Original Article

Prognostication of differentiated thyroid cancer recurrence: An explainable machine learning approach

Ghazi M. Idroes[1], Teuku R. Noviandy[2*], Ghalieb M. Idroes[3], Irsan Hardi[3], Teuku F. Duta[4], Lama MA. Hamoud[5] and Hala T. Al-Gunaid[6]

[1]Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar, Indonesia; [2]Department of Information Systems, Universitas Abulyatama, Aceh Besar, Indonesia; [3]Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar, Indonesia; [4]Medical Research Unit, School of Medicine, Universitas Syiah Kuala, Banda Aceh, Indonesia; [5]Department of Pharmacy Practice, College of Pharmacy, Taibah University, Madinah, Saudi Arabia; [6]Faculty of Medicine Kasr Al-Ainy, Cairo University, Cairo, Egypt

*Corresponding author: rizky_si@abulyatama.ac.id

Abstract

Differentiated thyroid cancer (DTC) generally has a favorable prognosis, but recurrence remains a concern for a subset of patients, highlighting the need for accurate predictive tools. While traditional methods, such as the American Thyroid Association (ATA) prognostic guideline, are widely used, they may not fully capture the complex patterns in clinical data. To address this, we developed a machine learning model using LightGBM and enhanced its interpretability with SHAP (SHapley Additive exPlanations) analysis. Our model, trained on data from 383 DTC patients, identified incomplete response to therapy as the most significant predictor of recurrence, alongside older age and a high-risk level. The model achieved an accuracy of 93.51%, with precision and sensitivity of 94.23% and 96.08%, respectively, using only five key features selected through Recursive Feature Elimination (RFE): age, physical examination, risk, tumor size and treatment response. SHAP analysis provided clear insights into how these features influenced predictions, offering a transparent and interpretable approach to risk stratification. These results highlight the potential of explainable machine learning to improve recurrence prediction, support personalized care, and build clinician trust, while laying the groundwork for further validation in diverse populations.

Keywords: LightGBM, SHAP, supervised learning, medical informatics, recurrence prediction

Introduction

Differentiated thyroid cancer (DTC) arises from the follicular cells of the thyroid gland, and its differentiated nature indicates that the cancer cells still retain some of the normal features and functions of thyroid cells [1,2]. The primary subtypes of DTC are papillary thyroid cancer and follicular thyroid cancer, with the first one being the more prevalent form [3]. Papillary thyroid cancer is characterized by distinctive papillary structures in the tumor tissue, while follicular thyroid cancer typically forms as a solid mass without papillary structures. Although distinct, both subtypes exhibit a differentiated nature, indicating the preservation of specific normal thyroid cell traits.

Despite a generally favorable prognosis associated with DTC and associated with relatively low mortality rates, there remains concern regarding the risk of recurrence [4]. The challenges lie in the unpredictable behavior of some cases, as a subset of patients may experience recurrence,

Received: November 28, 2024 - Accepted: December 28, 2024 - Published online: December 31, 2024

necessitating ongoing surveillance and management. Understanding the factors contributing to recurrence risk is crucial for tailoring effective follow-up strategies and interventions, thereby optimizing outcomes for individuals affected by DTC. To address this concern, the American Thyroid Association (ATA) has established prognostic guideline that plays a crucial role in assessing and estimating the risk of recurrence for individuals diagnosed with DTC [5]. The guideline offers a comprehensive and widely accepted framework in the medical community for evaluating the likelihood of disease recurrence based on various clinical and pathological factors [6].

The development of artificial intelligence (AI) and machine learning offers a promising way to enhance the prediction of recurrent DTC. Machine learning models can effectively examine large data sets, incorporate various factors, and detect subtle patterns that may signify an upcoming recurrence [7,8]. Researchers and medical professionals may be able to improve the predictive accuracy of models by integrating machine learning algorithms, opening the door to more individualized and successful treatment plans for thyroid cancer patients. Several studies have explored the application of machine learning algorithms in various cancer types, demonstrating improved predictive capabilities. For instance, a study focused on predicting breast cancer recurrence using multiple machine-learning algorithms, identifying the OneR algorithm as particularly effective for balancing sensitivity and specificity [9]. Similarly, a previous study utilized machine learning classification algorithms to predict cervical cancer recurrence, highlighting decision trees as especially useful for identifying predictive factors and linking improved socio-cultural conditions to reduced recurrence rates [10]. Another study employed five classifiers in early-stage endometrial cancer, including a Support Vector Machine, Random Forest, and Boosted Trees, achieving robust predictive accuracy with a multi-algorithmic approach [11]. For DTC, Boorzoei *et al*. [12] demonstrated the utility of classic machine learning models in effectively stratifying recurrence risk.

Recently, gradient-boosting algorithms have gained significant popularity in machine learning, often outperforming deep learning models in tabular data tasks [13]. One such example is Light Gradient Boosting Machine (LightGBM), a highly efficient gradient-boosting framework [14]. LightGBM has demonstrated exceptional performance in various predictive tasks, including medical applications, due to its ability to handle large datasets, capture complex patterns, and process categorical data efficiently [15-17]. Its ability to build accurate models quickly and with less computational resource demand, makes it an attractive option for predicting outcomes like cancer recurrence, offering the potential for more precise and individualized treatment plans for thyroid cancer patients.

As AI and machine learning become more common in medical research and clinical applications, there's a growing focus on the need for explainable models [18]. Machine learning algorithms can be quite complex, often resulting in "black box" models where it is hard to understand how predictions are made [19,20]. In healthcare, especially in critical areas like cancer prognosis, having clear explanations for predictions is crucial. One method that can help with this is SHAP (SHapley Additive exPlanations). SHAP explains model predictions by showing how much each feature contributes to the decision, making it is easier to understand how different factors affect the outcome [21]. This transparency helps healthcare professionals interpret AI predictions clearer, which is important for patient communication and decision-making. Ultimately, SHAP builds trust in AI models and supports their use in clinical settings.

