Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

Explainable Machine Learning Framework for Thyroid Cancer Recurrence Prediction

Tuti Alawiyah[1)], Taufik Wibisono[2)], Recha Abriana Anggraini[3)], Bambang Kelana Simpony[4)], Yesti Siti Nurjanah[5)]
[1,2,3,4)] Universitas Bina Sarana Informatika, Indonesia
[5)] Politeknik Triguna Tasikmalaya, Indonesia
[1)]tuti.tah@bsi.ac.id, [2)]taufik.tik@bsi.ac.id, [3)]recha.rcb@bsi.ac.id, [4)]bambang.bky@bsi.ac.id, [5)]yestisitinurjanah@poltektriguna.ac.id

ABSTRACT

Accurate prediction of thyroid cancer recurrence is essential for improving long-term patient management and supporting evidence-based clinical decision-making. Although machine learning has demonstrated promising predictive performance, limited model interpretability remains a major barrier to its clinical adoption. This study aims to develop an Explainable Machine Learning framework for thyroid cancer recurrence prediction by integrating Extreme Gradient Boosting (XGBoost) with SHapley Additive exPlanations (SHAP) using clinicopathological features. A publicly available dataset containing 383 patient records was preprocessed through label encoding, correlation analysis, Chi-Square-based feature selection, and Min-Max normalization. Logistic Regression, Decision Tree, Random Forest, and XGBoost were comparatively evaluated using 10-fold stratified cross-validation with Accuracy, Precision, Recall, F1-score, and ROC-AUC as evaluation metrics. The best-performing model was subsequently interpreted using global and local SHAP analyses. XGBoost achieved the highest performance, with an accuracy of 95.8% ± 4.4%, precision of 93.4% ± 8.3%, recall of 91.4% ± 9.9%, F1-score of 92.2% ± 8.3%, and ROC-AUC of 98.6% ± 2.5%, outperforming the other models. SHAP analysis identified Response, Risk, and N Stage as the most influential clinicopathological factors affecting recurrence prediction. This study contributes by developing a unified Explainable Machine Learning framework that integrates comparative model evaluation, XGBoost prediction, and global and local SHAP interpretation within a single workflow. The proposed framework provides accurate and clinically interpretable recurrence prediction, supporting trustworthy risk assessment and personalized decision-making in thyroid cancer management.

Keywords: explainable machine learning; thyroid cancer recurrence; xgboost; shapley additive explanations (shap); clinicopathological features

INTRODUCTION

Thyroid cancer is one of the most common endocrine malignancies worldwide, with its incidence continuing to increase over the past decades (Arista et al., 2023; Aulia et al., 2023; Nugraha et al., 2025). Although the prognosis is generally favorable, recurrence remains a major clinical concern because it often requires repeated treatment, increases healthcare costs, and negatively affects patients' quality of life. Consequently, early identification of patients at high risk of recurrence is essential for supporting personalized clinical decision-making.

Conventional statistical approaches have limitations in capturing the complex nonlinear relationships among clinicopathological variables. Recently, machine learning (ML) has demonstrated superior capability for disease diagnosis, prognosis, and recurrence prediction (Feng et al., 2026). Various algorithms, including Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost), have been successfully applied to clinical prediction tasks. Among them, XGBoost has received considerable attention because of its computational efficiency, robustness, regularization capability, and excellent predictive performance on structured clinical data (Gunasekara et al., 2024; Jiang et al., 2025).

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

Despite these advantages, predictive accuracy alone is insufficient for clinical implementation because many ML models operate as black boxes whose prediction mechanisms are difficult to interpret. This lack of transparency limits clinicians' trust in model predictions and hinders the adoption of artificial intelligence in healthcare (Abkar et al., 2026). Explainable Artificial Intelligence (XAI) has therefore emerged to improve model transparency by providing understandable explanations of prediction mechanisms (Takwim & Sulaeman, 2025). Among various XAI techniques, SHapley Additive exPlanations (SHAP) has become one of the most widely adopted approaches because it quantifies the contribution of each feature to prediction outcomes while preserving predictive performance (Carmona et al., 2022; Istiwana et al., 2026). Furthermore, SHAP provides both global explanations, which reveal overall feature importance, and local explanations, which explain individual predictions, thereby facilitating clinically interpretable decision-making (Ramadhan & Zeniarja, 2025).

Several studies have applied machine learning to thyroid cancer recurrence prediction. (Gong et al., 2021) compared multiple machine learning algorithms and employed SHAP to identify globally important predictive features. (Schindele et al., 2025) developed an XGBoost-based prediction model with global and local SHAP interpretation, whereas (Redlich et al., 2026) focused on explainable XGBoost for pediatric thyroid cancer recurrence. (Hanani et al., 2025) compared several machine learning algorithms before selecting CatBoost for SHAP-based interpretation, while (Hu et al., 2026) evaluated multiple tree-based models and utilized SHAP to investigate feature importance and simplified prediction models. These studies demonstrate that integrating explainability can improve model transparency without substantially sacrificing predictive performance.

However, as summarized in Table 1, existing studies generally emphasize either comparative evaluation of machine learning algorithms or explainable interpretation of the selected model. Comparative studies commonly identify the best-performing algorithm but provide limited multi-level interpretation, whereas explainability-oriented studies typically interpret a predefined model without establishing its superiority through systematic comparison. Consequently, the integration of comparative model evaluation and comprehensive global and local SHAP explanations within a unified explainable prediction framework remains limited.

