MEDICAL RECORDS-International Medical Journal

Research Article

Explainable Machine Learning Models for Predicting Recurrence in Differentiated Thyroid Cancer

Ahmet Kadir Arslan, Cemil Colak

İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye

Abstract

Aim: Differentiated thyroid cancer (DTC) is a common type of cancer that originates in the thyroid gland. This study aimed to predict the recurrence of differentiated thyroid carcinoma, in patient with well-DTC, using explainable machine learning (XAI) models.
Material and Method: The study utilized a dataset from the UCI Machine Learning Repository, which included 383 patients and 13 candidate predictors. After a variable selection process using distance correlation, only four predictors (Response, Risk, T, and N) were retained for model building. Two XAI models, Fast Interpretable Greedy-Tree Sums (FIGS) and Explainable Boosting Machines (EBM), were employed.
Results: The EBM model slightly outperformed the FIGS model in terms of accuracy. The study found that the most influential predictors of Well-DTC recurrence were the response to DTC treatment, risk status according to the American Thyroid Association classification, tumor size (T), and lymph node metastasis (N).
Conclusion: In conclusion, this study successfully identified key risk factors for DTC recurrence using XAI models, providing interpretable insights for clinical decision-making and potential for personalized treatment strategies.

Keywords: Differentiated thyroid cancer, explainable machine learning, risk factors, explainable boosting machine, fast interpretable greedy-tree sums

INTRODUCTION

Thyroid cancer is a type of cancer that begins in the thyroid gland. The thyroid gland is a small gland located at the front of the neck that produces hormones that regulate metabolism (1,2). Thyroid cancer is increasingly common worldwide. The reasons for this increase are not fully known, but it is thought to be due to improvements in diagnostic methods and environmental factors. Thyroid cancer is more common in women than men and is usually diagnosed early-50s (3).

Differentiated Thyroid Cancer (DTC) is the general name for the types of cancer that develop in the thyroid gland. The most common types of thyroid cancer are papillary thyroid cancer and follicular thyroid cancer. DTC usually grows slowly and responds well to treatment when detected early (4,5). Well-DTC is a type of cancer that occurs in the thyroid gland and consists of well-differentiated cancer cells, meaning they look like normal thyroid cells. Well-DTC is the most common type of thyroid cancer and is usually slow-growing and responds well to treatment (6,7).

Explainable Machine Learning (XAI) is an approach that makes it easier to understand the decisions and predictions of machine learning models. Traditional machine learning models are often referred to as "black boxes" because their inner workings and decision-making processes can be difficult to understand. Thanks to XAI, revealing what factors the model relies on and how it reaches its conclusions. This is especially critical in sensitive fields such as medicine to understand and trust the treatment decisions of the models by clinicians (8,9).

It is observed that there is an increase in the number of studies in the literature where thyroid cancer is predicted/classified using XAI methods. In a study (10), 9 classical machine learning methods were considered together with 3 XAI tools (SHAP, Shapash, LIME). In the study where the XGBoost model gave the best performance, the outputs of the relevant model were interpreted with XAI tools and it was determined that the TSH hormone was the variable that

CITATION
Arslan AK, Colak C. Explainable Machine Learning Models for Predicting Recurrence in Differentiated Thyroid Cancer. Med Records. 2024;6(3):468-73. DOI:1037990/medr.1525801

Received: 31.07.2024 **Accepted:** 28.08.2024 **Published:** 16.09.2024

Corresponding Author: Ahmet Kadir Arslan, İnönü University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Türkiye
E-mail: arslan.ahmet@inonu.edu.tr

contributed the most to the model performance. In another study (11), the outputs of a rule-based machine learning model were interpreted with the SHAP XAI technique. The use of the rule-based machine learning model and the SHAP technique together in explaining the patterns in the data set resulted in a more reliable prediction of thyroid cancer.

This study aimed to identify candidate predictors of DTC recurrence using two XAI methods and to obtain explainable/interpretable results.

