The Journal of Clinical Endocrinology & Metabolism, 2026, **111**, e844–e852
https://doi.org/10.1210/clinem/dgaf487
Advance access publication 1 September 2025
Clinical Research Article

Machine Learning Prediction of Recurrence in Pediatric Thyroid Cancer: Malignant Endocrine Tumors Cohort Analysis Using XGBoost and SHAP

Antje Redlich,[1] Elisabeth Pfaehler,[2] Marina Kunstreich,[1,3] Maximilian Schmutz,[4,5] Constantin Lapa,[5,6] and Michaela Kuhlen[3,5]

[1]Department of Pediatrics, Pediatric Hematology/Oncology, Otto von Guericke-University, Magdeburg 39120, Germany
[2]Institute for Neuroscience and Medicine 4, INM-4, Forschungszentrum Jülich GmbH, Juelich 52428, Germany
[3]Pediatrics and Adolescent Medicine, Faculty of Medicine, University of Augsburg, Augsburg 86156, Germany
[4]Haematology and Oncology, Faculty of Medicine, University of Augsburg, Augsburg 86156, Germany
[5]Bavarian Cancer Research Centre, Augsburg 86156, Germany
[6]Nuclear Medicine, Faculty of Medicine, University of Augsburg, Augsburg 86156, Germany

Correspondence: Antje Redlich, MD, Department of Pediatrics, Pediatric Hematology/Oncology, Otto von Guericke-University, Leipziger Str. 44, Magdeburg D-39120, Germany. Email: antje.redlich@med.ovgu.de; or Michaela Kuhlen, MD, Pediatrics and Adolescent Medicine, Faculty of Medicine, University of Augsburg, Stenglinstr. 2, Augsburg 86156, Germany. Email: Michaela.Kuhlen@uk-augsburg.de.

Abstract

Context: Pediatric differentiated thyroid carcinoma (DTC) often presents with advanced disease but generally has excellent long-term survival. However, recurrence or failure to achieve remission remains relatively frequent, underscoring the need for improved early risk stratification.

Objective: To develop and evaluate an interpretable machine learning model for predicting recurrence or nonremission in pediatric DTC using routine clinical and biochemical variables.

Design and Setting: Retrospective analysis of 250 pediatric patients (aged <18 years) enrolled in the German Pediatric Oncology Hematology-Malignant Endocrine Tumors Registry (1997-2023). Inclusion required known age at diagnosis and ≥24 months of follow-up. The composite study endpoint was structural recurrence or failure to achieve remission within 24 months of initial therapy.

Methods: An extreme gradient boosting classifier was trained on 80% of the data, with the remaining 20% used as an independent test set. Model generalizability was assessed via 50 randomized stratified train-validation splits of the training dataset. SHapley Additive exPlanations (SHAP) were used to interpret feature contributions.

Results: The final model achieved an area under the receiver operating characteristic curve (AUROC) of 0.86 on the independent test set. Across 50 validation splits, the mean AUROC was 0.82 (SD ± 0.05), sensitivity 0.81 (SD ± 0.09), and specificity 0.64 (SD ± 0.06). SHAP analysis identified younger age at diagnosis (<10 years), elevated postoperative thyroglobulin levels, and distant metastases as the most influential predictors.

Conclusion: This interpretable machine learning model reliably predicts early recurrence or nonremission in pediatric DTC and may complement current risk stratification systems to support personalized, risk-adapted treatment decisions.

Key Words: children and adolescents, differentiated thyroid carcinoma, XGBoost, SHAP, prediction

Abbreviations: AUROC, area under the receiver operating characteristic curve; DTC, differentiated thyroid carcinoma; GPOH-MET, German Pediatric Oncology Hematology-Malignant Endocrine Tumors; ML, machine learning; SHAP, SHapley Additive exPlanations; SMOTE, synthetic minority oversampling; TG, thyroglobulin; XGBoost, extreme gradient boosting.

Differentiated thyroid carcinoma (DTC) represents the most prevalent endocrine malignancy in the pediatric population, accounting for the vast majority of thyroid carcinomas in individuals under the age of 18 (1, 2). Its incidence, while still relatively low compared to adult populations, has shown a steady upward trend over the past decades (1, 3, 4). Pediatric DTC exhibits several biological and clinical differences from its adult counterpart. Most notably, children frequently present with advanced disease, including cervical lymph node involvement in approximately 60% to 80% of cases and distant metastases, particularly pulmonary, in 10% to 30% of patients at diagnosis (5, 6). Despite these aggressive features, overall survival remains excellent, often exceeding 95% at 10 years (7). However, the high rates of recurrence or persistent disease—reported in 20% to 40% of cases—constitute a significant clinical concern (8).

Given the excellent long-term prognosis, current management strategies for pediatric DTC aim to balance effective tumor control with minimization of treatment-related morbidity. In this context, accurate risk stratification is essential. Both the American Thyroid Association and the European Thyroid Association have issued pediatric-specific guidelines emphasizing the need for individualized, risk-adapted therapeutic approaches (9, 10). These include

Received: 2 July 2025. **Editorial Decision:** 23 August 2025. **Corrected and Typeset:** 9 September 2025

decisions on the extent of thyroid surgery and the use of postoperative radioactive iodine therapy. In particular, the potential to safely reduce radioactive iodine use in low-risk patients has garnered increasing attention, necessitating robust tools to predict the risk of recurrence or failure to achieve remission (11).

Traditional prognostic indicators include age at diagnosis, tumor size, presence of lymph node and distant metastases, extrathyroidal extension, and postoperative thyroglobulin (TG) levels (12). While informative, these individual variables often lack the predictive power required for personalized management. Moreover, the interplay between multiple clinical and pathological features is complex, underscoring the limitations of conventional statistical models.