Table 1. Comparison of Representative Studies

Study	Comparative ML	Multi-level SHAP	Integrated Explainable Prediction Framework
Gong et al.	√	X	X
Hanani et al.	√	√	X
Schindele et al.	X	√	X
Hu et al.	√	√	X
Redlich et al.	X	√	X
This Study	√	√	√

To address this gap, this study proposes an Explainable Machine Learning Framework for thyroid cancer recurrence prediction. Four supervised learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and XGBoost are systematically compared to identify the most reliable predictive model, after which the selected model is interpreted using both global and local SHAP explanations to improve transparency and clinical interpretability. Accordingly, this study addresses two research questions: RQ1, which machine learning algorithm provides the most reliable performance for thyroid cancer recurrence prediction; and RQ2, which clinicopathological variables contribute most significantly according to global and local SHAP explanations. The main contribution of this study is the integration of comparative model evaluation and multi-level explainability into a unified prediction framework that supports accurate and clinically interpretable recurrence prediction.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

METHOD

This study proposes an Explainable Machine Learning framework for thyroid cancer recurrence prediction by integrating comparative machine learning evaluation with SHapley Additive exPlanations (SHAP). The proposed framework consists of dataset acquisition, data preprocessing and feature selection, comparative model development, model evaluation, systematic best-model selection, SHAP-based model interpretation, and clinically interpretable prediction output. As illustrated in Figure 1, the workflow begins with dataset acquisition, followed by data preprocessing and feature selection. Multiple machine learning algorithms are then comparatively evaluated, after which the best-performing model is systematically selected. Finally, global and local SHAP explanations are applied to interpret prediction outcomes and support clinically meaningful decision-making. Unlike conventional prediction pipelines that primarily emphasize predictive accuracy, the proposed framework integrates comparative model evaluation, systematic best-model selection, and SHAP-based global and local explanations within a unified workflow, providing transparent and clinically interpretable thyroid cancer recurrence prediction.

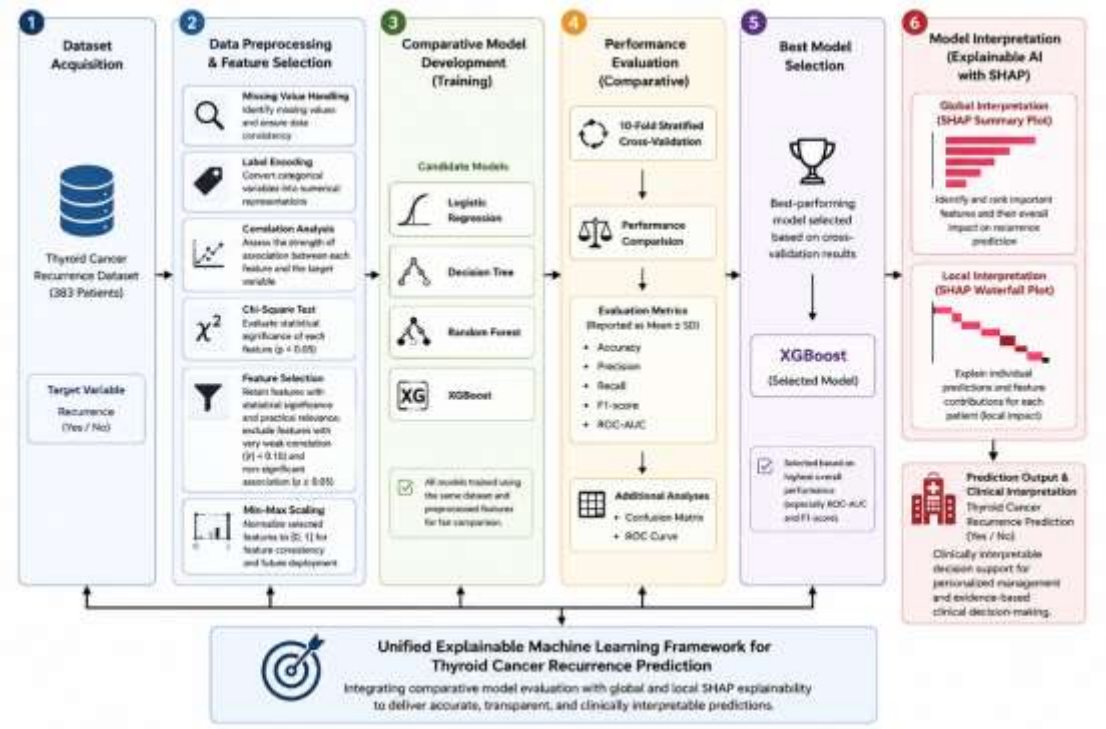

Figure 1. Proposed Explainable Machine Learning Framework for Thyroid Cancer Recurrence Prediction. The framework integrates dataset preprocessing, comparative machine learning evaluation, systematic best-model selection, and SHAP-based global and local interpretation to enable accurate and clinically interpretable thyroid cancer recurrence prediction.