MATERIAL AND METHOD

Data Set

The data set analyzed in this study is titled "Differentiated Thyroid Cancer Recurrence" obtained from the UCI Machine Learning Repository (12) (https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence). In addition, this dataset was originally generated in the study (13). Since this study was conducted on a publicly available clinical data set, Ethics Committee approval is not required. The dataset consisted of 13 candidate predictors for predicting well-DTC recurrence and 383 patients. In this dataset, collected over a 15-year period, each patient was followed up for 10 years. The mean age of the study participants was 40.86±15.13 years. The gender distribution was 312 (81.5%) females and 71 (18.5%) males. The distribution of the response variable well-DTC recurrence status (referred to as "Recurred") was 275 (71.8%) "No" and 108 (28.2%) "Yes". A comprehensive overview of the variables is given in Figure 1.

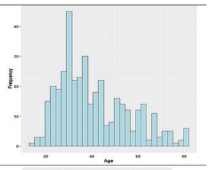
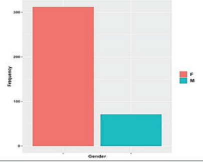
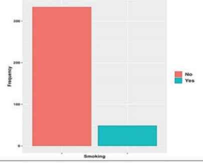
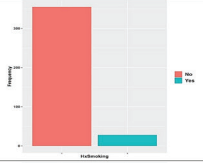
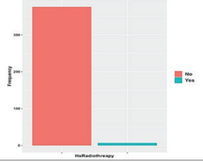
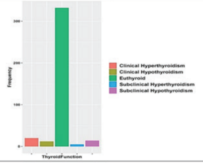
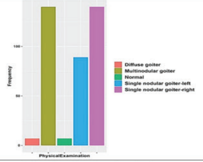
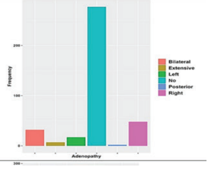
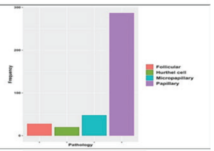
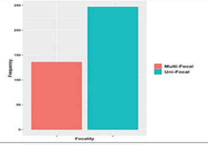
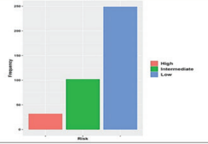
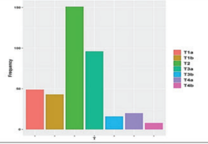
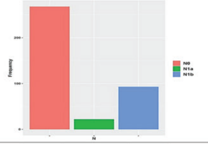
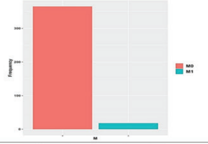
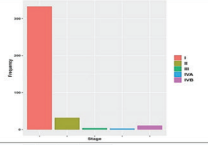
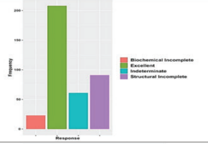
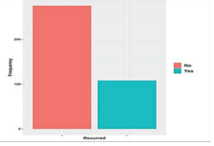
Features	Description of feature	Feature roles	Descriptive statistics	Distribution plots
Age	-	Predictor	N: 383 Mean: 40.87 Std. deviation: 15.13 Median: 37 Min: 15 Max: 82	
Gender	-	Predictor	F: 312 (%81) M: 71 (%19)	
Smoking	-	Predictor	No: 334 (%87) Yes: 49 (%13)	
HxSmoking	Smoking history	Predictor	No: 355 (%93) Yes: 28 (%7)	
IxRadiothreap	Radiation therapy history	Predictor	No: 376 (%98) Yes: 7 (%2)	
hyroidFunctio	-	Predictor	Clinical Hyperthyroidism: 20 (%5) Clinical Hypothyroidism: 12 (%3) Euthyroid: 332 (%87) Subclinical Hyperthyroidism: 5 (%1) Subclinical Hypothyroidism: 14 (%4)	
'sicalExamina	-	Predictor	Diffuse goiter: 7 (%2) Multinodular goiter: 140 (%37) Normal: 7 (%2) Single nodular goiter-left: 89 (%23) Single nodular goiter-right: 140 (%37)	
Adenopathy	-	Predictor	Bilateral: 32 (%8) Extensive: 7 (%2) Left: 17 (%4) No: 277 (%72) Posterior: 2 (%1) Right: 48 (%13)	
Pathology		Predictor	Follicular: 28 (%7) Hurthel cell: 20 (%5) Micropapillary: 48 (%13) Papillary: 287 (%75)	
Focality		Predictor	Multi-Focal: 136 (%36) Uni-Focal: 247 (%64)	
Risk		Predictor	High: 32 (%8) Intermediate: 102 (%27) Low: 249 (%65)	
T	TNM grading (Primary tumor)	Predictor	T1a: 49 (%13) T1b: 43 (%11) T2: 151 (%39) T3a: 96 (%25) T3b: 16 (%4) T4a: 20 (%5) T4b: 8 (%2)	
N	TNM grading (Regional lymph nodes)	Predictor	N0: 268 (%70) N1a: 22 (%6) N1b: 93 (%24)	
M	TNM grading (Distant metastasis)	Predictor	M0: 365 (%95) M1: 18 (%5)	
Stage	-	Predictor	I: 333 (%87) II: 32 (%8) III: 4 (%1) IVA: 3 (%1) IVB: 11 (%3)	
Response	-	Predictor	Biochemical Incomplete: 23 (%6) Excellent: 208 (%54) Indeterminate: 61 (%16) Structural Incomplete: 91 (%24)	
Recurred	-	Response	No: 275 (%72) Yes: 108 (%28)	

