Interpretable machine learning for thyroid cancer recurrence predicton: leveraging XGBoost and SHAP analysis

Andreas Schindele, Anne Krebold, Ursula Heiß, Kerstin Nimptsch, Elisabeth Pfaehler, Christina Berr, Ralph A. Bundschuh, Thomas Wendler, Olivia Kertels, Johannes Tran-Gia, Christian H. Pfob, Constantin Lapa

Angaben zur Veröffentlichung / Publication details:

Schindele, Andreas, Anne Krebold, Ursula Heiß, Kerstin Nimptsch, Elisabeth Pfaehler, Christina Berr, Ralph A. Bundschuh, et al. 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.

European Journal of Radiology 186 (2025) 112049

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Interpretable machine learning for thyroid cancer recurrence predicton: Leveraging XGBoost and SHAP analysis

Andreas Schindele [a], Anne Krebold [a], Ursula Heiß [a], Kerstin Nimptsch [a], Elisabeth Pfaehler [a], Christina Berr [b], Ralph A. Bundschuh [a], Thomas Wendler [c], Olivia Kertels [d], Johannes Tran-Gia [e], Christian H. Pfob [a], Constantin Lapa [a,*]

[a] Nuclear Medicine, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[b] Internal Medicine I, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[c] Diagnostic and Interventional Radiology, Faculty of Medicine, University of Augsburg, Augsburg, Germany
[d] Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany
[e] Department of Nuclear Medicine, University Hospital Würzburg, Oberdürrbacherstr. 6, Würzburg 97080, Germany

A B S T R A C T

Purpose: For patients suffering from differentiated thyroid cancer (DTC), several clinical, laboratory, and pathological features (including patient age, tumor size, extrathyroidal extension, or serum thyroglobulin levels) are currently used to identify recurrence risk. Validation and potential adjustment of their individual and combined prognostic values using a large patient cohort with several years of follow-up might improve the correct identification of patients at risk.
Methods: In this retrospective study, we developed an XGBoost model using clinical and biomarker features for accurate DTC recurrence prediction using a cohort of 1228 consecutive patients (965 papillary, and 263 follicular) that were treated at the Department of Nuclear Medicine at University Hospital Augsburg between 1976 and 2010. The dataset was split into 982 patients for model training, and 246 for independent testing. From the 982 patients, 200 different random combinations of 785 training and 197 validation patients were conducted. To identify critical risk factors and understand the model's decision-making process, we conducted Shapely Additive exPlanations (SHAP) analysis.
Results: The XGBoost model achieved an AUROC of 0.84 (95 % CI: 0.84–0.86; SD: 0.08), sensitivity of 0.79 (95 % CI: 0.77–0.81; SD: 0.17), and specificity of 0.78 (95 % CI: 0.77–0.79; SD: 0.04) on the validation datasets, and an AUROC of 0.88 (sensitivity 0.83, specificity 0.80) on the independent test set. Tumor size, maximal thyroglobulin values within six months after thyroidectomy, and maximal thyroglobulin antibody levels within 12 to 24 months after thyroidectomy were the most important factors. SHAP dependence plots suggested new recurrence risk thresholds for a tumor size of 25 mm, maximal serum thyroglobulin levels of 3 and 10 ng/mL, respectively, and maximal thyroglobulin antibody levels of 120 IU/mL.
Conclusion: Our XGBoost model, supported by SHAP analysis empowers clinicians with interpretable insights and defined risk thresholds and could facilitate informed decision-making and patient-centric care.

1. Introduction

Differentiated thyroid cancer (DTC) is the most common endocrine malignancy, with its incidence steadily increasing in the United States and many other countries around the world over the past few decades [1,2]. While surgical intervention and radioiodine therapy have significantly improved patient outcomes, a considerable number of patients still experience disease recurrence after initial treatment [3]. Identifying individuals at high risk of recurrence is of paramount importance for optimizing patient management. This tailored approach may involve more frequent follow-up appointments, additional diagnostic tests, and closer monitoring for patients deemed to be at higher risk. Conversely, patients identified as lower risk may require less intensive surveillance, potentially reducing unnecessary interventions and healthcare costs while still ensuring appropriate care. By stratifying patients based on their risk of recurrence, healthcare providers can optimize resources, improve patient outcomes, and enhance survival rates in DTC management.

