using feature selection with 97% accuracy. In addition, 10 different machine learning models created using feature selection were very successful in predicting thyroid cancer recurrence. In addition, 10 different machine learning models created using feature selection showed very successful performance such as 100% accuracy in predicting thyroid cancer recurrence. The findings obtained in the study regarding machine learning models in classifying thyroid cancer recurrence are in line with the aforementioned studies in the literature. However, unlike other studies, this study also presents possible clinical biomarkers that can be used to predict thyroid cancer recurrence. According to the variable significance scores obtained based on the optimal model, the five factors most likely to be associated with well-differentiated thyroid cancer recurrence are response, risk, node, tumor, and age.

CONCLUSION

In conclusion, the results suggest that recurrence of well-differentiated thyroid cancer may be associated with factors such as age, microscopic remnants after surgery, lymph node involvement, tumor aggressiveness and size, and response to iodine therapy. In this study, different machine learning methods were used to classify the risk of recurrence of well-differentiated thyroid cancer. The Random Forest algorithm, which is the best performing model, achieved high success in determining the risk of recurrence. Furthermore, similar results were obtained in other studies, and this study provides potential guidance on the prediction of thyroid cancer recurrence using clinical markers. These results may contribute to improving patient management and treatment planning (such as adjusting treatment intensity and determining appropriate follow-up intervals).

Acknowledgements: The authors would like to appreciate the patients participation in this study.

Ethics Committee Approval: Approval for this study was obtained from the İnonu University Non-Interventional Research Ethics Committee (28/05/2024 tarih ve 2024/5931).

We state that the parents have given their written informed consent to be involved in the study, in accordance with the Declaration of Helsinki.

Peer-review: Externally peer-reviewed

Author Contributions: Concept: ŞY, Design: ŞY, Data Collection and Processing: ŞY, Analysis and Interpretation: ŞY, Writing: ŞY, ŞY.

Conflict of Interest: The authors declared no conflict of interest.

Financial Disclosure: The authors declared that this study has not received no financial support.

REFERENCES

1. Adaş G, Adaş M, Özülker F, Akçakaya A. Thyroid cancers. Okmeydanı Tıp J. 2012;28: 26-34.

2. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. Ca Cancer J Clin 2023;73:17-48.

3. Vaish R, Mahajan A, Sable N, Dusane R, Deshmukh A, Bal M, D'cruz AK. Role of computed tomography in the evaluation of regional metastasis in well-differentiated thyroid cancer. Frontiers in Radiology 2023;3: 1243000.

4. Chen DW, Lang BH, McLeod DS, Newbold K, Haymart MR. Thyroid cancer. The Lancet 2023;401:1531-1544.

5. Boina R, Ganage D, Chincholkar YD, Wagh S, Shah DU, Chinthamu N, et al. Enhancing Intelligence Diagnostic Accuracy Based on Machine Learning Disease Classification. IJISAE (2023);11:765-774.

6. Mienye ID, Sun Y, Wang Z. An improved ensemble learning approach for the prediction of heart disease risk. IMU 2020;20:100402.

7. Yu L, Wang S, Lai KK. Credit risk assessment with a multistage neural network ensemble learning approach. Expert Syst. Appl. 2008;34:1434-1444.

8. Borzooei S, Briganti G, Golparian M, Lechien JR, Tarokhian A. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. Eur Arch Otorhinolaryngol 2024;281:2095-2104.

9. Misra P, Yadav AS. Improving the classification accuracy using recursive feature elimination with cross-validation. Int. J. Emerg. Technol, 2020;11:659-665.

10. Rigatti SJ. Random forest. J Insur Med 2017;47:31-39.

11. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. JCSS 1997;55:119-139.

12. Bulut F. Determining Heart Attack Risk Ration Through AdaBoost/AdaBoost ile Kalp Krizi Risk Tespiti. CBUJOS 2016;12:459-472.

13. George D, Mallery P. IBM SPSS statistics 26 step by step: A simple guide and reference. Routledge, 2019.

14. Alfaro E, Gamez M, Garcia N. Adabag: An R package for classification with boosting and bagging. Journal of Statistical Software 2013; 54:1-35.

15. Filetti S, Durante C, Hartl D, Leboulleux S, Locati LD, Newbold K, et al. Thyroid cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2019;30:1856-1883.

16. Ito Y, Onoda N, Okamoto T. The revised clinical practice guidelines on the management of thyroid tumors by the Japan Associations of Endocrine Surgeons: Core questions and

recommendations for treatments of thyroid cancer. Endocr J 2020;67:669-717.

17. Borzooei, S., Briganti, G., Golparian, M., Lechien, J. R., & Tarokhian, A. (2024). Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. European Archives of Oto-Rhino-Laryngology, 281(4), 2095-2104..

18. Setiawan KE. Predicting recurrence in differentiated thyroid cancer: a comparative analysis of various machine learning models including ensemble methods with chi-squared feature selection. Commun Math Biol Neurosci 2024;55: 1-29.