Figure 1. A comprehensive overview of the variables

Basic Statistical Analyses Phase

The variables considered in the study were summarized as frequency (percentage). Pearson chi-square tests were used to determine whether there was a statistically significant difference between the "Recurred" response variable groups. p≤0.05 was accepted as the statistical significance level.

Machine Learning Modeling Phase

Data preprocessing

In this study, the distance correlation-based variable selection method was applied to reduce model complexity and filter out variables that are not expected to contribute to the predictive performance of the machine learning models. Distance correlation is a statistical method used to measure the dependence relationship between two random variables. Unlike the classical Pearson correlation coefficient, it can detect not only linear relationships but also non-linear relationships. Thanks to this feature, it can reveal complex dependency structures between variables (14). This analysis, interpreted as a classical Pearson correlation coefficient, is applied sequentially between response and predictor variables. The cut-off value was set at 0.5 and variables with correlation values below this value were removed from the data set. In addition, to test validation of the predictive performance of machine learning models, the data set was randomly divided into two parts as training (80%) and test (20%) data sets. While the model training process was performed on the training data set, learning performance was evaluated on the test data set.

Machine learning models

Fast Interpretable Greedy-Tree Sums (FIGS)

FIGS is a tree-based model that generalizes Classification and Regression Trees (CART) to reduce bias and unexplained variance and aims to be both fast and interpretable. FIGS takes a greedy approach to the training process, building trees quickly. This helps compensate for the weaknesses of a single tree and constructs a more powerful and generalizable model. The generated trees have a simple and understandable structure to make it easier to understand the reasons for the model's predictions and bring transparency to decision-making processes. It can be used in both classification and regression problems, i.e. it is suitable for predicting both categorical and continuous features (15).

Explainable Boosting Machines (EBM)

Explainable Boosting Machines (EBM) is a machine learning model that offers both high prediction performance and the ability to explain the reasons for the model's decisions. It combines the power of traditional gradient boosting and generalized additive models with interpretability. EBM can model complex relationships using gradient boosting and achieve high prediction accuracy, making it suitable for a variety of classification and regression problems. EBM provides interpretability by visualizing and quantifying the impact of each feature on the prediction, thus making it easier to understand what factors the model's decisions are based on and bringing transparency to decision-making processes. It can work with different types of features (numeric, categorical, ordinal) and model various binary interactions, making it adaptable to different types of data. Among its advantages are its combination of high predictive performance and interpretability, its ability to work with different data types and features, its fast training process and scalability, and its use of various techniques to reduce the risk of overfitting. EBM is especially used in areas such as credit risk assessment, medical diagnosis, etc., where it is important to explain the reasons for the model's decisions (16,17).

Metrics for evaluating the predictive performance of the models

In this study, accuracy (ACC), the area under the receiver operating characteristics curve (AUC), F1-score (F1), logarithmic loss (Log-Loss), and Brier score (Brier) metrics were used to evaluate the binary classification performance of the models. When the model prediction performance is evaluated in the range of 0 to 1, ACC, AUC and F1 metrics with values of 1 or close to 1 indicate that the model has a high level of predictive performance, while Log-Loss and Brier metrics with values of 0 or close to 0 indicate that the model has a high level of predictive performance.