Over the years, several clinical and pathological factors have been investigated to predict thyroid cancer recurrence, including tumor size, patient age, tumor TNM staging, and various biomarkers such as serum thyroglobulin (HTG), thyroid-stimulating hormone (TSH), and thyroglobulin antibody (TgAb) levels [4–8]. Although these factors have shown some predictive capability, their individual and combined prognostic value remains challenging to be quantified accurately.

Machine learning (ML) has emerged as a powerful tool in medical

* Corresponding author.
E-mail address: Constantin.lapa@uk-augsburg.de (C. Lapa).

https://doi.org/10.1016/j.ejrad.2025.112049
Received 23 September 2024; Received in revised form 28 February 2025; Accepted 11 March 2025
Available online 14 March 2025

research and clinical decision-making, offering the potential to uncover complex patterns and interactions within large datasets. In recent years, ML approaches have been successfully applied to various cancer-related tasks, including survival prediction, treatment response assessment, and early diagnosis [9–12]. Leveraging the potential of ML for predicting thyroid cancer recurrences can lead to improved risk stratification and personalized treatment strategies.

While ML-based models have demonstrated promising results, they are not without challenges [13,14]. Optimal performance usually requires large datasets with comprehensive clinical information, which are not available in sufficient quantities for many indications and management strategies, in particular for therapeutic procedures. In addition, data annotation is associated with a considerable time expenditure and is error-prone. Finally, potential biases in the data [15] and the "black box" nature of most ML algorithms [16] raise concerns about the generalizability of the models and complicate the identification and interpretation of potential errors.

Building on the existing literature, this study develops an interpretable extreme gradient boosting (XGBoost) model [17] for thyroid cancer recurrence prediction. By training such a method with a comprehensive dataset comprising clinical, histopathological, and genetic features, we seek to enhance patient risk assessment. By integrating SHAP analysis, our objective is to unravel the model's decision-making process, to identify the key factors driving the predictions and to contribute to the growing body of literature on thyroid cancer recurrence prediction [18–22].

2. Methods

2.1. Dataset description

The dataset used in this study originates from patients diagnosed with DTC and treated by a surgical intervention at the Department of Nuclear Medicine at University Hospital Augsburg, Augsburg, Germany between 1976 and 2010. The data had been collected as part of routine clinical care and the cohort was retrospectively compiled for the study. Data dating back more than 30 years were included in more recent reports. The resulting initial dataset consisted of 2434 patients with primary DTC (papillary or follicular) confirmed by histopathology, who had undergone surgical intervention as the primary treatment. Electronic health records (EHRs) and pathology reports were used to extract relevant clinical and pathological information for each patient. The data were manually reviewed by trained medical personnel to ensure accuracy and consistency.

From the initial dataset of 2434 patients, 961 were excluded due to insufficient follow-up (i.e., less than 12 years), leaving 1473 patients. Of these, 245 patients were excluded because tumor size was missing (age was complete for all patients), resulting in a final sample of 1228 patients. While these exclusions were necessary to ensure reliable recurrence assessment, they might introduce selection bias if patients with shorter follow-up or missing tumor size differ systematically from those included.

The recurrence status of thyroid cancer was determined based on clinical and imaging assessments (neck ultrasound, radioiodine scintigraphy and [^{18}F]-FDG PET/CT), thyroglobulin (HTG) monitoring, and, if necessary, additional histopathological examinations. In addition, any measurable serum hTg was considered as tumor recurrence if thyroglobulin had not previously been detectable under TSH stimulation.

Patients were followed up for a minimum of 12.0 years, a maximum of 42.1 years, a mean of 20.5 years, and a standard deviation of 6.1 years following primary diagnosis.

This study adhered to the principles outlined in the Declaration of Helsinki and complied with local ethical guidelines. All patient data were pseudonymized to protect confidentiality and privacy. The use of the dataset for research purposes was approved by the institutional ethics committee of Ludwig-Maximilians-Universität München, Munich, Germany (approval number 22-1131).