In this study, our primary objective was to develop a machine-learning model for predicting the recurrence of DTC by prioritizing an explainable machine-learning approach. In contrast to conventional "black box" model, our method sought to produce discernible, interpretable insights into the fundamental elements influencing recurrence risk, in addition to accurate predictions. The aim of this study was to promote trust and usefulness among healthcare providers while facilitating efficient patient communication by addressing the crucial demand for transparency. This approach advances evidence-based and individualized management solutions for people with DTC by integrating AI into thyroid cancer prognosis.

Methods

Dataset

The dataset utilized in this study originated from the research conducted by Borzooei *et al*. [12], which involved a retrospective cohort of 383 patients diagnosed with DTC at a single medical center. The diagnoses encompassed histopathological subtypes such as papillary, micropapillary, follicular, and Hürthle cell carcinoma. The study spanned 15 years, with all patients undergoing a minimum follow-up period of 10 years from the time of surgery and initial diagnosis. Among the 383 patients, 275 experienced recurrences, while 108 did not. The dataset included 16 features, all categorical, which were encoded for analysis. A detailed description of each feature in the dataset is presented in **Table 1**.

Table 1. Description of features of the dataset used in this study

Feature name	Data type	Description (values)
Age	Integer	Patient's age in years (15–82 years old)
Sex	Categorical	Sex of the patient (0: Female, 1: Male)
Smoking	Categorical	Smoking status of the patient (0: No, 1: Yes)
History smoking	Categorical	History of smoking for the patient (0: No, 1: Yes)
History radiotherapy	Categorical	History of radiotherapy for the patient (0: No, 1: Yes)
Thyroid function	Categorical	Thyroid function status (0: Euthyroid, 1: Subclinical hypothyroidism, 2: Clinical hypothyroidism, 3: Subclinical hyperthyroidism, 4: Clinical hyperthyroidism)
Physical examination	Categorical	Results of the physical examination (0: Normal, 1: Single nodular goiter-right, 2: Single nodular goiter-left, 3: Multinodular goiter, 4: Diffuse goiter)
Adenopathy	Categorical	Presence or absence of adenopathy (0: No, 1: Right, 2: Left, 3: Bilateral, 4: Posterior, 5: Extensive)
Pathology	Categorical	Pathological characteristics (0: Papillary, 1: Micropapillary, 2: Follicular, 3: Hurthel cell)
Focality	Categorical	Focality of the thyroid cancer (0: Uni-focal, 1: Multi-focal)
Risk	Categorical	American Thyroid Association Risk (0: Low, 1: Intermediate, 2: High)
T	Categorical	Size of tumor (0: T1a, 1: T1b, 2: T2, 3: T3a, 4: T3b, 5: T4a, 6: T4b)
Node	Categorical	Lymph node (0: N0, 1: N1a, 2: N1b)
Metastasis	Categorical	Metastasis (0: M0, 1: M1)
Stage	Categorical	Cancer stage based on TNM classification (0: I, 1: II, 2: III, 3: IVa, 4: IVb)
Response	Categorical	Patient's response to treatment (0: Excellent, 1: Indeterminate, 2: Biochemical incomplete, 3: Structural incomplete)

The dataset was split into two subsets using stratified random sampling: 80% for training and 20% for testing. This approach ensures the model was trained on a large portion of the data, helping it learn patterns and relationships [22]. The training set was used to adjust the model's parameters with different machine learning algorithms, allowing it to make predictions on new data. The testing set was kept separate to evaluate how well the model performs on data it hadn't encountered before.

Feature selection

We performed feature selection to identify and retain the most predictive variables to improve the modeling process, refining the dataset for better model performance and interpretability [23,24]. We employed Recursive Feature Elimination (RFE), a systematic approach that evaluates features based on their contribution to the model's performance. Using an iterative process, one feature was removed at a time and recalculated the model's accuracy at each step, continuing this process until only one feature remained.

The best feature set was selected based on achieving the highest model accuracy with the smallest number of features, ensuring a balance between performance and simplicity. This approach ensures that the selected features contribute maximally to the model's predictive capability while reducing redundancy and complexity. By using accuracy as the fitness value and prioritizing minimal feature sets, the method enhances the final model's efficiency and interpretability [25].

LightGBM model

The LightGBM model was set up to balance performance and efficiency. It used the gradient-boosting decision tree method, which was known for its high accuracy and efficiency in training. The model allowed up to 31 leaves per tree to manage complexity, and trees could grow to any depth to capture detailed patterns in the data. The learning rate was set to 0.1 to maintain a good balance between training speed and model accuracy, with the model running for 100 boosting rounds [26].

Model explainability

To improve the explainability of our predictive model for DTC recurrence, SHAP was used to understand how each feature contributed to the model's predictions [21]. We visualized both the overall impact of features on the model's decisions and their individual effects. This approach helps make the model's decision-making process more transparent, allowing healthcare professionals and patients to understand the results better.

Performance evaluation

To assess the effectiveness of our predictive model for DTC recurrence, a comprehensive set of performance metrics was employed, including accuracy, precision, sensitivity, specificity, and F1-score. Accuracy measures the overall correctness of predictions, while precision gauges the proportion of correctly predicted positive instances among all predicted positives. Sensitivity assesses the ability of the model to capture all actual positive instances, and specificity assesses the ability of the model to capture all actual negative instances. The F1-score, which combines precision and recall, provides a balanced measure of the model's performance. The accuracy, precision, sensitivity, specificity, and F1-score equations are presented in Equations 1–5 [27,28].

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (4)$$

$$\text{F1} - \text{score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

where TP is true positive, FP is false positive, FN is false negative, and TN is true negative.

Experimental setup

Our predictive model for DTC recurrence was developed and implemented using Python version 3.10.9 (Python Software Foundation, Delaware, USA), along with open-source libraries such as Scikit-learn (version 1.2.0) for machine learning functionalities, LightGBM (version 4.1.0) as the gradient boosting framework, and SHAP (version 0.41.0) for enhanced model interpretability through SHAP values. Additionally, a fixed random state of 0 (zero) was employed throughout the process to maintain reproducibility.