The Environments Where the Analyses were Performed

In this study, R (version 4.1.2) was used for statistical analysis, and Python (version 3.10.0) was used for machine learning modeling. Python language-based XAI library PiML (18) was used to construct the modeling workflow.

RESULTS

There were no missing values in the data set. Since all variables except age are categorical, no transformation method was applied to the data set. After applying the distance correlation-based variable selection algorithm, 4 of the 13 predictor variables (Response, Risk, T, and N) were selected. The findings of the related variable selection analysis for the top 10 variables are given in Figure 2.

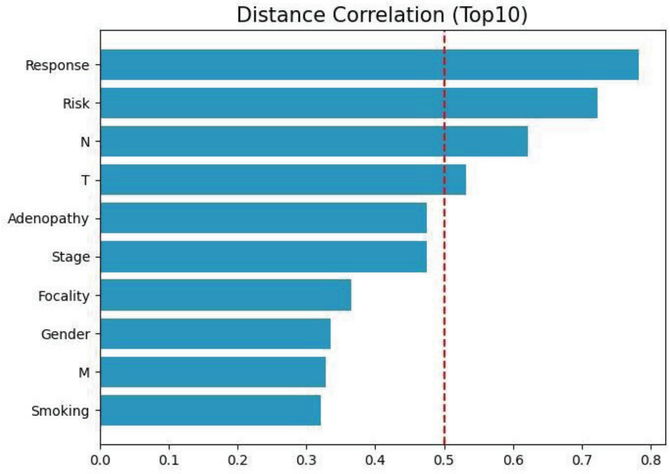


Figure 2. The findings of the related variable selection analysis for the top 10 variables

The inferential statistics results for the 4 variables obtained after the variable selection analysis are given in Table 1.

Table 1. Inferential statistics findings of Well-DTC groups in terms of predictor variables

Predictor		Well-DTC recurrence		Pearson chi square statistics	p
		No (n=275)	Yes (n=108)		
Risk	High	0 (0.0%)	32 (29.6%)	208.83	<0.001
	Intermediate	38 (13.8%)	64 (59.3%)		
	Low	237 (86.2%)	12 (11.1%)		
T	T1a	48 (17.5%)	1 (0.9%)	141.29	<0.001
	T1b	38 (13.8%)	5 (4.6%)		
	T2	131 (47.6%)	20 (18.5%)		
	T3a	55 (20.0%)	41 (38.0%)		
	T3b	2 (0.7%)	14 (13.0%)		
	T4a	1 (0.4%)	19 (17.6%)		
	T4b	0 (0.0%)	8 (7.4%)		
N	N0	241 (87.6%)	27 (25.0%)	153.19	<0.001
	N1a	12 (4.4%)	10 (9.3%)		
	N1b	22 (8.0%)	71 (65.7%)		
Response	Biochemical incomplete	12 (4.4%)	11 (10.2%)	309.47	<0.001
	Excellent	207 (75.3%)	1 (0.9%)		
	Indeterminate	54 (19.6%)	7 (6.5%)		
	Structural incomplete	2 (0.7%)	89 (82.4%)		

The overall accuracy (ACC) values obtained for both FIGS and EBM models as a result of training with Risk, T, N, and Response variables are presented in Figure 3. In addition, detailed classification performance metric values obtained from both training and test datasets for the two models are reported in Table 2.

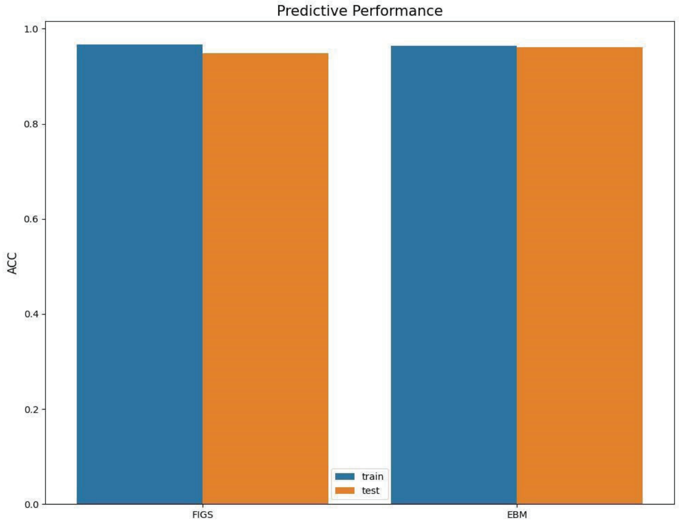


Figure 3. The overall accuracy (ACC) values obtained for both FIGS and EBM models