In recent years, machine learning (ML) has emerged as a transformative approach in clinical oncology (13). ML algorithms can identify nonlinear relationships and complex feature interactions in large datasets, thereby offering superior predictive accuracy compared to traditional methods. Among these, extreme gradient boosting (XGBoost) has gained prominence for its high performance, scalability, and ability to handle heterogeneous data. XGBoost is an ensemble method that builds decision trees sequentially, optimizing for predictive accuracy by minimizing a specified loss function with regularization to prevent overfitting (14).

To enhance interpretability, SHapley Additive exPlanations (SHAP) have been developed as a model-agnostic framework that quantifies the contribution of each input feature to a specific prediction (15). SHAP values are based on cooperative game theory and provide both global and local interpretability, making them particularly useful in clinical applications where understanding the rationale behind a model's prediction is critical for trust and adoption.

In the present study, we aimed to evaluate the feasibility and clinical utility of an ML model based on XGBoost, complemented by SHAP analysis, to predict recurrence or failure to achieve remission within 24 months in a cohort of pediatric DTC patients. Using a well-characterized, multicenter registry dataset from the German Society for Pediatric Oncology and Hematology-Malignant Endocrine Tumors (GPOH-MET) Registry (16), we sought to identify key prognostic factors and assess the model's potential to inform risk-adapted management strategies in pediatric patients with DTC.

Methods

Study Design and Patient Cohort

The study population includes 562 pediatric patients (aged <18 years) diagnosed with DTC and enrolled in the German GPOH-MET Registry between 1997 and 2023. For the present analysis, patients were included only if age at diagnosis was available and if a minimum follow-up of 24 months was documented. Patients without these mandatory data elements were excluded to ensure consistency in outcome classification. Following this filtering, a final cohort of 250 patients was retained for ML analysis.

Ethical Approval and Patient Consent

This study was conducted in accordance with the principles outlined in the Declaration of Helsinki. The GPOH-MET Registry received ethical approval by the ethics committees of the University of Luebeck (IRB 97125) and Otto von Guericke University Magdeburg (IRB 174/12 and 52/22), Germany. All data were pseudonymized prior to analysis. Written informed consent was obtained at the time of original inclusion into the GPOH-MET Registry from patients and/or their legal guardians, in compliance with applicable national regulations.

Outcome Definition

The primary endpoint was defined as either (1) recurrence of newly detected structural disease (identified via imaging modalities) or (2) failure to achieve complete remission within 24 months of initial therapy. Complete remission was assessed according to the criteria established by Tuttle et al, including imaging findings and biochemical parameters, notably thyroglobulin levels (17). Follow-up time was standardized to 24 months postinitial treatment, and outcomes were assessed within this fixed interval.

Feature Selection and Variable Definitions

The following features were included based on established or suspected prognostic relevance in pediatric DTC:

- Tumor size (cm): maximum tumor diameter as reported in histopathology.
- Age at diagnosis: categorized as <10 years vs ≥10 years based on prior prognostic stratification literature (16).
- pN and cM status: categorized as binary variables (presence vs absence of regional or distant metastasis).
- Postoperative stimulated TG level (ng/mL): collected prior to initial radioactive iodine therapy.
- Year of diagnosis: categorized into three periods: 1997-2005 (1), 2006-2014 (2), and 2015-2023 (3).
- Histological subtypes (eg, papillary vs follicular DTC) and histopathological features (eg, extrathyroidal extension) were excluded from model training due to negligible predictive contribution in exploratory analysis.

Handling of Missing Data

To maximize dataset utility, the following imputation strategy was applied:

- Patients with missing age at diagnosis were excluded.
- Missing tumor size (available in 212 of 250 cases; 38 missing) and postoperative TG values (available in 159 of 250 cases; 91 missing) were imputed using the median value calculated across the entire cohort.
- Missing entries for pN and cM were assumed absent and coded accordingly (ie, 0).

ML Model

We used the XGBClassifier implementation of XGBoost to develop a supervised classification model for predicting the defined event outcome. The dataset was randomly split into a training set (80%) and an independent test set (20%) using stratified sampling to preserve the class distribution.

To address class imbalance (23.6% event rate), synthetic minority oversampling [SMOTE (18)] was applied exclusively to the training subset after the initial train-test split. A SMOTE sampling strategy of 0.5 was used, increasing the number of minority class samples to 50% of the majority class. Importantly, when the training data were further

divided into training and validation folds during cross-validation, SMOTE was applied only within each training fold to prevent data leakage into validation sets.

Baseline characteristics of the training and test sets were compared and showed no statistically significant differences across key clinical variables, including age, tumor size, staging, and outcome distribution, supporting the internal validity of the data split.

XGBoost is based on decision tree ensembles and, unlike many linear or kernel-based-models, does not require feature standardization or normalization. Therefore, no feature scaling was applied prior to model training.

Hyperparameter tuning was conducted using randomized grid search with 5-fold cross-validation applied exclusively to the training data. The area under the precision-recall curve was used to address the primary optimization objective, reflecting the class imbalance and the clinical priority of minimizing false negatives. The following parameter ranges were explored:

- Learning rate: 0.01 to 0.2 (optimal: 0.01)
- Maximum tree depth: 3 to 6 (optimal: 6)
- Gamma (minimum loss reduction): 0 to 1 (optimal: 0.3)

Model Validation and Statistical Analysis

To assess the generalizability of the model, performance was evaluated using 50 stratified shuffle-splits of the training data, implemented via the StratifiedShuffleSplit method. In each split, 20% of the training data was reserved as a validation set. The following performance metrics were calculated for each split and are reported as mean values with SD:

- Area under the receiver operating characteristic curve (AUROC)
- Sensitivity (recall)
- Specificity
- Precision

To prioritize sensitivity—given the clinical importance of minimizing missed events—a fixed classification threshold of 0.2 was applied. This tradeoff was accepted to ensure high sensitivity, even at the cost of increased false-positive predictions.