Results

Before developing the LightGBM models, a thorough analysis of the dataset was conducted to gain a deeper understanding of its structure and key characteristics. Principal component analysis (PCA) was used as a dimensionality reduction technique, allowing the data to be visualized in a two-dimensional space. The results of this analysis are presented in **Figure 1**, where the data points are represented in the reduced-dimensional space. In this analysis, the first two principal components (PC), PC-1 and PC-2, were selected for visualization. PC-1 accounted for the majority of the variance in the dataset, capturing 96.23%, while PC-2 was orthogonal to PC-1 and explained

an additional 1.25% of the variance. The scattered distribution of points underscored the potential presence of complex, non-linear relationships among the features in the dataset. These complexities would likely be difficult for traditional linear models to capture, and suggested the potential benefits of employing LightGBM, which is known for its effectiveness in handling non-linear relationships within data. LightGBM's adaptability to such complexities in the dataset makes it a suitable choice for predictive modeling, offering the capability to capture intricate patterns that may contribute to the recurrence prediction of DTC.

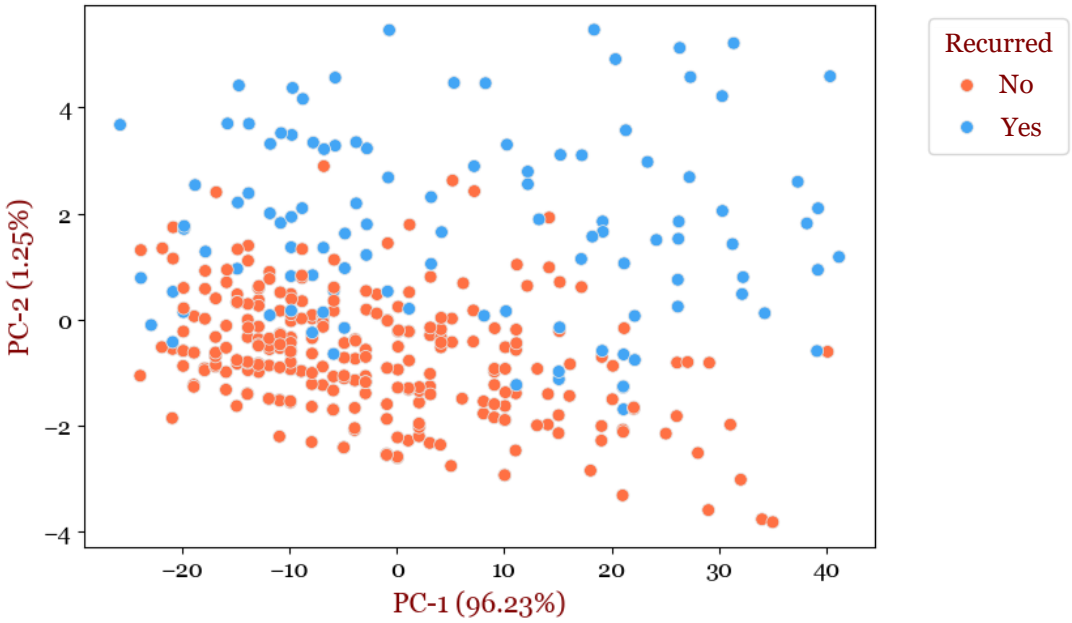

Figure 1. Principal component analysis (PCA) visualization of the differentiated thyroid cancer (DTC) dataset. The figure indicates the lack of linear separability between the two classes. The data points are scattered across the plot with significant overlap between recurrence and non-recurrence cases. This indicates that the dataset does not have clear linear boundaries, which could make it challenging for simple linear models, such as logistic regression, to effectively distinguish between the two classes. The majority of the data points are concentrated near the center of the plot, with some spreading along principal component 1 (PC-1), reflecting its dominant role in capturing variance. Principal component 2 (PC-2), while accounting for a smaller percentage of variance, still contributes to separating certain data points, particularly those on the periphery of the plot.

The next step in the analysis involved performing RFE to select the most relevant features for the model. The results of this process are presented in **Figure 2**. Interestingly, the accuracy of the model using a subset of just five features was found to be equal to that of the model using all 16 features. This suggests that the additional 11 features do not significantly improve the model's predictive performance. In other words, the five selected features—'Age,' 'Physical Examination,' 'Risk,' 'T,' and 'Response'—appear to capture the essential information needed for effective prediction, while the other features may either be redundant or irrelevant for the task at hand.

The five features identified by RFE were then used to train the LightGBM model. By focusing on this smaller, more efficient set of features, the model was trained more quickly, with less computational overhead, while still maintaining high predictive accuracy. Additionally, using a reduced set of features enhances the model's explainability, making it easier to interpret how individual features contributed to the predictions. This step highlights the importance of feature selection in improving model efficiency and performance, while also making the model more transparent and interpretable.

The performance metrics for the LightGBM model trained using the top five features are presented in **Figure 3**. The model achieved a high accuracy of 93.51%, indicating that it made

correct predictions in a large majority of cases. This level of accuracy reflects the model's ability to generalize well to unseen data, providing a solid foundation for its predictive reliability. The precision of 94.23% indicates that when the model predicted a positive outcome, it was correct 94.23% of the time. This suggests that the model effectively minimizes false positives, making it reliable when the cost of incorrectly predicting a positive case is high. The sensitivity of 96.08% shows that the model correctly identified 96.08% of the actual positive instances in the dataset. This high sensitivity is particularly important in medical or high-stakes applications, where missing a positive case (false negative) could have significant consequences. On the other hand, the specificity of 88.46% indicates that the model could identify 88.46% of the actual negative cases correctly. While not as high as sensitivity, this is still a solid result, reflecting the model's ability to avoid false positives, which is crucial for maintaining a balance between detecting positives and minimizing unnecessary interventions.