Table 2. Classification performance metrics for both EBM and FIGS models

Models	Data source	Performance metrics				
		ACC	AUC	F1	LogLoss	Brier
FIGS	Train	0.9673	0.9964	0.9444	0.1438	0.0283
	Test	0.9481	0.9964	0.9000	0.1534	0.0319
EBM	Train	0.9641	0.9922	0.9364	0.1015	0.0284
	Test	0.9610	0.9927	0.9189	0.0939	0.0268

Figure 4 shows the global effect importance levels of the single and binary interaction states of the variables obtained from the EBM model.

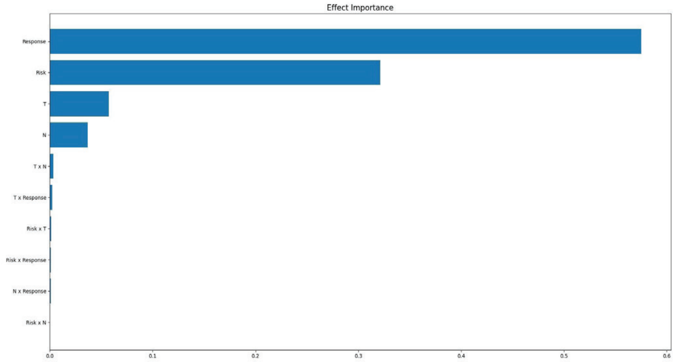


Figure 4. Global effect importance levels of the single and binary interaction states of the variables obtained from the EBM model

Figure 5 shows the local effect findings of the EBM model for two randomly selected patients with positive and negative recurrence labeling. Here,

- For the "Risk" variable, 0, 1, and 2 values indicate high, medium, and low risk respectively.
- For the "Response" variable, values 0, 1, 2, and 3 represent the categories Biochemical Incomplete, Excellent, Indeterminate, and Structural Incomplete, respectively.
- For the "T" variable, 0, 1, 2, 3, 4, 5, and 6 values indicate T1a, T1b, T2, T3a, T3b, T4a, and T4b categories, respectively.
- For the "N" variables values 0, 1, and 2 represent N0, N1a, and N1b categories, respectively.

The classification rules and uncalibrated recurrence risks based on the two samples shown in Figure 5 are presented in Table 3.

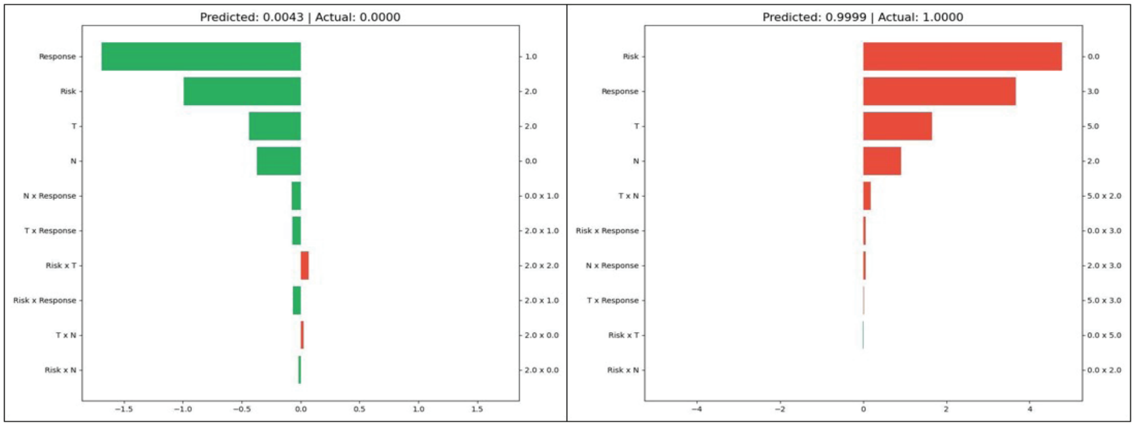


Figure 5. The local effect findings of the EBM model for two randomly selected patients with positive and negative recurrence labeling