2.2. Feature definitions

Features were included in this study based on their clinical relevance and their potential impact on thyroid cancer recurrence as reported by previous work. By incorporating multiple features, we aim to capture the complexity of interactions that may influence thyroid cancer recurrence.

- **Tumor size:** The tumor size is defined as the largest extension (diameter) of the primary tumor in millimeters. A larger tumor size has been associated with a higher risk of recurrence in thyroid cancer [7]. Tumor size is a well-established clinical parameter and is widely used for prognostic assessment in DTC patients [23].
- **Age:** Patient age in years at the time of diagnosis has been identified as an important prognostic factor for thyroid cancer recurrence [24,25]. Older patients tend to have a higher likelihood of experiencing recurrence, possibly due to a more aggressive tumor behavior and different underlying molecular mechanisms.
- **Nodal and metastatic staging**: The 2017 TNM staging system for thyroid cancer (8th Edition) is a standardized method used to describe the extent and spread of the disease [25,26]. It involves the three key components tumor size and invasion (T), lymph node involvement (N), and metastases to distant organs (M). This system helps clinicians determine the prognosis and guide treatment decisions for patients with thyroid cancer. We only consider N and M since T is already contained in the tumor size and the extrathyroidal extension. For N, the ordinal values 0 (absent), 1a (close nodal involvement), and 1b (extended nodal involvement) were used, while for M, 0 (absent) and 1 (present) were used.
- **Extrathyroidal extension (ETE):** ETE refers to the spread of thyroid cancer beyond the confines of the thyroid gland into surrounding tissues or structures in the neck (1 for T staging 3b, 4a, 4c, 0 otherwise). This extension is a critical factor in determining the stage and prognosis of thyroid cancer, as it signifies a more advanced and potentially aggressive disease [27].
- **Thyroglobulin (HTG):** Serum thyroglobulin (HTG) level (in ng/mL) is a well-established biomarker for thyroid cancer recurrence monitoring [4,5,7,8]. Elevated HTG levels after initial treatment are indicative of persistent or recurrent disease [28]. In this study, we employed the maximum HTG value observed within a six-month period after thyroidectomy and the corresponding TSH and TgAb values for predictive modeling. These features will be referred to as *HTG, TSH,* and *TgAb* in the following paper.
- **Thyroid-Stimulating Hormone (TSH) Value:** It plays a crucial role in thyroid cancer progression and recurrence, and is measured in mIU/L. TSH suppression therapy is commonly administered to thyroid cancer patients after surgery to reduce the risk of recurrence [23]. In this study, after identifying the highest TSH measurement following thyroidectomy, we selected the minimum TSH value observed during the subsequent twelve months and the corresponding HTG and TgAb values. This strategy is intended to capture the effect of TSH suppression on recurrence risk. These features will be referred to as *HTG_2, TSH_2*, and *TgAb_2*.
- **Thyroglobulin Antibody (TgAb) Value:** A high TgAb is often detected in patients with autoimmune thyroid disease. The presence and quantity of TgAb measured in IU/mL can affect thyroid function and potentially influence thyroid cancer prognosis [6]. For predictive modeling, we used the maximum TgAb value observed between 12 and 24 months after thyroidectomy. This feature will be referred to as *TgAb12-24*.
- **Number of Radioiodine Therapies:** In this study, we include the feature 'Number of Radioiodine Therapies' within a year after thyroidectomy in our machine learning model. This feature quantifies

the frequency of radioiodine therapies administered following thyroidectomy, providing data on postoperative management.
- **Subtype:** Papillary and follicular thyroid tumors are the two most common histological subtypes of thyroid cancer.

2.3. Data preprocessing

After data collection, the dataset was scanned for missing feature values. Missing data can impact the integrity of the analysis and model performance. To ensure the integrity and quality of the dataset, several preprocessing steps were performed prior to model training to address missing and inconsistent values in our dataset:

- **Tumor Size, Subtype and Age:** As stated in the exclusion criteria, patients with missing age, subtype, or tumor size were excluded from the analysis.
- **HTG, TSH, and TgAb Values:** Given the potential clinical significance of these features in predicting thyroid cancer recurrence median imputation was applied. The median was chosen because it is robust to outliers and better represents the central tendency in skewed distributions, common in clinical laboratory values, compared to the mean. The percentage of missing values imputed varied across features, with rates ranging from 1.8 % for HTG to 22.5 % for TgAb_12-24.
- **ETE, N, M, Number of therapies:** For these features, missing values were interpreted as the absence of the respective condition or therapy; hence, no imputation was performed.