The dataset used in this study was obtained from the publicly available Thyroid Disease Data repository on Kaggle and consists of 383 thyroid cancer patient records containing demographic, clinicopathological, and treatment-related characteristics. The target variable, Recurred, indicates whether thyroid cancer recurrence occurred after treatment and was formulated as a binary classification problem comprising 275 non-recurrence cases (71.8%) and 108 recurrence cases (28.2%). Although the dataset exhibits a moderate class imbalance, this issue was addressed during model evaluation using 10-fold Stratified Cross-Validation, which preserves the original class distribution in each fold and provides a reliable estimate of

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

model generalization performance. The dataset contains sixteen predictor variables, namely Age, Gender, Smoking, History of Smoking, History of Radiotherapy, Thyroid Function, Physical Examination, Adenopathy, Pathology, Focality, Risk Classification, Tumor Stage (T), Lymph Node Stage (N), Metastasis Stage (M), Clinical Stage, and Treatment Response.

An initial data quality assessment confirmed that all 383 records were complete, with no missing values or data inconsistencies; therefore, no records were removed during preprocessing. All categorical variables were subsequently transformed into numerical representations using Label Encoding. Feature selection was then performed through a two-stage procedure consisting of correlation analysis and Chi-Square feature selection. Correlation analysis was first conducted to measure the strength of association between each predictor and the target variable. Subsequently, Chi-Square testing was applied to evaluate the statistical significance of each feature. Following the correlation strength guideline proposed by Sudjana (2022) (Patimah et al., 2025), features exhibiting very weak correlation ($|r| < 0.10$) together with non-significant Chi-Square association ($p \geq 0.05$) were excluded from model development. Based on these criteria, only Thyroid Function and Pathology were excluded, whereas the remaining fourteen features were retained for predictive modeling.

Following feature selection, the retained features were normalized using Min-Max Scaling, which transforms feature values into the range of 0–1 according to

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x denotes the original feature value, x_{min} and x_{max} represent the minimum and maximum values of the feature, respectively, and x' denotes the normalized value. Although XGBoost is generally insensitive to feature scaling, Min-Max normalization was applied to ensure consistent feature ranges across all evaluated machine learning algorithms, particularly Logistic Regression, thereby enabling a fair comparative evaluation and improving numerical stability during model optimization.

To identify the most appropriate predictive model, four supervised machine learning algorithms were comparatively evaluated, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). Logistic Regression was selected as a representative linear classifier, Decision Tree as an interpretable tree-based model, Random Forest as a robust ensemble learning algorithm, and XGBoost as an advanced gradient boosting algorithm widely recognized for its high predictive performance on structured clinical datasets.

Prior to model development, the dataset was divided into training (72%) and testing (28%) subsets using stratified sampling to preserve the original class distribution. The training subset was exclusively used for XGBoost hyperparameter optimization, while the testing subset was reserved for final performance visualization using the confusion matrix and ROC curve. The XGBoost classifier was initialized using the binary:logistic objective function, logloss as the evaluation metric, and random_state = 42 to ensure reproducibility. Hyperparameter optimization was subsequently performed using RandomizedSearchCV, which evaluated 30 randomly sampled hyperparameter combinations under 10-fold cross-validation using ROC-AUC as the optimization criterion. The predefined search space included the number of boosting trees (n_estimators = 100, 200, 300, and 400), maximum tree depth (max_depth = 3–6), learning rate (learning_rate = 0.01, 0.05, 0.10, and 0.20), minimum loss reduction (gamma = 0, 0.10, and 0.30), row subsampling ratio (subsample = 0.70, 0.80, and 0.90), and column sampling ratio (colsample_bytree = 0.70, 0.80, and 0.90). The optimal hyperparameter configuration obtained from this optimization process is presented in Table 2.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

Table 2. Best Hyperparameter Configuration Obtained Using RandomizedSearchCV

Parameter	Optimal Value	Description
n_estimators	100	Number of boosting trees
max_depth	4	Maximum depth of each decision tree
learning_rate	0.05	Learning rate controlling the contribution of each tree
subsample	0.90	Fraction of training samples used to build each tree
colsample_bytree	0.70	Fraction of predictor variables randomly selected for each tree
gamma	0.10	Minimum loss reduction required to create a new split

The optimized hyperparameter configuration indicates that a moderately complex XGBoost model was sufficient to capture the nonlinear relationships among clinicopathological variables while maintaining good generalization capability. Specifically, a relatively low learning rate (0.05) enables gradual learning during the boosting process, whereas subsampling (0.90) and column sampling (0.70) introduce randomness that reduces overfitting and improves model robustness. Furthermore, a gamma value of 0.10 prevents unnecessary tree growth by requiring a minimum reduction in loss before additional splits are created. The optimized model was subsequently adopted as the final XGBoost classifier for comparative evaluation and SHAP-based explainability analysis.

After hyperparameter optimization, the predictive performance of all four machine learning algorithms was comparatively evaluated using 10-fold Stratified Cross-Validation, which preserves the class distribution across all folds and provides a reliable estimate of model generalization performance. In each iteration, nine folds were used for model training and one fold for testing until every fold had served as the testing set exactly once. Model performance was assessed using Accuracy, Precision, Recall, F1-Score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). To evaluate both predictive performance and model stability, the mean and standard deviation of each evaluation metric across the ten folds were reported. The best-performing classifier was subsequently evaluated using a confusion matrix and ROC curve to provide a more comprehensive assessment of classification performance.