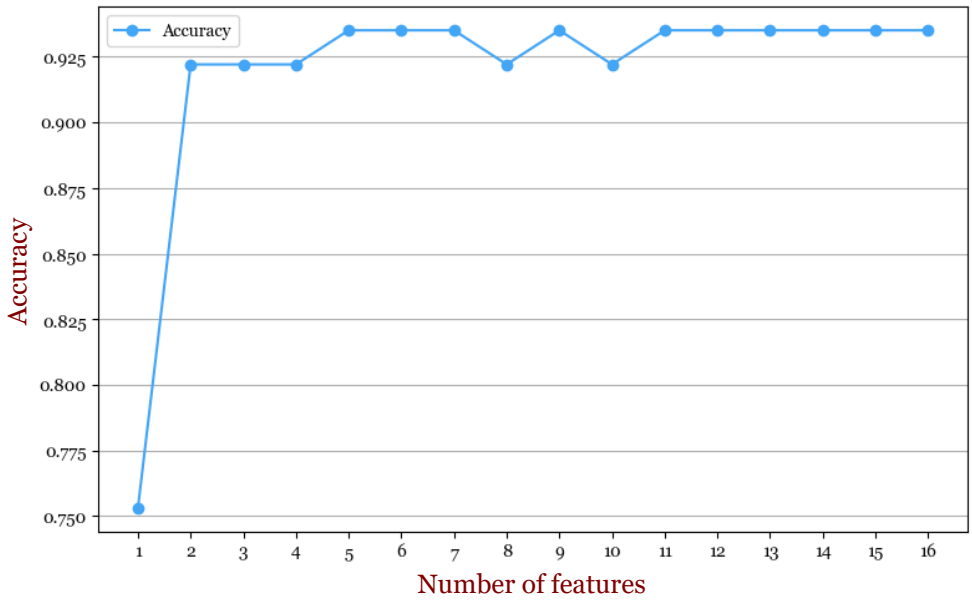

Figure 2. Recursive Feature Elimination (RFE) results show model accuracy as features are iteratively reduced.

Finally, the F1-score, which balances both precision and sensitivity, achieved a value of 95.15%. This score provided an overall measure of the model's ability to correctly identify positive cases while minimizing false positives and false negatives. A high F1-score like this demonstrated that the model was performing well in both identifying true positives and maintaining a low false positive rate. These metrics suggested that the model is well-calibrated, effectively identifying positive instances while maintaining a strong balance between sensitivity and specificity. This highlighted the model's robustness and potential for use in predictive tasks where accuracy and reliability are important.

To assess the prediction results, we present the confusion matrix in **Table 2**, which provides a detailed breakdown of the model's predictions and enables an evaluation of its classification accuracy. The matrix revealed that the model successfully predicted 49 cases of no recurrence, with only two misclassifications. Furthermore, it correctly identified 23 cases of recurrence, although three instances were misclassified. This outcome highlighted the model's strong performance in distinguishing between recurrence and no recurrence, particularly in its ability to identify patients without recurrence accurately. The relatively low number of misclassifications further demonstrated the model's overall effectiveness in making precise predictions.

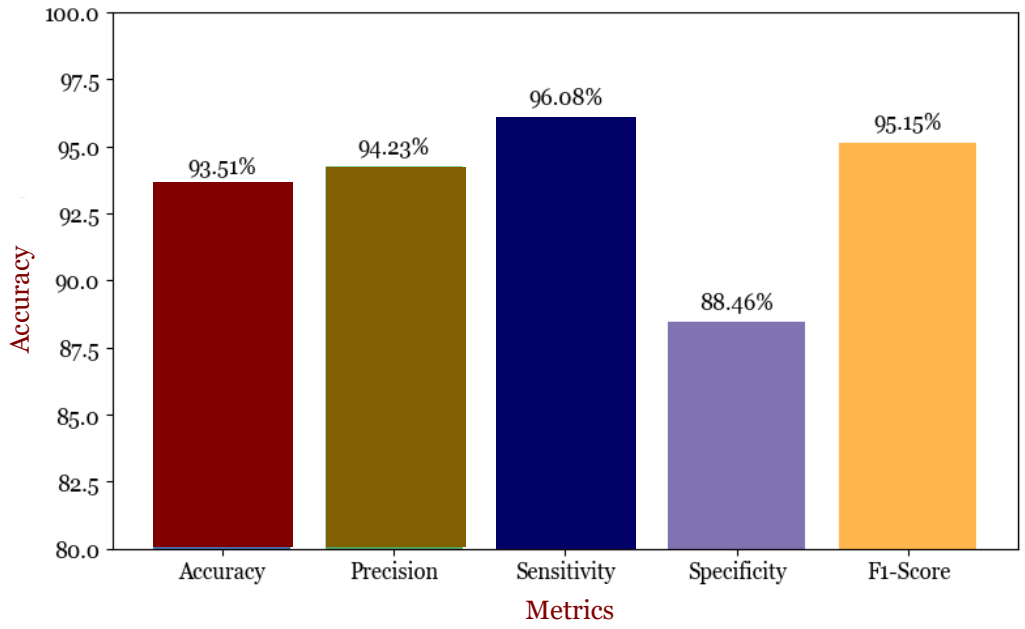

Figure 3. Performance metrics for Light Gradient Boosting Machine (LightGBM) model with top five features (age, physical examination, risk, tumor size and treatment response).

To further understand the model's decision-making process, the SHAP values were then examined, which provided a deeper insight into how individual feature contributed to the model's predictions for DTC recurrence. The SHAP value bar plot demonstrating the impact of various features on the model's predictions for DTC recurrence is presented in **Figure 4**. The SHAP values indicate the magnitude and direction of each feature's effect on the model's output. A positive SHAP value implies that the feature increases the likelihood of predicting a recurrence of DTC. In contrast, a negative SHAP value suggests a decrease in the likelihood of recurrence.

Table 2. Confusion matrix of the Light Gradient Boosting Machine (LightGBM) model with top five features

Actual	Predicted	
	Recurrent – No	Recurrent - Yes
Recurrent - No	49	2
Recurrent - Yes	3	23

The 'Response', the clinical response to the treatment, feature from the bar plot substantially impacts the model's predictions, with a mean absolute SHAP value of approximately +0.35 (**Figure 4**). This suggested that the patient's response to treatment is a significant predictor of recurrence, with a positive response correlating with a lower likelihood of recurrence. 'Risk' and 'Age' follow, each with a mean absolute SHAP value of +0.03, indicating that these factors moderately impact the prognosis. Higher risk levels and older age contribute to an increased probability of recurrence. 'Physical examination' and 'T' (tumor) were the least impactful features, each with a mean absolute SHAP value of +0.01. Their lower values suggest that while contributing to the model's predictions, their influence is relatively minor compared to factors like 'Response.'