As decision trees and ensemble trees are not sensitive to feature scaling, no feature scaling was used for this algorithm. In summary, the final input to the binary machine learning algorithm consists of 1228 samples and 12 features. Of the 1228 patients enrolled in this study, 33 patients experienced a recurrence of thyroid cancer. The remaining 1195 subjects did not suffer from a recurrence during the follow-up period.

2.4. Model training

Before training the machine learning models, we randomly divided the preprocessed dataset into two subsets: 982 patients for the training set (80 % of the data) and 246 patients for the test set (20 % of the data). The training set, which contained 26 recurrence cases, was used for model training, while the test set, with 7 recurrence cases, was held out for model evaluation. This split ensured that the models were tested on unseen data, providing a reliable estimate of their generalization performance. Stratified cross-validation was employed to ensure balanced representation of class labels across the train-test split, which is particularly important given the imbalanced nature of the data.

We then conducted a grid search to optimize hyperparameters using k-fold cross-validation within the training set. The hyperparameters yielding the best average performance across the folds were selected to build the final models.

To address class imbalance, we adopted a weighted sampling strategy. Specifically, for the majority class (patients without recurrence), the weight was computed as follows:

$$class\_weight_0 = \frac{n_{minority}}{N}$$

where:

- $n_{minority}$ is the number of minority class samples (patients with recurrence),
- N is the total number of samples.

Minority class samples were assigned a weight of 1. During model training, these weights were provided to the classifier via the sample_weight parameter in XGBoost, effectively scaling the loss function so that misclassifications of minority class samples incurred a higher penalty. This adjustment helped to mitigate the bias toward the majority class and improved the sensitivity of the model.

To robustly estimate model performance, we generated 200 validation datasets from the training set. This was achieved by repeated random subsampling (a Monte Carlo cross-validation approach): in each iteration, 20 % of the training data was randomly selected as a validation set while the remaining 80 % was used for training. Repeating this process 200 times allowed us to obtain a distribution of evaluation metrics, thereby quantifying variability and ensuring that performance estimates were not dependent on a single split.

2.5. Machine learning algorithm

XGBoost (Extreme Gradient Boosting) is an ensemble learning method that has recently gained popularity for its excellent performance in various machine learning tasks, including classification and regression [29]. It is an extension of the gradient boosting algorithm, which sequentially builds multiple weak learners (decision trees) and combines them to create a strong predictive model [17]. Additionally, L1 (Lasso) and L2 (Ridge) regularization terms are added to the loss function. This regularization helps to prevent overfitting and enhances the model's generalization performance. The algorithm builds multiple decision trees in parallel, which are combined to make the final prediction. By leveraging the strength of multiple trees, XGBoost can capture complex non-linear relationships between features and the target variable. XGBoost employs tree pruning techniques to control the depth of each decision tree, preventing overfitting and reducing computational complexity. Setting appropriate tree depth is essential to avoid capturing noisy patterns in the data.

To optimize the performance of the XGBoost model, we conducted hyperparameter tuning using randomized search and cross-validation on the training set. The following hyperparameters were tuned:

- **Learning Rate:** The step size shrinkage used to prevent overfitting. We explored values in the range [0.01, 0.40] and ultimately set it to 0.07.
- **Maximum Depth:** The maximum depth of a tree. We considered values in the range [6,30] and eventually settled on 25.
- **Gamma:** The minimum loss reduction required to make a further partition on a leaf node. We explored values in the range [0.0, 0.3], finally set to 0.3.

2.6. Performance evaluation

To assess the predictive performance of the machine learning models in predicting thyroid cancer recurrence, we used the following evaluation metrics: AUROC, sensitivity, and specificity.

The AUROC quantifies the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at various classification thresholds with 1 indicating perfect discrimination and 0.5 representing random classification.