After identifying the best-performing classifier, SHapley Additive exPlanations (SHAP) were employed to improve model transparency and interpretability. SHAP, introduced by Lundberg and Lee, is based on Shapley values from cooperative game theory and quantifies the contribution of each feature to individual model predictions. SHAP analysis was performed only on the best-performing model (XGBoost). Global model interpretation was conducted using SHAP beeswarm plots to identify the clinicopathological variables that most strongly influenced recurrence prediction across the entire dataset. Local interpretation was subsequently performed using SHAP waterfall plots, which explain how individual feature contributions increase or decrease the predicted recurrence risk for specific patients. By integrating comparative machine learning evaluation with both global and local explainability, the proposed framework provides not only accurate recurrence prediction but also transparent and clinically interpretable evidence to support personalized decision-making in thyroid cancer management.

The proposed framework was implemented in Python using the Pandas, NumPy, Scikit-learn, XGBoost, SHAP, Matplotlib, and Seaborn libraries.

RESULT

Prior to predictive modeling, feature selection was performed to identify the most relevant clinicopathological variables associated with thyroid cancer recurrence. The selection process combined correlation analysis to evaluate the strength of association between each predictor and the target variable with Chi-Square testing to assess statistical significance. Variables exhibiting very weak correlation ($|r| < 0.10$) together with non-significant Chi-Square association ($p \geq 0.05$) were excluded from subsequent model development. The complete feature selection results are presented in Table 3.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

Table 3. Feature Selection Results Based on Correlation Analysis and Chi-Square Test

Feature	Correlation	Chi2 Score	p-value	Correlation Strength	Statistical Significance	Decision
N	0.632323	206.909723	6.487838e-47	Strong	Significant	Selected
Stage	0.449137	189.759689	3.587451e-43	Moderate	Significant	Selected
Age	0.258897	143.510002	4.546977e-33	Weak	Significant	Selected
Response	0.708957	102.677741	3.943506e-24	Very Strong	Significant	Selected
T	0.556201	96.849704	7.479471e-23	Strong	Significant	Selected
Risk	-0.733376	54.262695	1.754003e-13	Very Strong	Significant	Selected
M	0.354360	45.833333	1.287539e-11	Moderate	Significant	Selected
Smoking	0.333243	37.090915	1.127478e-09	Moderate	Significant	Selected
Gender	0.328189	33.604934	6.752124e-09	Moderate	Significant	Selected
Focality	-0.383776	20.030655	7.621058e-06	Moderate	Significant	Selected
Hx Radiothreapy	0.174407	11.437056	7.199376e-04	Weak	Significant	Selected
Hx Smoking	0.136073	6.573165	1.035277e-02	Weak	Significant	Selected
Adenopathy	-0.182530	5.979275	1.447495e-02	Weak	Significant	Selected
Physical Examination	-0.131801	4.722485	2.977069e-02	Weak	Significant	Selected
Thyroid Function	0.067758	0.357936	5.496550e-01	Very Weak	Not Significant	Excluded
Pathology	0.003272	0.001270	9.715669e-01	Very Weak	Not Significant	Excluded

As shown in Table 3, fourteen of the sixteen clinicopathological variables satisfied the predefined selection criteria and were retained for predictive modeling. Among these variables, Response, Risk, N Stage, T Stage, and Clinical Stage exhibited the strongest associations with thyroid cancer recurrence, as indicated by both high correlation coefficients and highly significant Chi-Square statistics ($p < 0.001$). In contrast, Thyroid Function and Pathology showed very weak correlations with the target variable and non-significant Chi-Square associations ($p \geq 0.05$). Consequently, these two variables were excluded from further analysis, while the remaining fourteen features were used for model development and evaluation.

The predictive performance of four machine learning algorithms, namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), was comparatively evaluated using 10-fold stratified cross-validation. Model performance was assessed using Accuracy, Precision, Recall, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). For each evaluation metric, the mean and standard deviation across the ten folds were calculated to evaluate both predictive performance and model stability. The comparative results are presented in Table 4.

Table 4. Performance Comparison of Machine Learning Models (Mean ± SD)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.903 ± 0.038	0.873 ± 0.095	0.777 ± 0.074	0.819 ± 0.069	0.952 ± 0.043
Decision Tree	0.937 ± 0.044	0.880 ± 0.100	0.914 ± 0.121	0.890 ± 0.082	0.930 ± 0.061
Random Forest	0.958 ± 0.052	0.947 ± 0.101	0.905 ± 0.118	0.921 ± 0.096	0.983 ± 0.029
XGBoost	0.958 ± 0.044	0.934 ± 0.083	0.914 ± 0.099	0.922 ± 0.083	0.986 ± 0.025

As presented in Table 4, XGBoost and Random Forest achieved the highest mean accuracy (95.8%). However, XGBoost obtained the highest ROC-AUC (0.986 ± 0.025) and F1-score (0.922 ± 0.083), indicating superior discriminative ability while maintaining a balanced trade-off between precision and recall. Logistic Regression produced the lowest predictive performance among the evaluated models, whereas Decision Tree outperformed Logistic Regression but remained inferior to the ensemble-based methods. Overall, the results demonstrate that ensemble learning approaches consistently outperformed the conventional linear classifier on this dataset. Based on its superior overall performance, particularly in terms of ROC-AUC and F1-score, XGBoost was selected as the final prediction model for subsequent explainability analysis.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

To further evaluate the classification performance of the selected model, a confusion matrix was generated using the testing dataset. As shown in Figure 2, the majority of recurrence and non-recurrence cases were correctly classified, with relatively few false-positive and false-negative predictions.