Evaluation on Independent Test Data

Finally, the held-out test dataset, which was not involved in model training or validation, was used to evaluate the performance of the final trained model.

Model Interpretability

To interpret the influence of individual features on model predictions, SHAP were employed. SHAP summary plots and dependence plots were generated to visualize the marginal effect of key variables—including distant metastasis (cM), pN status, age category, tumor size, postoperative TG level, and period of diagnosis—on predicted event risk. These visualizations were produced separately for the training and independent test datasets to confirm that model interpretability was preserved across both internal and external evaluation settings.

All analyses were performed using Python 3.12 with the following libraries: scikit-learn, pandas, NumPy, XGBoost, SHAP, matplotlib, pyreadstat, and imbalanced-learn.

Results

Patient Characteristics

Of the 562 patients enrolled in the GPOH-MET Registry, 250 met the inclusion criteria for this analysis, which required availability of age at diagnosis and at least 24 months of follow-up. The median age at diagnosis was 13.1 years (range 3.6-17.9), with 74 male (29.6%) and 176 female (70.4%) patients.

Most patients (87.2%) were diagnosed with papillary thyroid carcinoma, while 9.2% had follicular thyroid carcinoma; the remainder included rare subtypes. The median tumor diameter was 2.0 cm (range 0.1-7.0 cm).

Postoperative stimulated TG levels ranged from 0 to 43 200 ng/mL, with a median of 9.0 ng/mL (mean 451.0 ng/mL, SD 3488.4 ng/mL). Distant metastases were documented in 60 patients (24.0%), and lymph node metastases were present in 160 patients (64.0%).

Demographic and clinical characteristics of the study cohort are presented in Table 1.

Model Performance

Following data preprocessing and SMOTE-based rebalancing of the training cohort, an XGBoost classification model was trained and evaluated. To assess model robustness and generalizability, we performed 50 randomized stratified train-validation splits, using 20% of the training data as validation

Table 1. Demographic and clinical characteristics of the study cohort

Characteristics	Number of patients
Sex, n (%)	
Female	176 (70.4)
Male	74 (29.6)
Age at diagnosis, years	
Mean	12.5
Median	13.1
Range	3.6-17.9
Diagnosis, n (%)	
PTC	218 (87.2)
FTC	23 (9.2)
DTC with low differentiated parts	3 (1.2)
DTC NOS	5 (2.0)
PTC + FTC	1 (0.4)
Lymph node metastases, n (%)	160 (64.0)
Distant metastases, n (%)	60 (24.0)
Event, n (%)	59 (23.6)
Follow-up, years	
Mean	6.8
Median	6.2
Range	2.0-20.6

Abbreviations: DTC, differentiated thyroid carcinoma; FTC, follicular thyroid carcinoma; NOS, not otherwise specified; PTC, papillary thyroid carcinoma.

in each iteration. The model demonstrated consistent performance across splits, with the following mean metrics:

- AUROC: 0.824 (SD ± 0.051)
- Sensitivity: 0.814 (SD ± 0.090)
- Specificity: 0.644 (SD ± 0.058)
- Precision: 0.421 (SD ± 0.042)

These cross-validation results are shown in Fig. 1.

Performance was then evaluated on the independent hold-out test set, which had not been used during model training, tuning, or resampling. On this test set, the final model achieved the following metrics (Fig. 2):

- AUROC: 0.857
- Sensitivity (recall): 1.000
- Specificity: 0.658
- Area under the precision-recall curve: 0.599

These results indicate strong generalizability, particularly in terms of sensitivity, which was prioritized during model calibration using a fixed classification threshold of 0.2.

Feature Importance and SHAP Analysis

To understand the contribution of individual features to model predictions, we performed both internal XGBoost feature importance analysis and SHAP. XGBoost's internal feature importance (Fig. 3) quantifies how often and how effectively each variable was used to split decision trees within the ensemble.

According to this measure, the most impactful variables were the presence of distant metastases, younger age at diagnosis, and lymph node involvement. Postoperative-stimulated TG levels, period of diagnosis, and tumor size contributed less frequently to decision splits but still played a role in overall model construction.

To complement the model's internal feature importance, we used SHAP to interpret both global and individual-level contributions of each variable to model predictions. SHAP summary plots (Fig. 4A and 4B) illustrate the direction and strength of each feature's impact on prediction output across all patients.

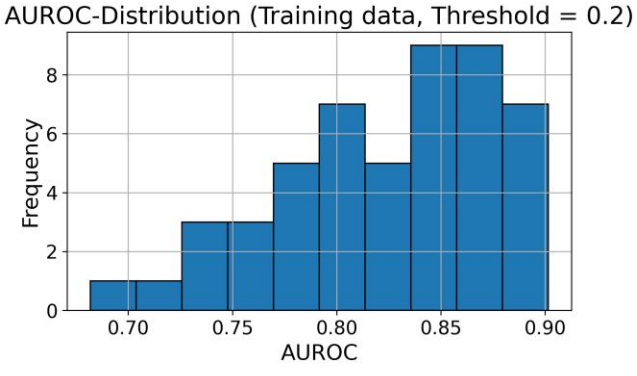

Figure 1. Distribution of AUROC values across 50 stratified train-validation splits. Histogram illustrating the consistency of the model's discriminative performance (AUROC) across 50 randomized stratified train-validation partitions of the training data. This demonstrates robustness and low variance in model performance across subsamples.

Abbreviation: AUROC, area under the receiver operating characteristic curve.

The presence of distant metastases, younger age at diagnosis (<10 years), and higher postoperative TG levels were associated with positive SHAP values—indicating increased risk for recurrence or failure to achieve remission. Conversely, absence of distant metastases, older age (≥10 years), and lower TG levels contributed negatively, pushing predictions toward remission (class 0). These relationships were consistent across both the training and independent test datasets, underscoring model stability and interpretability.