Sensitivity, also known as the true positive rate, measures the proportion of correctly identified positive cases (patients with thyroid cancer recurrence) out of all true positive and false negative cases. Specificity, also known as the true negative rate, measures the proportion of correctly identified negative cases (patients without thyroid cancer recurrence) out of all true negative and false positive cases. Since sensitivity and specificity are threshold-dependent metrics, we determined the threshold such that sensitivity and specificity were similarly high on the validation datasets.

For each of the 200 validation datasets, we computed the AUROC, sensitivity, and specificity using the predetermined classification threshold. We then report both the mean and the standard deviation (SD) of these metrics across all iterations to quantify performance

variability. Additionally, we explicitly state the chosen threshold for sensitivity and specificity to ensure complete transparency in the evaluation process.

2.7. Model interpretability, SHAP analysis

To enhance the interpretability of the XGBoost model's predictions, we utilized SHAP (SHapley Additive exPlanations) values [30]. SHAP values quantify the contribution of each feature to a particular prediction by measuring the change in the model output from the expected value when a feature is present.

The SHAP analysis was conducted on a final model that was retrained using the complete training set. This approach avoids the need to aggregate SHAP values across different splits and ensures that the interpretability analysis reflects the model learned from the overall dataset.

To investigate the non-linear and interactive effects of individual features on the model's predictions, we generated SHAP dependence plots. In these plots, the x-axis represents the raw values of a given feature, and the y-axis displays the corresponding SHAP values, which indicate the impact on the model's output.

2.8. Software

All data preprocessing, model training, and performance evaluation were performed in Python 3.8 using the libraries scikit-learn, pandas, NumPy, Matplotlib, XGBoost, and SHAP.

3. Results

3.1. Descriptive statistics

A statistical summary of the selected numerical features used in the model are presented in Table 1:

- **Nodal and metastatic staging:** N: 1079 patients presented without lymph node involvement, 117 patients with N1a, and 32 patients with N1b disease. M: 1143 patients without hematologic spread, 125 patients with M1 disease.
- **ETE:** 166 of 1228 patients.
- **Num_Therapies:** Among the cohort of 1228 patients, 631 had not received radioiodine therapy, 586 had undergone a single therapy, 10 had undergone two therapies, and only 1 patient had received three successive therapies within one year.

Table 1
Statistical summary of numerical features included in this study.

Feature	Minimum	Maximum	Expectation value	Standard deviation
Tumor size [mm]	1	220	20.0	19.7
Age [years]	9	90	51.6	10.7
HTG [ng/mL]	0.1	120,000	475	5627
HTG_2 [ng/mL]	0	17,000	72.3	948
TSH [mIU/L]	0	350	31.1	34.6
TSH_2 [mIU/L]	0	180	1.14	8.00
TGAb [IU/mL]	0	35,239	179	1344
TGAb_2 [IU/mL]	0	28,775	91.9	856
TGAb_12-24 [IU/mL]	0	50,000	114	1499

- **Subtype:** 965 patients suffered from papillary thyroid cancer, and 263 patients had follicular thyroid cancer.

3.2. Threshold selection for sensitivity and specificity

The chosen threshold for the XGBoost model was 0.35. By selecting this threshold, we aimed to strike a balance between correctly identifying patients at risk of thyroid cancer recurrence (high sensitivity) and minimizing false positive predictions (high specificity).

3.3. Model performance

The XGBoost model demonstrated promising performance in predicting thyroid cancer recurrence on the validation datasets (created by randomly sampling 20 % of the training data 200 times) as well as on the test datasets. The XGBoost model's AUROC varied across the 200 validation datasets, reflecting its performance under different data splits. The performance values are displayed in Table 2.

The relatively low standard deviations prove stability and generalization of the XGBoost model across different validation datasets. This indicates that the model's performance is consistent and robust, even when trained on different subsets of the training data.

A histogram displaying the distribution of the validation AUROC is provided in Fig. 1. The histogram illustrates the range and frequency of AUROC values, providing a visual representation of the model's performance variability.

3.4. Feature importance analysis, SHAP

A bar plot illustrating the feature importance scores based on SHAP values is shown ion Fig. 2. This visualization reveals that Tumor Size had the most significant impact on the model's predictions. Similarly, TGAb_12-24 and HTG played crucial roles in the model's decision-making process. In contrast, the significance of the quantity of administered radioiodine therapies plays a limited role in predicting the likelihood of recurrence.