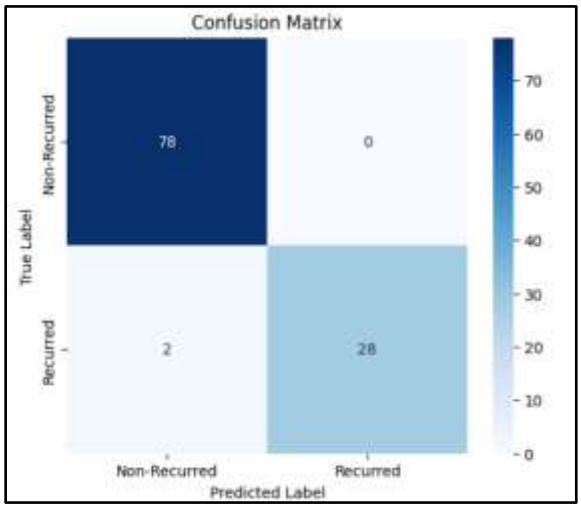

Figure 2. Confusion Matrix of the XGBoost Model

The discriminative ability of the selected model was further assessed using the Receiver Operating Characteristic (ROC) curve. As illustrated in Figure 3, the ROC curve is positioned close to the upper-left corner of the graph, corresponding to an ROC-AUC value of approximately 0.99, indicating excellent discrimination between recurrence and non-recurrence cases across different classification thresholds.

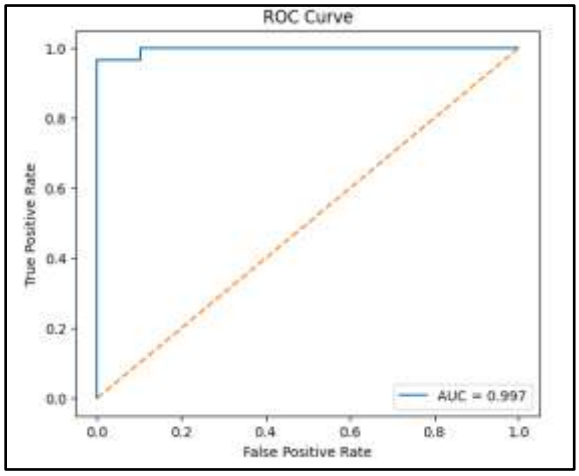

Figure 3. ROC Curve of the Best-Performing XGBoost Model

To improve model interpretability, SHapley Additive exPlanations (SHAP) were applied to the best-performing XGBoost classifier. The global explanation generated by the SHAP summary plot is presented in Figure 4. Response was identified as the most influential predictor, followed by Risk, Age, N Stage, and T Stage. The beeswarm plot further illustrates how individual feature values influence model predictions. Higher values of Response are generally associated with positive SHAP values, indicating an increased predicted probability of recurrence, whereas lower values tend to decrease the prediction. Conversely, higher Risk values are predominantly associated with negative SHAP values, while lower risk levels

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

contribute positively to the prediction, reflecting the encoded category ordering used in the dataset. Overall, the wide dispersion of SHAP values demonstrates that the influence of clinicopathological features varies across patients, confirming that the XGBoost model captures complex nonlinear interactions between predictors.