A SHAP decision plot is presented in **Figure 5**, visually explaining how each feature contributed to the model's output value for a specific prediction. The decision plot showed the cumulative effect of each feature on the model's output, starting from the base value (the average model output over the dataset when no features are considered) and ending at the actual model output value for a particular instance.

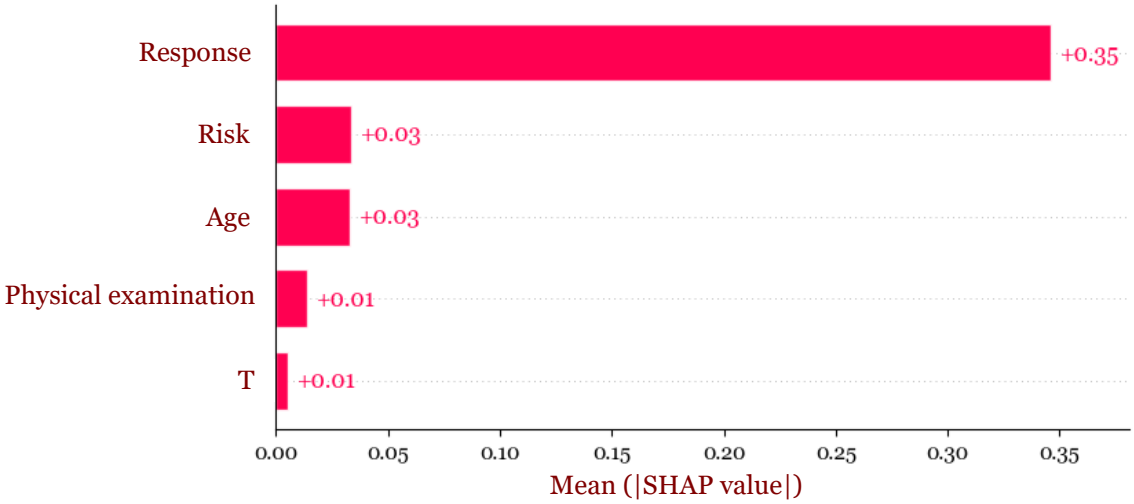

Figure 4. Figure 4. Bar plot of SHAP (SHapley Additive exPlanations) analysis for Light Gradient Boosting Machine (LightGBM) model with top five features.

The 'Response' feature had a strong red indication, suggesting that it significantly increases the model's output value, thus raising the predicted risk of recurrence (**Figure 5**). Conversely, blue starting points for many instances, which suggested that when no other information was provided, the model might tend to predict a lower risk of recurrence. The overlapping lines for 'Risk,' 'Age,' 'Physical Examination,' and 'T' show varying degrees of influence on the model's output. Some lines for 'Risk' and 'Age' veer into the red area, indicating that, in some cases, these features also contribute to a higher prediction of recurrence risk. Notably, the decision plot showed individual prediction paths, which can diverge significantly even if they start or end at similar values. This divergence illustrated the interplay and relative weights of the different features in the context of individual predictions (**Figure 5**).

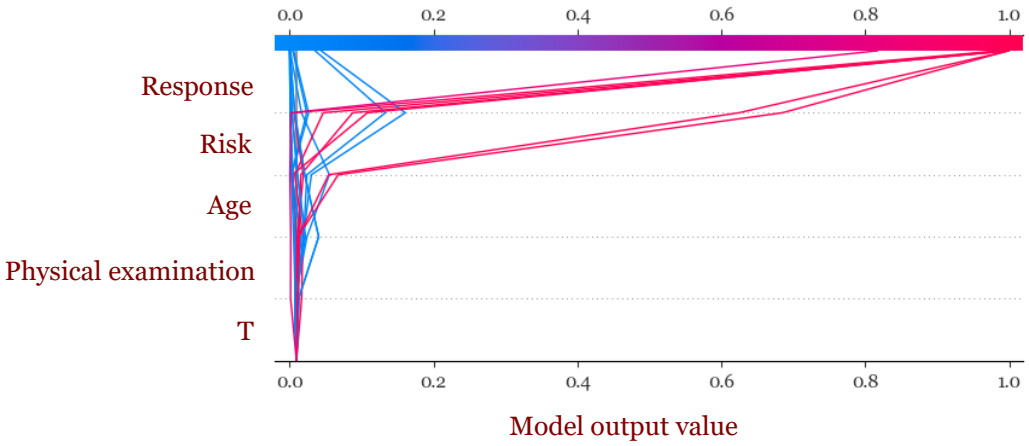

Figure 5. SHAP (SHapley Additive exPlanations) decision plot for Light Gradient Boosting Machine (LightGBM) model with top five features. Each line represents an individual prediction for a patient, tracing how the combined features lead to the final output value. The plot starts on the left with the base value and moves to the right, showing the impact of each feature. Features that increase the model output are presented in red, indicating a higher likelihood of differentiated thyroid cancer (DTC) recurrence. In contrast, features that decrease the prediction are presented in blue, indicating a lower likelihood of recurrence.

The SHAP dependence plot, presented in **Figure 6**, visualized the individual contributions of the top five features to the model's predictions of DTC recurrence. These plots provided a detailed view of the relationship between each feature's categorical or numerical values and their SHAP values, which quantified the feature's impact on the prediction outcome.