Table 3. The classification rules obtained from the EBM model

Real case	Rules	Model prediction	Uncalibrated recurrence risk (%)
Recurrence positive	Risk = “High” & Response = “Structural Incomplete” & T= “T4a” & N = "N1b”	Recurrence positive	0.9999
Recurrence negative	Response = “Excellent” & Risk = “Low” & T = “T2” & N = “N0”	Recurrence negative	0.0043

DISCUSSION

In this study, we aimed to identify risk factors that can be used as decision support in the prediction of recurrence of well-differentiated thyroid carcinoma with XAI models such as FIGS and EBM, which have gained increasing popularity in recent years. The data set considered in the study consisted of 13 predictor variables in the first stage. This number decreased to 4 after the variable selection analysis. Statistically significant differences were observed between the replication groups in terms of the relevant variables (Table 1).

When the classification performances of the EBM and FIGS models are evaluated, it is observed that both models give similar results, but the EBM model performs slightly better than FIGS. Since the EBM model is the best-performing model, the prediction explanations of the relevant model have been considered. In fact, when the global effect importance graph in Figure 4 is evaluated, it is seen that the order of the predictors affecting the classification performance of the model is "Response", "Risk", "T" and "N". The variable pairs seen under the effect values of single variables and connected with the "x" symbol are interaction terms. As can be observed from Figure 4, the single effects of the relevant risk factors were more effective on the classification performance of the EBM model. It was observed that the effect of the binary interaction terms of the relevant variables on the classification performance remained weak.

When Figure 5 and Table 3 are examined together, it is observed that if the rule combination, Risk= “High” & Response= “Structural Incomplete” & T= “T4a” & N= ”N1b” the risk of well-DTC recurrence has a high probability of occurring.

In the current study, the greatest contribution to the risk of recurrence occurred in patients at high risk according to the American Thyroid Association (ATA) classification (19). In thyroid cancer, the term "structural incomplete" is often used in a pathology report and refers to the fact that microscopic examination after surgical removal of the thyroid gland shows that there is uncertainty as to whether the tumor was completely removed (20). The prevalence of structural incomplete is correlated with thyroid cancer risk stratification (21). The T4a classification is part of the TNM system for staging thyroid cancer (22). This classification indicates that the cancer is at an advanced stage and treatment options may be more limited. One study (23) also reported that tumor sizes over 4 cm are a risk factor for recurrence of follicular thyroid cancer, a subtype of DTC. The N0, N1a and N1b classifications for thyroid cancer are part of the TNM staging system, which indicates the spread of cancer to lymph nodes (metastasis). N1b classification indicates that the cancer is at a more advanced stage and treatment options may be more limited. One study concluded that nodal involvement in DTCs may increase the risk of recurrence (24).

Similarly, if the resulting classification rule is such, *Response= “Excellent” & Risk= “Low” & T= “T2” & N= “N0”* the risk of well-DTC recurrence has a very low probability. As expected, this suggests that the risk of well-DTC is very low in the presence of a good response to treatment, low recurrence, tumors smaller than 2 cm, and no cancer in regional lymph nodes.

When other machine learning-based studies using this dataset are evaluated, the support vector machine model (SVM) showed better classification performance than other classification models (sensitivity=0.99, specificity=0.97, and AUC=0.99) in the study carried out by Borzooei et al. (13). The results obtained are close to the findings of the present study, and it is a disadvantage for clinicians that the SVM model is not within the scope of explainable models such as the EBM and FIGS models considered in this study.

Therefore, the results obtained from the model are only related to classification performance. Moreover, the high classification performance obtained in our study was achieved with only 4 variables, which may suggest that modeling the remaining 9 variables is unnecessary.

In another study (25) dealing with the same dataset, after various preprocessing analyses, the dataset was modeled with the ensemble stacking algorithm and the related model showed a classification accuracy of 97%. This finding is less than the EBM and FIGS models considered in this study.

CONCLUSION

In this study, candidate risk factors that can be used to predict the risk of recurrence in patients with well-DTC were determined by XAI methods such as EBM and FIGS. According to the outputs obtained from the EBM model, which has a better classification performance, the response to DTC treatment, risk status, tumor size, and location, and the spread of cancer to nearby lymph nodes were determined as the most important risk factors for recurrence. This study has some limitations. The use of data obtained from a single center with relatively small sample size and the absence of an external cohort to increase the generalizability of the results are the main limitations of this study. As further research, it is recommended that researchers construct a meta-model using more XAI models together to obtain outputs with higher validity and reliability.