To evaluate overall feature impact, global SHAP importance plots (Fig. 5A and 5B) display the mean absolute SHAP values across patients. Distant metastases, postoperative TG levels, and age at diagnosis emerged as the most influential predictors in both datasets, followed by lymph node status, tumor size, and diagnosis period. While distant metastases showed the highest average SHAP value, they were present in only 24% of patients. In contrast, age and biochemical response (TG levels) influenced predictions across the entire cohort and therefore contributed consistently to individualized risk estimates. To further explore directional and nonlinear effects, we generated SHAP dependence plots for the training and independent test datasets (Fig. 6A-6F). These plots illustrate how variation in specific feature values affects the SHAP values and thereby the predicted probability of recurrence or nonremission:

- Age at diagnosis (A + B): patients diagnosed under the age of 10 years (encoded as 0) consistently exhibited high positive SHAP values, indicating strong contributions toward predicted recurrence or nonremission. In contrast, patients aged ≥10 years, encoded as 1, showed predominantly negative SHAP values, suggesting a protective association with older age.
- Presence of distant metastases (C + D): the SHAP dependence plots for distant metastases (encoded as 0 for absence and 1 for presence) demonstrate a clear directional impact on model predictions. In both the training (Fig. 6C) and independent test datasets (Fig. 6D), the presence of distant metastases (value = 1) is consistently associated with high positive SHAP values, indicating a strong contribution to increased predicted risk of recurrence or nonremission. In contrast, patients without distant metastases (value = 0) predominantly show negative SHAP values, suggesting a protective association. These results underscore the significant prognostic value of metastatic spread in the predictive model.
- Postoperative stimulated TG levels (log-transformed, E + F): a nonlinear relationship was observed between TG levels and SHAP values. Low TG levels (<10 ng/mL) were associated with negative SHAP values (indicating remission), whereas higher TG values were linked to progressively increasing SHAP values and predicted risk.

The shape and direction of these associations were highly consistent across both training and test datasets, underscoring the model's generalizability. These reinforce the clinical relevance of younger age, presence of distant metastases, and elevated postoperative TG levels as key contributors to predicted adverse outcomes in pediatric DTC.

SHAP dependence plots for postoperative TG levels revealed a steep rise in SHAP values between ~10 and 100 ng/mL, followed by a plateau at higher levels (Fig. 6C and 6D). This

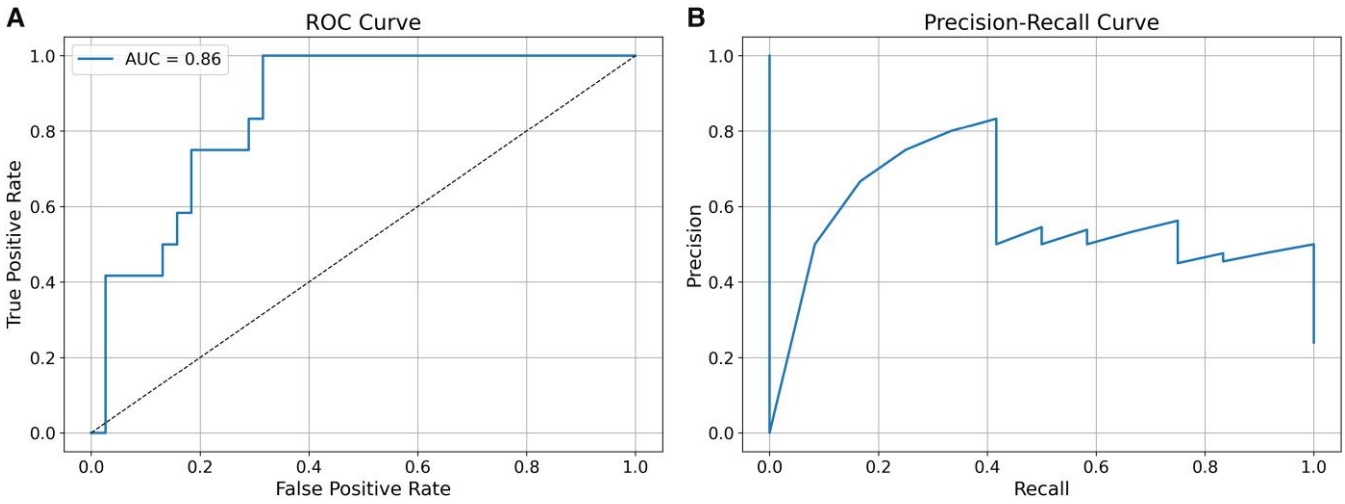

Figure 2. ROC and PR curves of the final model on the independent test set. (A) ROC curve: the model achieved an AUROC of 0.86, reflecting strong discriminative ability in distinguishing patients with vs without an event (recurrence or failure to achieve remission). (B) PR curve: this plot illustrates the tradeoff between precision and recall, particularly relevant in the context of class imbalance. The model maintains good precision across a broad range of recall values. Together, these plots demonstrate the generalizability and recall-oriented performance of the final extreme gradient boosting model on unseen data.

Abbreviations: AUROC, area under the receiver operating characteristic curve; PR, precision recall; ROC, receiver operating characteristic.