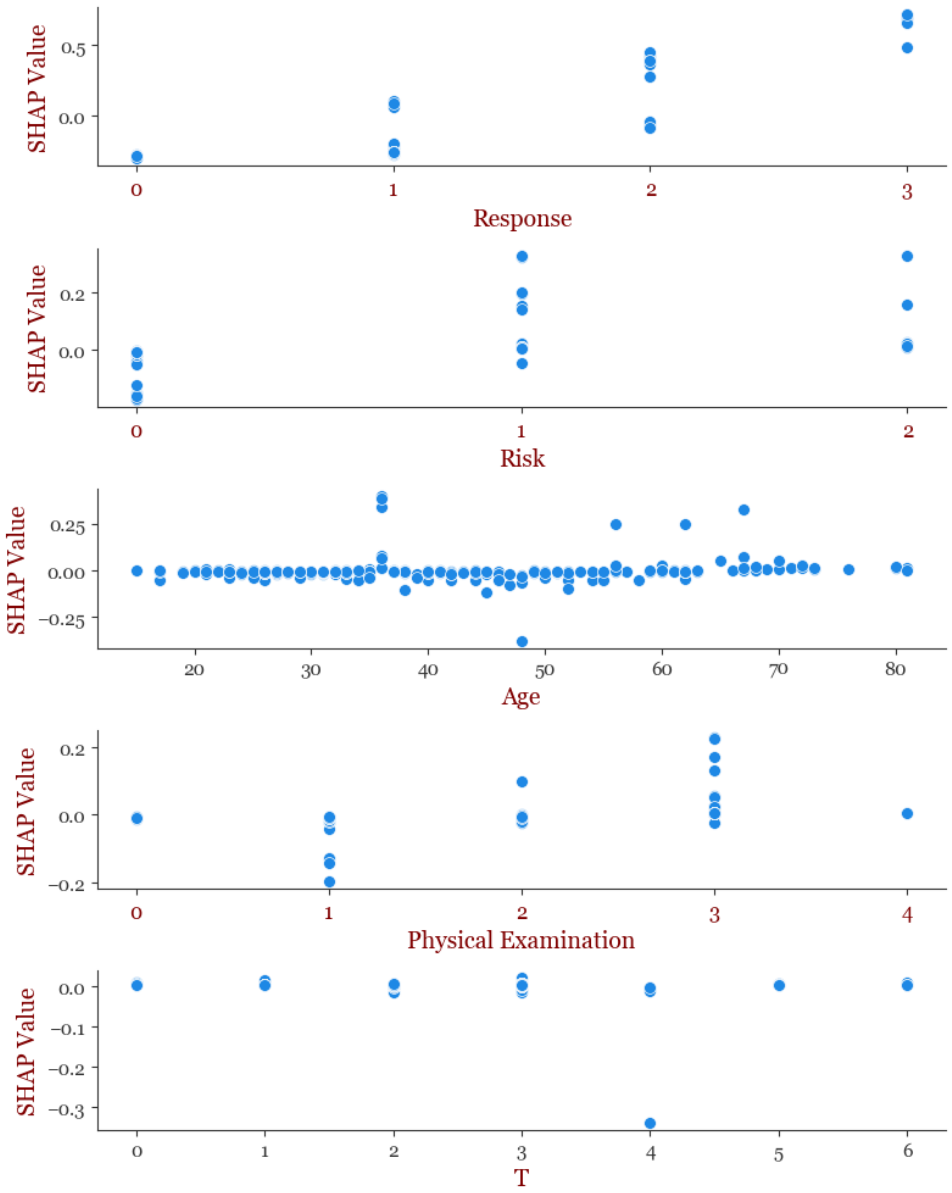

Figure 6. Dependence plot of SHAP (SHapley Additive exPlanations) analysis for Light Gradient Boosting Machine (LightGBM) models with top five features.

The 'Response' feature, categorized into four levels (0: Excellent, 1: Indeterminate, 2: Biochemical incomplete, 3: Structural incomplete), demonstrated the most substantial influence on the model's predictions (**Figure 6**). The dependence plot revealed a clear trend where higher response categories, indicating poorer treatment outcomes, were associated with larger positive SHAP values. This signified that as the patient's response moved toward 'Structural incomplete' (3), the likelihood of recurrence significantly increased. This strong association underscored the clinical importance of treatment response in predicting DTC outcomes.

Similarly, the 'Risk' feature, classified into three levels (0: Low, 1: Intermediate, 2: High), exhibited a moderate yet consistent impact on predictions (**Figure 6**). The dependence plot showed that patients with a high-risk classification (2) tended to have higher positive SHAP

values, reflecting an increased probability of recurrence. This finding aligned with established clinical evidence that higher-risk patients were more likely to experience adverse outcomes, validating the model's reliance on this feature for prediction. The 'Age' feature, represented as a continuous variable ranging from 15 to 82 years, showed a subtler influence on the model's predictions (**Figure 6**). The dependence plot indicated that older age was generally associated with slightly higher positive SHAP values, suggesting a modest increase in recurrence likelihood for older patients. However, the overall distribution of SHAP values for this feature remained narrower compared to 'Response' and 'Risk,' indicating that while age contributed to predictions, its impact was less pronounced (**Figure 6**). The 'Physical Examination' feature, categorized based on examination findings (0: Normal, 1: Single nodular goiter-right, 2: Single nodular goiter-left, 3: Multinodular goiter, 4: Diffuse goiter), exhibited a smaller range of SHAP values. The dependence plot suggested that findings such as multinodular or diffuse goiter might have contributed modestly to an increased risk of recurrence, but their overall influence was relatively minor. This feature appeared to have played a more supportive role in refining the model's predictions rather than being a primary driver. Finally, the 'T' feature, which represented tumor size and was categorized from 0 (T1a) to 6 (T4b), showed a limited range of SHAP values in its dependence plot (**Figure 6**). Larger tumor sizes, particularly in the higher categories (e.g., T4a and T4b), were associated with slightly positive SHAP values, indicating a small contribution to recurrence likelihood. However, this effect was minimal compared to the influence of features like 'Response' and 'Risk.'

A SHAP summary plot of an individual prediction from the model (**Figure 7**), specifically for a case where recurrence of DTC did not occur. The plot showed the individual SHAP values for each feature and their impact on the model's prediction for this specific case. In this example, the 'Response' feature had the most substantial negative impact on the prediction, with a SHAP value of -0.26. The 'Age' of the patient contributed a smaller negative impact with a SHAP value of -0.01. Similarly, 'Risk' and 'Physical Examination' negatively influenced the prediction with SHAP values of -0.01. The feature 'T' had a SHAP value of 0, meaning it did not shift the prediction away.

A SHAP summary plot for an individual prediction example where DTC recurrence occurred is presented in **Figure 8**. For this individual, 'Response' had a positive SHAP value of +0.72, making it was the most influential factor in predicting recurrence. 'T' has a small positive contribution with a SHAP value of +0.01, slightly increasing the likelihood of predicting a recurrence. Conversely, 'Physical Examination' and 'Risk' had negative SHAP values (-0.01), indicating a minor contribution toward predicting no recurrence. 'Age' had a SHAP value of 0, indicating no effect on the prediction outcome.