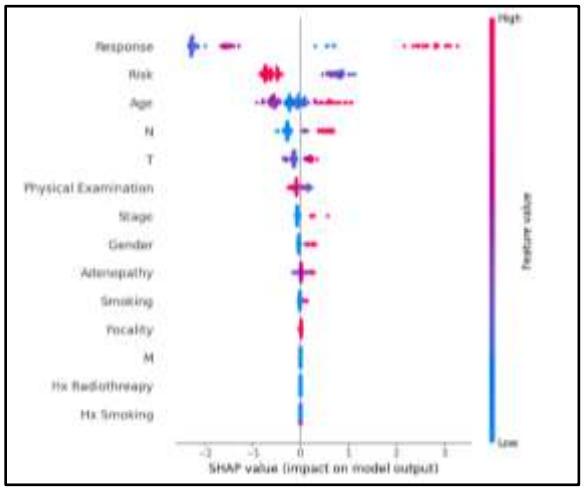

Figure 4. Global SHAP Beeswarm Plot of the Best-Performing XGBoost Model

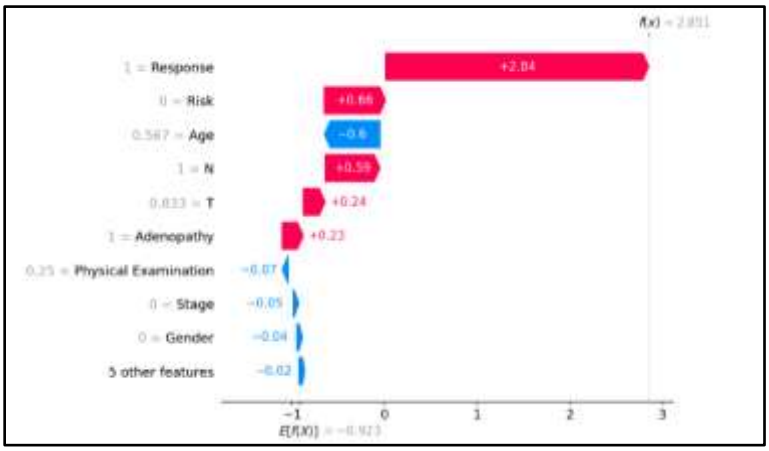

Figure 5. Local SHAP Waterfall Plot for an Individual Patient

To complement the global interpretation, local explainability was performed using a SHAP waterfall plot. As illustrated in Figure 5, the prediction for an individual patient is decomposed into the contribution of each clinicopathological feature relative to the model's baseline prediction. The model produced a final prediction score of 2.851, substantially higher than the baseline value of −0.923, indicating a high predicted probability of thyroid cancer recurrence. Among all variables, Response was the dominant contributor, increasing the prediction by 2.84, followed by Risk (+0.66), N Stage (+0.59), T Stage (+0.24), and Adenopathy (+0.23). In contrast, Age (−0.60), Physical Examination (−0.07), Stage (−0.05), and Gender (−0.04) reduced the predicted recurrence risk. These patient-specific explanations demonstrate how individual clinicopathological characteristics collectively influence the final prediction, thereby improving model transparency and supporting clinically interpretable decision-making.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

DISCUSSIONS

The comparative evaluation demonstrated that XGBoost achieved the best overall predictive performance among the four evaluated machine learning algorithms. Although Random Forest achieved the same mean accuracy (95.8%), XGBoost obtained the highest ROC-AUC (0.986 ± 0.025) and F1-score (0.922 ± 0.083), indicating superior discriminative ability while maintaining a better balance between precision and recall. Logistic Regression produced the lowest predictive performance among all evaluated classifiers, suggesting that the relationship between clinicopathological variables and thyroid cancer recurrence cannot be adequately represented by a linear decision boundary. Decision Tree outperformed Logistic Regression but remained inferior to the ensemble-based approaches. These findings demonstrate that ensemble learning algorithms, particularly XGBoost, are more capable of capturing the complex nonlinear interactions among clinicopathological variables than conventional linear or single-tree classifiers.

The superior performance of XGBoost can be attributed to its gradient boosting mechanism, which sequentially constructs decision trees to minimize prediction errors while incorporating regularization techniques that improve model generalization. Furthermore, hyperparameter optimization using RandomizedSearchCV enabled the selection of an appropriate model configuration before comparative evaluation, thereby enhancing predictive performance. Unlike conventional machine learning algorithms, XGBoost effectively models nonlinear relationships and interactions among demographic, pathological, and treatment-related variables that commonly occur in clinical datasets. These characteristics are reflected in the superior ROC-AUC and F1-score achieved by XGBoost, confirming its suitability for thyroid cancer recurrence prediction.

The feature selection procedure also contributed to improving the transparency and reproducibility of the proposed framework. Rather than relying solely on correlation analysis, this study combined correlation strength assessment with Chi-Square statistical testing to evaluate both the strength and statistical significance of each predictor. Based on predefined selection criteria, only Thyroid Function and Pathology were excluded because they exhibited both very weak correlation with recurrence ($|r| < 0.10$) and non-significant Chi-Square associations ($p \geq 0.05$). The remaining fourteen clinicopathological variables were retained for model development, reducing unnecessary model complexity while preserving statistically and clinically relevant predictors.

While predictive performance is essential, successful clinical implementation also requires transparent and interpretable predictions. Machine learning models with high predictive accuracy but limited interpretability are often difficult for clinicians to trust because the rationale underlying individual predictions remains unclear. To address this challenge, SHapley Additive exPlanations (SHAP) were incorporated into the proposed framework. Unlike conventional feature importance methods that merely rank variables according to their overall contribution, SHAP provides both global and local explanations, enabling interpretation of model behavior at the population level as well as for individual patients. This dual-level explainability enhances model transparency and supports evidence-based clinical decision-making by revealing how each clinicopathological variable contributes to the predicted recurrence risk.

The SHAP analysis consistently identified Response and Risk as the most influential predictors of thyroid cancer recurrence, followed by Age, N Stage, and T Stage. These findings are clinically meaningful because treatment response directly reflects the effectiveness of therapeutic interventions, whereas clinical risk classification summarizes multiple prognostic characteristics associated with disease progression. Likewise, lymph node involvement (N Stage) and primary tumor extent (T Stage) are well-established prognostic indicators of recurrence. The SHAP beeswarm plot further demonstrated that both the magnitude and direction of feature contributions varied across patients, highlighting the complex nonlinear relationships learned by the XGBoost model. Moreover, the SHAP waterfall plot provided patient-specific explanations by decomposing individual predictions into feature-level contributions, illustrating how clinicopathological variables jointly increased or decreased the predicted recurrence risk. The consistency

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

between these explanations and established clinical knowledge suggests that the proposed framework generates predictions that are not only accurate but also clinically plausible and interpretable.