In Fig. 3, SHAP dependence plots are displayed for features with high predictive value. As displayed, tumors with sizes larger than 25 mm were associated with higher risk prediction of thyroid cancer recurrence in the SHAP dependence plot for tumor size. Similarly, TGAb12-24 showed a higher risk of recurrence at levels greater than 120 IU/mL, and HTG appeared to have a threshold of approximately 3 ng/mL for low to intermediate risk of recurrence and 10 ng/mL for increased risk of recurrence. Extrathyroidal extension is also a significant risk factor for recurrence and, as expected, showed a higher risk of recurrence. Age showed a threshold of 60 years, and HTG_2 of 0.6 ng/mL.

4. Discussion

The XGBoost model demonstrated promising performance in predicting thyroid cancer recurrence, achieving an AUROC of 0.88, a sensitivity of 0.83, and a specificity of 0.80 on the test dataset. The AUROC of 0.88 indicates that the model has a strong discriminative ability, with a higher likelihood of correctly ranking patients with and without thyroid cancer recurrence and is in a similar range to the

Table 2
Performance values for XGBoost Model. Standard deviation is not shown for test dataset since only one test set has been analyzed.

Perfomance measure	Test dataset values	Validation dataset mean ± standard deviation, 95 % confidence intervall
AUROC	0.88	0.84 ± 0.08, 0.84–0.86
Sensitivity	0.83	0.79 ± 0.17, 0.77–0.81
Specificity	0.80	0.78 ± 0.04, 0.77–0.79

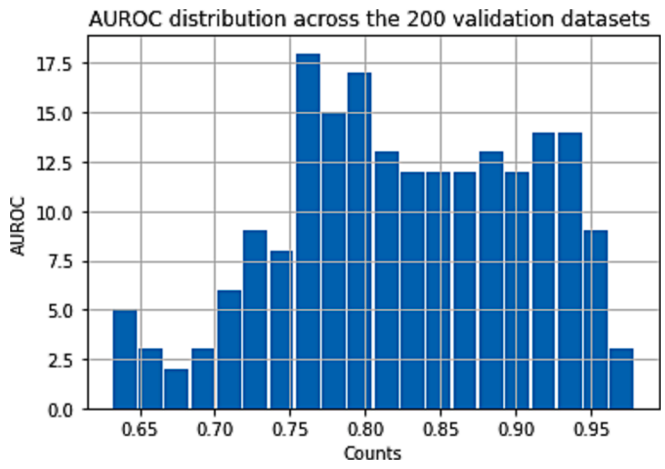

Fig. 1. Distribution of AUROC performance Distribution of AUROC performance across 200 randomly sampled validation datasets during cross-validation.

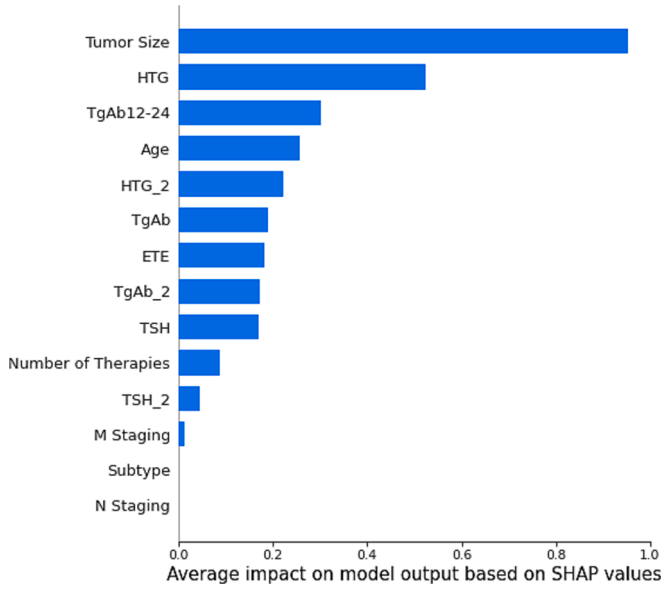

Fig. 2. Feature importance based on SHAP analysis HTG: maximum HTG value observed within six-month period after thyroidectomy, TSH and TgAb: corresponding TSH and antibody values. TSH_2: minimal TSH value observed within twelve-month period after highest measured TSH value after thyroidectomy. HTG_2 and TgAb: corresponding HTG and antibody values. TgAb12-24: highest antibody level within 12 to 24 months after thyroidectomy. M: metastasis staging. N: nodal staging. ETE: extrathyroidal extension.