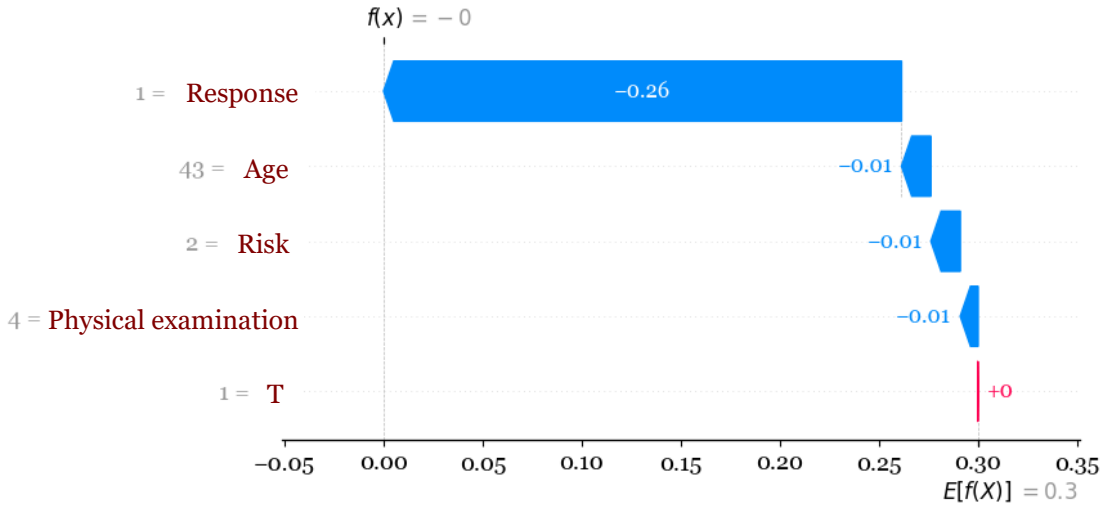

Figure 7. Bar plot of SHAP (SHapley Additive exPlanations) analysis for individual prediction where differentiated thyroid cancer (DTC) recurrence did not occur.

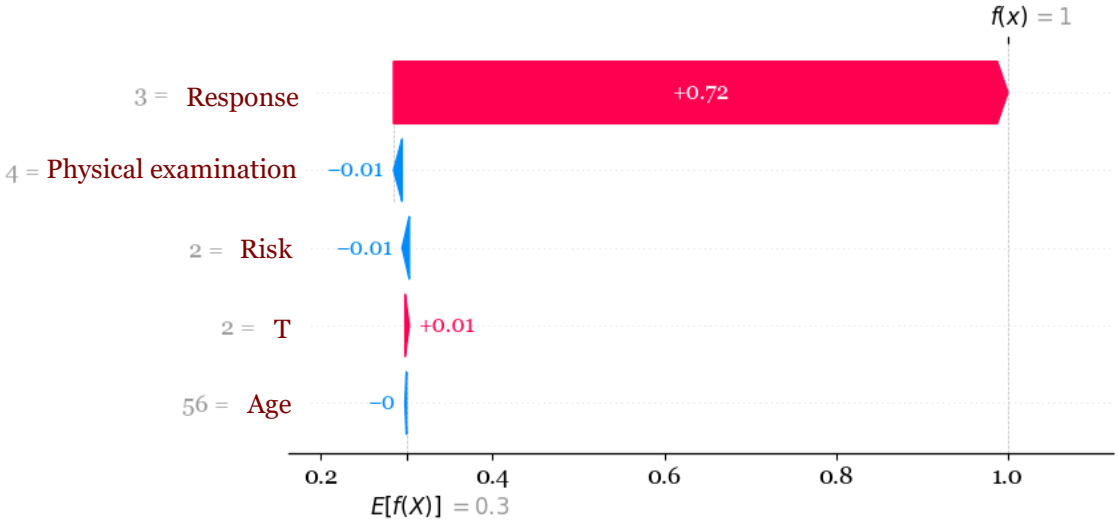

Figure 8. Bar plot of SHAP (SHapley Additive exPlanations) analysis for individual prediction where differentiated thyroid cancer (DTC) recurrence occurs.

Discussion

This study explored the use of machine learning, specifically the LightGBM model, for predicting recurrence in DTC. By incorporating SHAP values, we understood how different clinical factors contributed to the model's predictions. This approach makes the results more transparent and easier for clinicians to interpret.

Our findings showed that the patient's response to initial therapy was the most important factor in predicting recurrence. This matches what is already known in clinical practice—patients with poor treatment responses are more likely to have their cancer come back. Other factors, like older age and being classified as high risk, also played a role, but their effects were smaller. These results confirm that the model aligns with medical knowledge while also identifying patterns from data that may not be immediately obvious.

One key strength of this study is the focus on explainability. Using SHAP, we could clearly see how each feature, such as treatment response or age, influenced the model's predictions [29]. This is important because doctors need to understand why the model makes certain predictions, especially in critical decisions like cancer follow-up care. For example, our decision plots showed that incomplete responses to treatment and older age consistently increased the risk of recurrence. This kind of insight can help doctors decide which patients need closer monitoring [30].

Another benefit of explainable models is improved communication with patients. When doctors can explain the reasons behind a prediction, patients may feel more confident in their care plan [31]. For instance, a patient with a high predicted risk of recurrence due to poor treatment response might better understand why they need more frequent follow-ups or additional treatments.

The results of this study also support personalized care for DTC patients. By identifying high-risk patients based on specific factors like treatment response and age, doctors can create follow-up plans tailored to each patient's needs. For example, patients at high risk may need closer monitoring, while those at low risk could avoid unnecessary tests or interventions. This approach can improve patient outcomes and make better use of healthcare resources.

While the results are promising, there are some limitations. The data for this study came from one medical center, so the findings might not apply to all patient groups [12]. Future studies should test the model with data from different hospitals and populations to confirm its accuracy and usefulness. Also, the model only used clinical features; adding genetic, imaging, or other types of data could improve its predictions.