The findings of this study are generally consistent with previous research demonstrating the effectiveness of machine learning for thyroid cancer recurrence prediction. Gong et al. reported that ensemble learning methods achieved high predictive accuracy but did not provide interpretable explanations for model predictions (Gong et al., 2021). Schindele et al. and Hanani et al. incorporated SHAP to improve model transparency (Hanani et al., 2025; Schindele et al., 2025), whereas Hu et al. and Redlich et al. further demonstrated the usefulness of explainable machine learning in clinical prediction tasks (Hu et al., 2026; Redlich et al., 2026). However, most previous studies primarily emphasized predictive performance or applied explainability without systematically comparing multiple machine learning algorithms before selecting the final predictive model. In contrast, the present study integrates comparative evaluation of multiple classifiers, objective selection of the best-performing model, and comprehensive SHAP interpretation within a unified Explainable Machine Learning framework. This integrated approach enables both objective model selection and transparent interpretation at the global and individual patient levels, thereby strengthening the reliability and clinical applicability of the proposed framework.

From a clinical perspective, the proposed framework has the potential to support personalized thyroid cancer management. By combining accurate recurrence prediction with interpretable explanations, clinicians can identify high-risk patients while simultaneously understanding the clinicopathological factors driving each prediction. Such information may facilitate individualized follow-up strategies, optimize surveillance planning, and improve communication between clinicians and patients. Furthermore, transparent prediction models are more likely to gain acceptance in clinical practice because their recommendations can be interpreted and validated using established medical knowledge rather than functioning solely as black-box algorithms.

Despite these promising findings, several limitations should be acknowledged. First, the proposed framework was developed and evaluated using a single publicly available dataset comprising only 383 patient records, which may limit the generalizability of the findings. Second, external validation using independent datasets from different healthcare institutions was not conducted and should be considered in future studies to evaluate model robustness across more diverse patient populations. Third, although four widely used machine learning algorithms were comparatively evaluated, future research may include additional state-of-the-art methods, such as LightGBM and CatBoost, to further investigate potential improvements in predictive performance. Finally, statistical significance testing, such as McNemar's test or the Wilcoxon signed-rank test, was not performed to determine whether the observed performance differences between classifiers were statistically significant. Incorporating these statistical analyses in future studies would provide stronger evidence regarding comparative model performance.

Overall, the findings demonstrate that the proposed Explainable Machine Learning framework effectively combines comparative machine learning evaluation with SHAP-based explainability to produce accurate, transparent, and clinically interpretable thyroid cancer recurrence predictions. Rather than focusing solely on predictive performance, the framework provides objective model selection together with both global and patient-level explanations, thereby enhancing the trustworthiness of machine learning predictions and supporting the development of reliable clinical decision support systems for thyroid cancer management.

CONCLUSION

This study proposed an Explainable Machine Learning framework for thyroid cancer recurrence prediction that integrates transparent feature selection, comparative machine learning evaluation, optimized XGBoost modeling, and SHapley Additive exPlanations (SHAP) within a unified workflow. The proposed framework systematically combines data preprocessing, objective model selection, predictive modeling,

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

and multi-level interpretation to provide both accurate recurrence prediction and clinically interpretable explanations.

The experimental results demonstrated that XGBoost achieved the best overall predictive performance among the evaluated classifiers, outperforming Logistic Regression, Decision Tree, and Random Forest in terms of ROC-AUC and F1-score while achieving comparable accuracy. The comparative evaluation strategy enabled the objective selection of the final prediction model rather than assuming XGBoost as the optimal classifier. Furthermore, SHAP successfully explained the prediction process through both global and local interpretations, consistently identifying Response, Risk, Age, N Stage, and T Stage as the most influential predictors of thyroid cancer recurrence.

Accordingly, the findings answer the research questions by demonstrating that the proposed framework can accurately predict thyroid cancer recurrence while simultaneously providing clinically interpretable explanations of the underlying prediction process. By integrating comparative model evaluation with explainable artificial intelligence, the framework improves both the transparency and trustworthiness of machine learning predictions, thereby increasing their potential applicability in evidence-based clinical decision support.

This study contributes to the growing field of Explainable Artificial Intelligence in healthcare by presenting a reproducible framework that objectively identifies the most suitable prediction model while providing transparent global and patient-level explanations. Such integration helps bridge the gap between predictive performance and interpretability, facilitating the adoption of trustworthy machine learning models in clinical practice.

Despite these promising findings, several limitations should be acknowledged. The proposed framework was evaluated using a single publicly available dataset, and external validation was not performed. Future research should therefore validate the framework using larger multicenter datasets, incorporate statistical significance testing to compare competing models, and investigate additional state-of-the-art machine learning algorithms together with multimodal clinical data to further improve predictive performance and generalizability.

REFERENCES

Abkar, A., Mehrabi, M., Golabpour, A., & Shayegan, M. A. (2026). Designing an explainable algorithm based on XGBoost and genetic algorithm for predicting hospitalization needs of COVID-19 patients. *Scientific Reports 2026 16:1*, *16*(1), 10210-. https://doi.org/10.1038/s41598-026-40120-6

Arista, R. D., Karima, K., Anugrah, M. F., Widyastuti, P., & Triani, E. (2023). Thyroid Cancer : an Overview of Epidemiology, Risk Factor, and Treatment. *Lombok Medical Journal*, *2*(3), 90–96. https://doi.org/10.29303/LMJ.V2I2.2791

Aulia, N., Kasprata, H. N., Priyahita, P. D., Syahla, T., & Triani, E. (2023). Clinical Diagnosis and Management of Thyroid Cancer. *Jurnal Kedokteran (Unram Medical Journal)*, *12*(3), 240–246. https://doi.org/10.29303/JK.V12I3.4493