Financial disclosures: *The authors declared that this study has received no financial support.*

Conflict of interest: *The authors have no conflicts of interest to declare.*

Ethical approval: Since this study was conducted on a publicly available clinical data set, Ethics Committee approval is not required.

REFERENCES

1. Cabanillas ME, McFadden DG, Durante C. Thyroid cancer. The Lancet. 2016;388:2783-95.
2. Nguyen QT, Lee EJ, Huang MG, et al. Diagnosis and treatment of patients with thyroid cancer. Am Health Drug Benefits. 2015;8:30-40.
3. Chen DW, Lang BH, McLeod DS, et al. Thyroid cancer. Lancet. 2023;401:1531-44.
4. Burns WR, Zeiger MA. Differentiated thyroid cancer. Semin Oncol. 2010:557-66.
5. Schmidbauer B, Menhart K, Hellwig D, Grosse J. Differentiated thyroid cancer—treatment: state of the art. Int J Mol Sci. 2017;18:1292.
6. Panagiotakopoulos T, Chorti A, Pliakos I, et al. Thyroid cancer and pregnancy: a systematic ten-year-review. Gland surgery. 2024;13:1097-107.
7. Caron N, Clark O. Well differentiated thyroid cancer. Scand J Surg. 2004;93:261-71.
8. Belle V, Papantonis I. Principles and practice of explainable machine learning. Frontiers in big Data. 2021;4:688969.
9. Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. Ieee Access. 2020;8:42200-16.
10. Akter S, Mustafa HA. Analysis and interpretability of machine learning models to classify thyroid disease. Plos One. 2024;19:e0300670.
11. Sankar S, Sathyalakshmi S. A study on the explainability of thyroid cancer prediction: SHAP values and association-rule based feature integration framework. Computers, Materials & Continua. 2024;79:3111-38.
12. Borzooei S, Tarokhian A. Differentiated Thyroid Cancer Recurrence (Dataset). UCI Machine Learning Repository. 2023. doi: 10.24432/C5632J
13. Borzooei S, Briganti G, Golparian M, et al. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. Eur Arch Otorhinolaryngol. 2024;281:2095-104.
14. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. 2007;35:2769-94.
15. Tan YS, Singh C, Nasseri K, et al. Fast interpretable greedy-tree sums (figs). arXiv. 2023;arXiv:2201.11931.
16. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. 2013:623-31.
17. Nori H, Jenkins S, Koch P, Caruana R. Interpretml: a unified framework for machine learning interpretability. arXiv. 2019;arXiv:190909223.
18. Sudjianto A, Zhang A, Yang Z, et al. PiML toolbox for interpretable machine learning model development and diagnostics. arXiv. 2023;arXiv:230504214.
19. Smallridge RC, Ain KB, Asa SL, et al. American Thyroid Association guidelines for management of patients with anaplastic thyroid cancer. Thyroid. 2012;22:1104-39.
20. Steinschneider M, Pitaro J, Koren S, et al. Differentiated thyroid cancer with biochemical incomplete response: clinico-pathological characteristics and long term disease outcomes. Cancers. 2021;13:5422.
21. Campopiano MC, Ghirri A, Prete A, et al. Active surveillance in differentiated thyroid cancer: a strategy applicable to all treatment categories response. Frontiers in Endocrinol. 2023;14:1133958.
22. Onitilo AA, Engel JM, Lundgren CI, et al. Simplifying the TNM system for clinical use in differentiated thyroid cancer. J Clin Oncol. 2009;27:1872-8.
23. Grønlund MP, Jensen JS, Hahn CH, et al. Risk factors for recurrence of follicular thyroid cancer: a systematic review. Thyroid. 2021;31:1523-30.
24. Taboni S, Paderno A, Giordano D, et al. Differentiated thyroid cancer: the role of ATA nodal risk factors in N1b patients. Laryngoscope. 2021;131:E1029-34.
25. Sibarani IJB, Suharjito S. Enhancing predictive accuracy for differentiated thyroid cancer (DTC) recurrence through advanced data mining techniques. TIN: Terapan Informatika Nusantara. 2024;5:11-22.