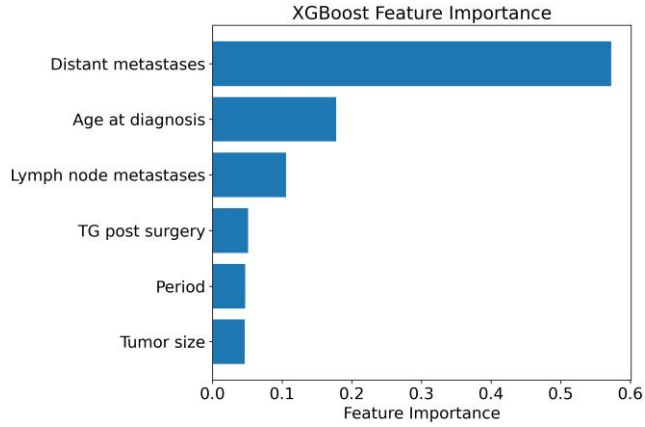

Figure 3. XGBoost feature importance of the final model. Bar plot showing the relative importance of each input feature based on its frequency and contribution to decision splits within the XGBoost ensemble. Distant metastases, age at diagnosis, and lymph node metastases were the most influential features in model construction.

Abbreviation: XGBoost, extreme gradient boosting.

pattern indicates that moderately elevated TG values contribute strongly to predicted risk, while further increases do not significantly change the model's output.

To further quantify TG's predictive relevance, we performed a threshold analysis using Youden's index, which identified an optimal postoperative TG cutoff of approximately 9.0 ng/mL (data not shown). At this threshold, the model achieved a sensitivity of 91.3% and specificity of 38.7%, supporting the clinical utility of TG in early risk stratification.

SHAP-based Individual Risk Explanation

To demonstrate individualized model interpretability, we visualized SHAP-based prediction decompositions for 2 hypothetical patients who were identical in all clinical features except for age at diagnosis (Fig. 7). Both patients had distant and lymph node metastases, a tumor size of 3 cm, postoperative TG of 10 ng/mL, and diagnosis during the same period.

The model predicted a recurrence probability of 43.0% for the patient aged ≥10 years, compared to 67.4% for the patient aged <10 years, illustrating how a single variable can meaningfully shift the individualized risk estimates. These SHAP waterfall plots further highlight how the same feature (eg, TG or distant metastases) can contribute differently depending on the overall feature context. For example, distant metastases had a greater influence in the older patient, emphasizing SHAP's ability to capture interaction effects and explain model behavior transparently.

Discussion

This study demonstrates that an interpretable ML model based on XGBoost can effectively predict early recurrence or failure to achieve remission in pediatric DTC. Despite the challenge of class imbalance—only 23.6% of patients in our cohort experienced the primary event—the model exhibited strong and consistent performance, achieving a mean AUROC of 0.82 across 50 stratified train-validation splits and maintaining robustness on an independent test dataset. Importantly, the model was trained exclusively on routinely collected clinical variables, emphasizing its translational potential.

We applied SHAP to enhance the model's interpretability and transparency. Unlike XGBoost's internal feature importance—which reflects how frequently a variable is used for tree-based splits—SHAP provides a directionally informative, individualized breakdown of each feature's influence on the model's output. In both global summary and patient-level analyses, the most influential variables included younger age at diagnosis, elevated postoperative TG levels, and the presence of distant metastases. Our findings support and expand upon existing risk stratification frameworks. While the American Thyroid Association guidelines emphasize distant metastases as a core risk determinant (9), our SHAP-based

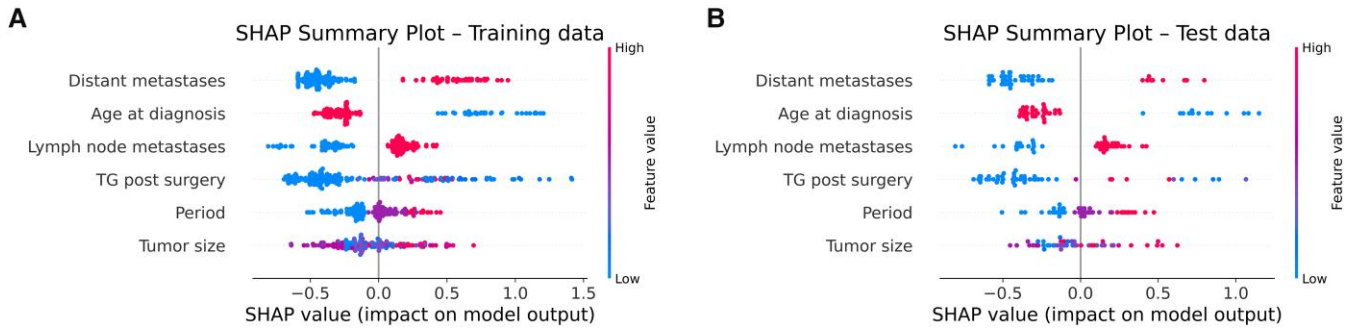

Figure 4. SHAP summary plots for training (A) and test (B) datasets. SHAP dot plots show the individual-level impact and directionality of each feature on model predictions. (A) In the training set, younger age, presence of distant metastases, and higher thyroglobulin levels were associated with positive SHAP values (predicting recurrence). (B) The test set shows a similar pattern, demonstrating the consistency of feature effects across datasets. SHAP values >0 indicate contributions toward predicting recurrence or nonremission.

Abbreviation: SHAP, SHapley Additive exPlanations.

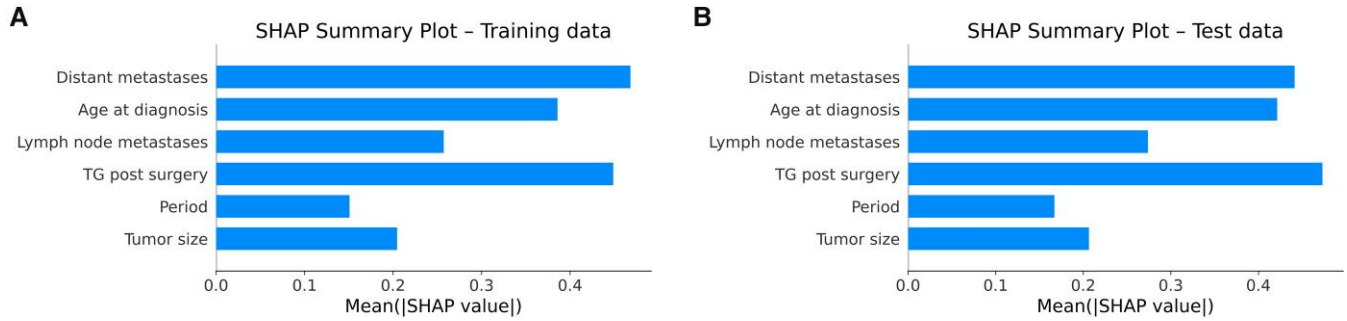

Figure 5. SHAP global feature importance based on mean absolute SHAP values. Bar plots showing the mean SHAP value for each feature, representing its average contribution to model output. (A) Training set. (B) Independent test set. Distant metastases, postoperative thyroglobulin levels, and age at diagnosis were the most influential predictors in both cohorts.