Carmona, P., Dwekat, A., & Mardawi, Z. (2022). No more black boxes! Explaining the predictions of a machine learning XGBoost classifier algorithm in business failure. *Research in International Business and Finance*, *61*, 101649. https://doi.org/10.1016/J.RIBAF.2022.101649

Feng, Y., Hu, Y., Li, T., Li, M., & Zhang, M. (2026). XGBoost-Cox modeling with SHAP analysis for survival prediction in ovarian cancer patients: a retrospective cohort study. *BMC Cancer 2026 26:1*, *26*(1), 573-. https://doi.org/10.1186/S12885-026-15921-7

Gong, X., Zheng, B., Xu, G., Chen, H., & Chen, C. (2021). Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer. *Journal of Thoracic Disease*, *13*(11), 6240. https://doi.org/10.21037/JTD-21-1107

Gunasekara, N., Pfahringer, · Bernhard, Gomes, · Heitor, Bifet, · Albert, Pfahringer, B., Gomes, H., Bifet, A., & Nz, A. (2024). Gradient boosted trees for evolving data streams. *Machine Learning 2024 113:5*,

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 3, July 2026
https://doi.org/ 10.47709/cnahpc.v8i3.8894

Submitted : 2026-06-19
Accepted : 2026-07-01
Published : 2026-07-01

113(5), 3325–3352. https://doi.org/10.1007/S10994-024-06517-Y

Hanani, A. A., Donmez, T. B., Kutlu, M., & Mansour, M. (2025). Predicting thyroid cancer recurrence using supervised CatBoost A SHAP-based explainable AI approach. *Medicine (United States)*, *104*(22). https://doi.org/10.1097/MD.0000000000042667

Hu, G., Huang, B., Cai, L., Zhang, Y., Zhang, Y., Liu, Y., & Wu, G. (2026). Machine learning prediction of thyroid cancer recurrence for early screening and clinical decision pathways: a retrospective cohort study. *Discover Oncology 2026 17:1*, *17*(1), 239-. https://doi.org/10.1007/S12672-025-04293-2

Istiwana, A. P., Sani, R. R., & Pramudi, Y. T. C. (2026). Pendekatan Explainable Machine Learning Untuk Analisis Faktor Drop Out Mahasiswa Menggunakan XGBoost. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *11*(1), 1074–1083. https://doi.org/10.36341/RABIT.V11I1.7218

Jiang, H., Ji, L., Zhu, L., Wang, H., & Mao, F. (2025). XGBoost model for predicting erectile dysfunction risk after radical prostatectomy: development and validation using machine learning. *Discover Oncology 2025 16:1*, *16*(1), 810-. https://doi.org/10.1007/S12672-025-02685-Y

Nugraha, W., Sabaruddin, R., Abdul Rahman Saleh No, J., Belitung Laut, B., Pontianak Tenggara, K., Pontianak, K., & Barat, K. (2025). Evaluasi Performa Algoritma Klasifikasi dalam Prediksi Kekambuhan Kanker Tiroid Pasca Terapi RAI: Studi Kasus Dataset RAI Therapy. *Teknik: Jurnal Ilmu Teknik Dan Informatika*, *5*(1), 27–35. https://doi.org/10.51903/TEKNIK.V5I1.717

Patimah, P., Zulpan, Z., Alfansuri, D. U., Munawaroh, E., & Ilyas, M. (2025). Memahami Dan Menerapkan Uji Korelasi Dalam Analisis Data Penelitian Pendidikan. *Journal Education Innovation (JEI)*, *3*(4), 740–752. https://doi.org/10.65474/F7VD8Y11

Ramadhan, M. E., & Zeniarja, J. (2025). Implementation of Deep Transfer Learning and Explainable AI in Skin Cancer Classification. *Sistemasi: Jurnal Sistem Informasi*, *14*(5), 2266–2279. https://doi.org/10.32520/STMSI.V14I5.5425

Redlich, A., Pfaehler, E., Kunstreich, M., Schmutz, M., Lapa, C., & Kuhlen, M. (2026). Machine Learning Prediction of Recurrence in Pediatric Thyroid Cancer: Malignant Endocrine Tumors Cohort Analysis Using XGBoost and SHAP. *The Journal of Clinical Endocrinology & Metabolism*, *111*(3), e844–e852. https://doi.org/10.1210/CLINEM/DGAF487

Schindele, A., Krebold, A., Heiß, U., Nimptsch, K., Pfaehler, E., Berr, C., Bundschuh, R. A., Wendler, T., Kertels, O., Tran-Gia, J., Pfob, C. H., & Lapa, C. (2025). Interpretable machine learning for thyroid cancer recurrence predicton: Leveraging XGBoost and SHAP analysis. *European Journal of Radiology*, *186*, 112049. https://doi.org/10.1016/J.EJRAD.2025.112049

Takwim, A., & Sulaeman, H. (2025). Explainable AI Sebagai Solusi Black Box Effect Dalam Kecerdasan Buatan. *SEMINAR TEKNOLOGI MAJALENGKA (STIMA)*, *9*, 622–627. https://doi.org/10.48550/ARXIV.2409.00265

* Corresponding author