Although SHAP helps make the model's predictions understandable, the results must be carefully interpreted. For example, the importance of certain features might vary in different

patient groups or settings. Overemphasizing a single feature without considering the full clinical picture could lead to errors.

Conclusion

This study has successfully demonstrated the application of machine learning and explainability techniques like SHAP in improving predictions for DTC recurrence. By using a LightGBM model, we achieved high accuracy while also providing insights into how clinical factors, such as response to initial therapy, age, and risk levels, influence predictions. Explainability enhances trust and usability in healthcare settings, enabling clinicians to effectively understand and communicate the rationale behind model predictions. This approach supports personalized care and better-informed decision-making, demonstrating that machine learning when paired with interpretable tools, can be a valuable asset in advancing cancer prognosis and management.

Ethics approval

Not required.

Acknowledgments

None to declare.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

The data used in this study was retrieved from: https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence.

Declaration of artificial intelligence use

We hereby confirm that no artificial intelligence (AI) tools or methodologies were utilized at any stage of this study, including during data collection, analysis, visualization, or manuscript preparation. All work presented in this study was conducted manually by the authors without the assistance of AI-based tools or systems.

How to cite

Idroes GM, Noviandy TR, Idroes GM, *et al*. Prognostication of differentiated thyroid cancer recurrence: An explainable machine learning approach. Narra X 2024; 2 (3): e183 - https://doi.org/10.52225/narrax.v2i3.183.

References

1. Burns WR, Zeiger MA. Differentiated Thyroid Cancer. Semin Oncol 2010;37(6):557-566.
2. Schlumberger M, Leboulleux S. Current practice in patients with differentiated thyroid cancer. Nat Rev Endocrinol 2021;17(3):176-188.
3. Lloyd R V, Buehler D, Khanafshar E. Papillary thyroid carcinoma variants. Head Neck Pathol 2011;5:51–56.
4. Li M, Brito JP, Vaccarella S. Long-term declines of thyroid cancer mortality: An international age–period–cohort analysis. Thyroid® 2020;30(6):838-846.
5. Lee J, Lee SG, Kim K, *et al.* Clinical value of lymph node ratio integration with the 8th edition of the UICC TNM classification and 2015 ATA risk stratification systems for recurrence prediction in papillary thyroid cancer. Sci Rep 2019;9(1):13361.
6. Schmidbauer B, Menhart K, Hellwig D, *et al.* Differentiated thyroid cancer-treatment: State of the art. Int J Mol Sci 2017;18(6):1292.

7. Maulana A, Faisal FR, Noviandy TR, *et al.* Machine learning approach for diabetes detection using Fine-Tuned XGBoost algorithm. Infolitika J Data Sci 2023;1(1):1-7.
8. Noviandy TR, Maulana A, Emran TB, *et al.* QSAR classification of beta-secretase 1 inhibitor activity in alzheimer's disease using ensemble machine learning algorithms. Heca J Appl Sci 2023;1(1):1-7.
9. Alzu'bi A, Najadat H, Doulat W, *et al.* Predicting the recurrence of breast cancer using machine learning algorithms. Multimed Tools Appl 2021;80(9):13787-13800.
10. Asadi F, Salehnasan C, Ajori L. Supervised algorithms of machine learning for the prediction of cervical cancer. J Biomed Phys Eng 2020;10(4):513-522.
11. Akazawa M, Hashimoto K, Noda K, *et al.* The application of machine learning for predicting recurrence in patients with early-stage endometrial cancer: A pilot study. Obstet Gynecol Sci 2021;64(3):266-273.
12. 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(4):2095-2104.
13. Carlens H. State of competitive machine learning in 2022. ML Contests Research; 2023.
14. Ke G, Meng Q, Finley T, *et al.* Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.
15. Noviandy TR, Nainggolan SI, Raihan R, *et al.* Maternal health risk detection using light gradient boosting machine approach. Infolitika J Data Sci 2023;1(2):48-55.
16. Yang H, Chen Z, Yang H, *et al.* Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. IEEE Access 2023;11:23366-23380.
17. Rufo DD, Debelee TG, Ibenthal A, *et al.* Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). diagnostics 2021;11(9):1714.
18. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics 2022;12(2):237.
19. Noviandy TR, Maulana A, Khowarizmi F, *et al.* Effect of CLAHE-based enhancement on bean leaf disease classification through explainable AI. 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE). IEEE; 2023.
20. Meng C, Trinh L, Xu N, *et al.* Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Sci Rep 2022;12(1):7166.
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
22. Noviandy TR, Idroes GM, Maulana A, *et al.* Credit card fraud detection for contemporary financial management using XGBoost-driven machine learning and data augmentation techniques. Indatu J Manag Account 2023;1(1):29-35.
23. Li J, Cheng K, Wang S, *et al.* Feature selection: A data perspective. ACM Comput Surv CSUR 2017;50(6):1-45.
24. Noviandy TR, Maulana A, Sasmita NR, *et al.* The prediction of kovats retention indices of essential oils at gas chromatography using genetic algorithm-multiple linear regression and support vector regression. J Eng Sci Technol 2022;17(1):306-326.
25. Bahl A, Hellack B, Balas M, *et al.* Recursive feature elimination in random forest classification supports nanomaterial grouping. NanoImpact 2019;15:100179.
26. Noviandy TR, Idroes GM, Mohd Fauzi F, *et al.* Application of ensemble machine learning methods for QSAR classification of leukotriene A4 hydrolase inhibitors in drug discovery. Malacca Pharm 2024;2(2):68-78.
27. Idroes GM, Maulana A, Suhendra R, *et al.* TeutongNet: A fine-tuned deep learning model for improved forest fire detection. Leuser J Environ Stud 2023;1(1):1-8.
28. Agustia M, Noviandy TR, Maulana A, *et al.* Application of fuzzy support vector regression to predict the Kovats retention indices of flavors and fragrances. 2022 International Conference on Electrical Engineering and Informatics (ICELTICs). IEEE; 2022.
29. Amann J, Blasimme A, Vayena E, *et al.* Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak 2020;20(1):310.
30. Antoniadi AM, Du Y, Guendouz Y, *et al.* Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Appl Sci 2021;11(11):5088.
31. Yang CC. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. J Healthc Inform Res. 2022;6(2):228–239.