Abbreviation: SHAP, SHapley Additive exPlanations.

analysis revealed that age at diagnosis and biochemical response (postoperative TG levels) contributed more consistently to model predictions across the entire cohort. Although the presence of distant metastases exerted a strong influence when present—evident by high SHAP values in affected patients—this feature applied to only a minority of the cohort. In contrast, age and TG levels provided a predictive signal in all patients and exhibited broader variability in SHAP values, making them more consistently influential in the model's multivariable decision-making. These findings align more closely with the dynamic risk stratification framework proposed by Tuttle et al (17), which prioritizes early treatment response over static baseline staging. In our cohort, younger children (<10 years) were consistently assigned higher predicted risk, supporting prior evidence of age-related differences in tumor biology and immune response in pediatric DTC (16).

One key strength of this study is the use of SHAP to capture nonlinear and context-dependent effects. For instance, postoperative TG showed a steep rise in associated SHAP values between ~10 and 100 ng/mL, followed by a plateau—likely reflecting saturation in model decision trees, where further increases in TG no longer significantly shift predictions. This pattern mirrors clinical intuition: once TG exceeds a concerning threshold, the risk of persistent or recurrent disease is already high. An exploratory receiver operating characteristic curve analysis confirmed that a TG cutoff of approximately 9.0 ng/mL provided high sensitivity (91.3%) for predicting events, though specificity remained limited.

SHAP-based individual risk explanations illustrated how even a single variable, such as age, can shift risk predictions meaningfully when all other features are held constant. Furthermore, the same feature (eg, TG) contributed differently across patients depending on the presence of other variables (eg, metastases), emphasizing the added clinical value of ML explainability.

Unlike many previous ML models in thyroid cancer that focus on adult populations (19-21), our study addresses a critical gap in pediatric oncology. Pediatric DTC is biologically distinct from its adult counterpart—characterized by higher rates of metastases and excellent survival but significant recurrence risk. Few prior studies have applied ML to this setting (22). A notable exception is the recent work by Schindele et al (19), who demonstrated promising performance of an XGBoost model with SHAP analysis in adults. Our findings extend this methodology to pediatric patients, confirming the relevance of early treatment response and baseline risk factors in individualized outcome prediction.

Several limitations must be acknowledged. First, although the dataset was large for a rare pediatric cancer, the sample size limited the inclusion of additional variables, such as histopathology subtypes or autoimmune markers. Second, the study's retrospective design and single-registry source may introduce bias. While we validated the model using repeated stratified splits and tested on an independent hold-out set, external validation in independent, preferably multinational cohorts is essential.

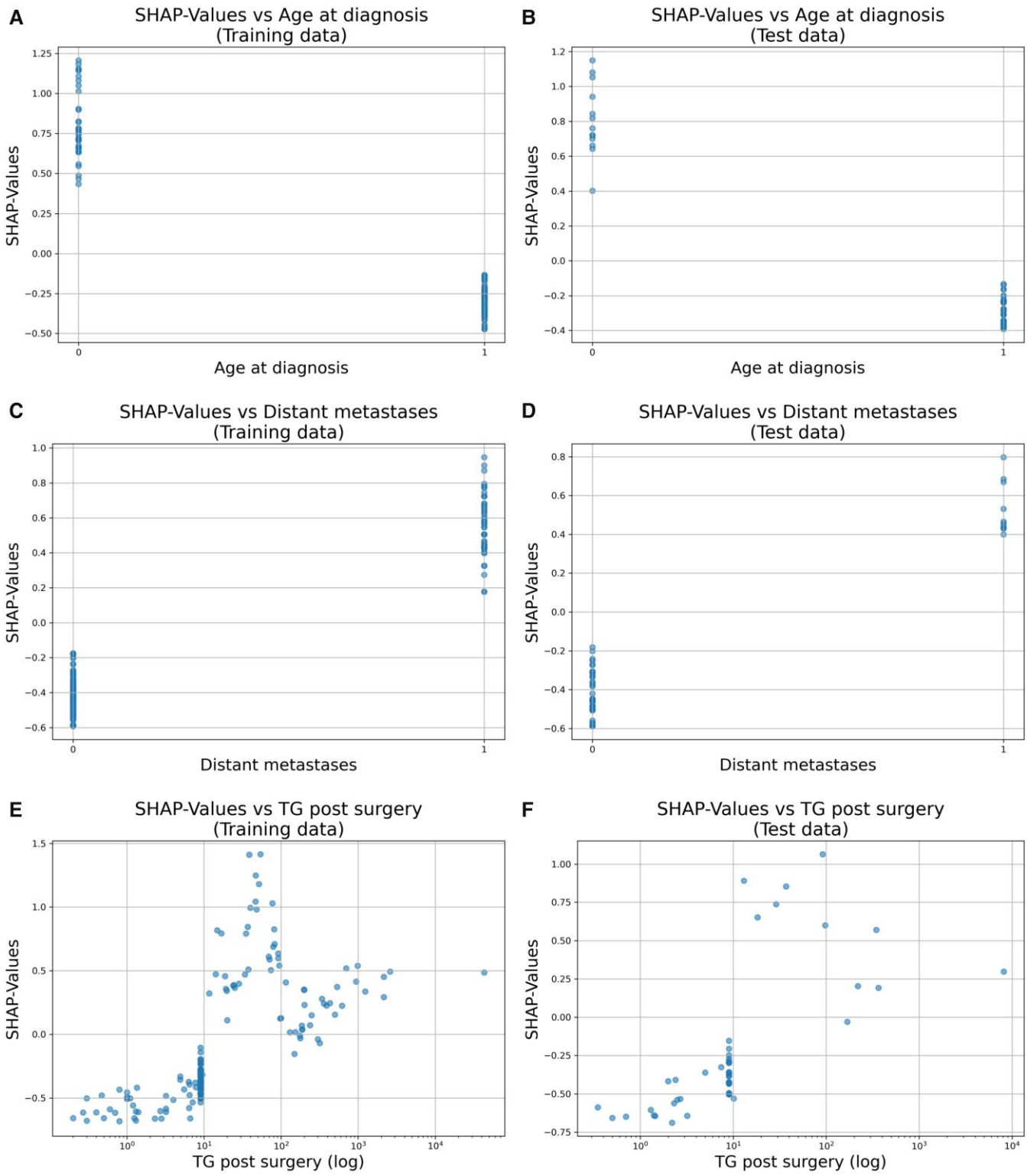

Figure 6. SHAP dependence plots for key predictive features in the training (left) and test (right) sets. (A + B) Age at diagnosis: younger patients (<10 years, encoded as 0) were consistently associated with positive SHAP values, indicating higher predicted risk of recurrence or nonremission. Older age (≥10 years, encoded as 1) was associated with negative SHAP values (protective effect). (C + D) Presence of distant metastases: presence of distant metastases (encoded as 1) was associated with high positive SHAP values, indicating a strong contribution toward predicting recurrence or nonremission. In contrast, absence of metastases (encoded as 0) corresponded to negative SHAP values, suggesting a protective effect. (E + F) Postoperative TG (log-transformed): A nonlinear pattern was observed, with low TG values linked to negative SHAP values (class 0, remission) and higher values associated with increased SHAP values (class 1, recurrence). Postoperative TG levels showed a steep increase in SHAP value between 10 and 100 ng/mL, plateauing at higher values. This reflects model saturation at high-risk thresholds rather than a decline in predicted risk. SHAP values >0 indicate a contribution toward predicting recurrence or nonremission (class 1); SHAP values <0 indicate a contribution toward remission (class 0). Consistency across both datasets supports the robustness of the model's interpretability.

Abbreviations: SHAP, SHapley Additive exPlanations; TG, thyroglobulin.

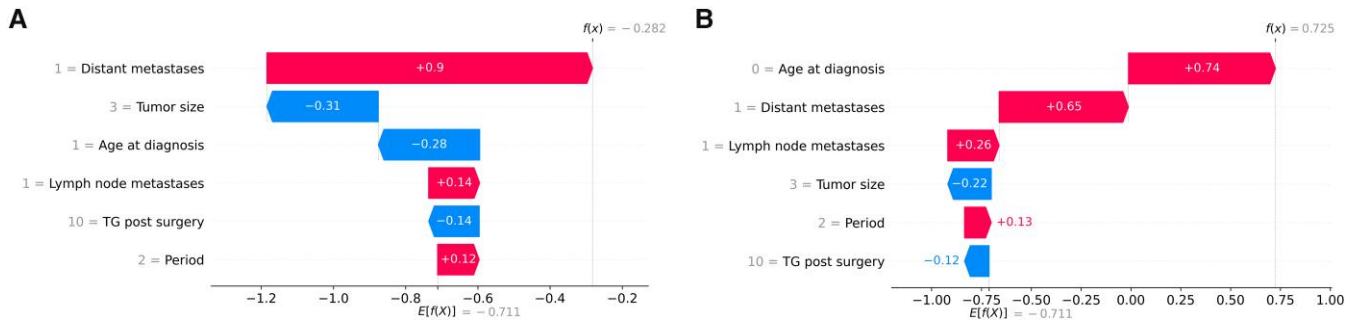

Figure 7. SHAP waterfall plot for individualized risk prediction in 2 hypothetical patients differing only in age at diagnosis. Each plot visualized the SHAP-based contribution of individual features to the model's predicted risk of recurrence or nonremission. Both patients had identical clinical characteristics except for age at diagnosis (≥10 vs <10 years). Patient A (≥10 years): predicted event risk: 43.0%. Patient B (<10 years): predicted event risk: 67.4%. The base value reflects the average model prediction across the training dataset. Red bars/arrows to the right indicate features that increased predicted risk (toward class 1), and blue bars/arrows to the left indicate features that decreased it (toward class 0). Notably, the influence of features such as distant metastases differed between the two patients, illustrating how SHAP captures context-dependent and interaction effects in individualized predictions.

Abbreviation: SHAP, SHapley Additive exPlanations.

Additionally, while SHAP improves model transparency, it reflects statistical associations—not causality. Features such as TG or metastases may also capture downstream effects of other unmeasured variables. As with any ML application in medicine, clinical judgment remains paramount in interpreting and acting upon model outputs.

Looking forward, several avenues merit exploration. Larger datasets may allow separation of recurrence and nonremission into distinct predictive models. Dynamic ML approaches incorporating longitudinal TG trends or imaging findings could improve prediction over time. Embedding these tools into clinical workflows, such as electronic health records or web-based calculators, may support real-time decision-making. Lastly, federated learning could allow secure, collaborative model development across institutions, enhancing generalizability while preserving patient privacy.

Conclusions

This study provides robust evidence that interpretable ML can identify pediatric DTC patients at increased risk of poor early outcomes. By highlighting key predictive features—particularly age and early biochemical response—and demonstrating model transparency using SHAP, our findings support the integration of ML-based tools into personalized pediatric thyroid cancer care.

Funding

The German malignant endocrine tumors studies were funded by Deutsche Kinderkrebsstiftung (grant numbers DKS 2021.11, DKS 2024.16, and DKS 2025.07). The research on pediatric differentiated thyroid carcinoma was funded by Mitteldeutsche Kinderkrebsforschung.

Author Contributions

A.R.: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Data Curation, Writing—Review & Editing, Visualization, Supervision, Project administration, Funding acquisition; E.P.: Methodology, Validation, Writing—Review & Editing; Ma.K.: Investigation, Data Curation, Writing—Review & Editing; M.S.: Investigation, Writing—Review & Editing; C.L.: Conceptualization, Investigation, Writing—Review & Editing; Mi.K.: Conceptualization, Methodology, Resources, Writing—Original Draft, Supervision, Project administration, Funding acquisition.

Disclosures

The authors have no conflicts of interest to declare.

Data Availability

Restrictions apply to the availability of some or all data generated or analyzed during this study to preserve patient confidentiality or because they were used under license. The corresponding author will on request detail the restrictions and any conditions under which access to some data may be provided.

References

1. Gupta S, Harper A, Ruan Y, *et al.* International trends in the incidence of cancer among adolescents and young adults. *J Natl Cancer Inst*. 2020;112(11):1105-1117.
2. Lebbink CA, van den Broek MFM, Kwast ABG, *et al.* Opposite incidence trends for differentiated and medullary thyroid cancer in young Dutch patients over a 30-year time span. *Cancers (Basel)*. 2021;13(20):5104.
3. Kuhlen M, Wellbrock M, Kunstreich M, *et al.* Incidence and temporal patterns of differentiated thyroid carcinoma in children and adolescents in Germany: a pooled analysis based on data from the German malignant endocrine tumor registry and the German childhood cancer registry. *Thyroid*. 2024;34(12):1540-1550.
4. Vaccarella S, Lortet-Tieulent J, Colombet M, *et al.* Global patterns and trends in incidence and mortality of thyroid cancer in children and adolescents: a population-based study. *Lancet Diabetes Endocrinol*. 2021;9(3):144-152.
5. Lazar L, Lebenthal Y, Steinmetz A, Yackobovitch-Gavan M, Phillip M. Differentiated thyroid carcinoma in pediatric patients: comparison of presentation and course between pre-pubertal children and adolescents. *J Pediatr*. 2009;154(5):708-714.
6. Ren PY, Liu J, Xue S, Chen G. Pediatric differentiated thyroid carcinoma: the clinicopathological features and the coexistence of Hashimoto's thyroiditis. *Asian J Surg*. 2019;42(1):112-119.
7. Pawelczak M, David R, Franklin B, Kessler M, Lam L, Shah B. Outcomes of children and adolescents with well-differentiated thyroid carcinoma and pulmonary metastases following (1)(3)(1)I treatment: a systematic review. *Thyroid*. 2010;20(10):1095-1101.

8. Welch Dinauer CA, Tuttle RM, Robie DK, *et al.* Clinical features associated with metastasis and recurrence of differentiated thyroid cancer in children, adolescents and young adults. *Clin Endocrinol (Oxf)*. 1998;49(5):619-628.
9. Francis GL, Waguespack SG, Bauer AJ, *et al.* Management guidelines for children with thyroid nodules and differentiated thyroid cancer. *Thyroid*. 2015;25(7):716-759.
10. Lebbink CA, Links TP, Czarniecka A, *et al.* 2022 European thyroid association guidelines for the management of pediatric thyroid nodules and differentiated thyroid carcinoma. *Eur Thyroid J*. 2022;11(6):e220146.
11. Damasio I, Dias D, Pinheiro S, Esteves S, Leite V, Santos R. Differentiated thyroid cancer in a pediatric population: estimating the risk of recurrence and evolution over time. *Cureus*. 2023;15(1):e34313.
12. Winder O, Lazar L, Hod R, Shpitzer T, Mizrachi A, Bachar G. Age stratification and prognostic factor analysis in pediatric differentiated thyroid cancer. *Laryngoscope*. 2024;134(11):4818-4825.
13. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358.
14. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016, San Francisco, California, USA.
15. Lundberg SM, Erion G, Chen H, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67.
16. Redlich A, Luster M, Lorenz K, *et al.* Age, American thyroid association risk group, and response to therapy are prognostic factors in children with differentiated thyroid cancer. *J Clin Endocrinol Metab*. 2022;107(1):e165-e177.
17. Tuttle RM, Tala H, Shah J, *et al.* Estimating risk of recurrence in differentiated thyroid cancer after total thyroidectomy and radioactive iodine remnant ablation: using response to therapy variables to modify the initial risk estimates predicted by the new American thyroid association staging system. *Thyroid*. 2010;20(12):1341-1349.
18. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
19. Schindele A, Krebold A, Heiss U, *et al.* Interpretable machine learning for thyroid cancer recurrence predicton: leveraging XGBoost and SHAP analysis. *Eur J Radiol*. 2025;186:112049.
20. Hanani AA, Donmez TB, Kutlu M, Mansour M. Predicting thyroid cancer recurrence using supervised CatBoost: a SHAP-based explainable AI approach. *Medicine (Baltimore)*. 2025;104(22):e42667.
21. Wang H, Zhang C, Li Q, *et al.* Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches. *BMC Cancer*. 2024;24(1):427.
22. Wang Z, Wang H, Zhou Y, *et al.* An individualized protein-based prognostic model to stratify pediatric patients with papillary thyroid carcinoma. *Nat Commun*. 2024;15(1):3560.