AUROC of 0.9 reported previously [19].In this work, however, a deep learning model was trained on ultrasound images of the thyroid. Our results demonstrate that a similar accuracy can be achieved without any imaging, but instead with clinical parameters only.

Previously, a sensitivity of 0.71 and a specificity of 0.98 was reported [20]. However, it is noteworthy that all false predictions were categorized by the algorithm as abstentions. Furthermore, the study highlighted the significance of HTG levels 5 years post-surgery as the most crucial feature, yet the analysis in this paper only considered data up to 2 years post-surgery.

Our feature importance analysis provided insights into the factors contributing to thyroid cancer recurrence predictions. Tumor Size emerged as the most influential feature affecting the model's predictions, consistent with its established role as a critical prognostic factor. Additionally, TGAb12-24 and HTG were identified as prominent contributors, highlighting their significance in recurrence risk assessment. The model also considered TSH Value, Patient Age, and Tumor Staging to contribute to accurate predictions.

SHAP dependence plots provided valuable information on feature interactions. In particular, concrete cut-off values were found for Tumor Size, HTG, and TGAb12-24. Our SHAP dependence plots indicated that patients with a tumor size of 25 mm had a higher risk of recurrence. This threshold is stricter than the threshold reported in the AJCC guidelines [26], according to which tumors with sizes above 40 mm carry a higher risk than those with lower sizes. These thresholds give physicians the possibility to assess the patients' risk of recurrence at an early stage. Furthermore, thresholds are independent of the ML model, especially since de-novo TGAb can develop following a tumor recurrence diagnosis, as reported by Yin et al., making them potentially unsuitable for clinical recurrence diagnosis [31].

The ability to accurately predict thyroid cancer recurrence has significant clinical relevance. In particular, the thresholds identified in this study can help clinicians to make better informed decisions. To improve patient outcomes, patients identified as high-risk can benefit from more frequent monitoring, restaging and surveillance. In case of a recurrence, early therapeutic interventions can be performed, which can improve treatment outcomes.

Despite the promising results, our study has several limitations. First, the dataset is retrospective, and prospective validation is necessary to confirm the model's utility in routine clinical practice. Second, a potential selection bias due to retrospective data collection and the possibility of missing data despite imputation efforts cannot be excluded. Awareness of these limitations is essential when interpreting the study results.

Additionally, the model's performance may vary across different populations and healthcare settings, warranting external validation in diverse cohorts.

While SHAP analysis provides valuable interpretability, it has certain limitations. Interpreting complex interactions among high-dimensional features may be challenging. Future research should explore advanced techniques to address this limitation and further improve the model's interpretability.

Further research could use the model in a clinical setting to further enhance the model's performance and usability. Investigating the inclusion of additional data sources, such as genetic profiles or lifestyle factors, may improve the model's predictive accuracy. Data from different health centers are needed to validate the model's generalizability and reliability.

5. Conclusion

Our XGBoost model for predicting thyroid cancer recurrence demonstrated a strong discriminative ability. The SHAP analysis gave insight in model interpretability, highlighting tumor size, thyroglobulin antibody level 12 to 24 months after primary treatment, tumor staging, and maximum thyroglobulin values within six months after thyroidectomy as key predictors. SHAP dependence plots identified new cut-off values for recurrence prediction, such as a tumor size of 25 mm or greater, a thyroglobulin antibody level of 120 IU/mL, and a maximum thyroglobulin level of 3 and 10 ng/mL, respectively.

CRediT authorship contribution statement

Andreas Schindele: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Anne Krebold:** Writing – review & editing, Investigation, Data curation. **Ursula Heiß:** Writing – review & editing, Investigation, Data curation. **Kerstin Nimptsch:** Writing – review & editing, Investigation, Data curation. **Elisabeth Pfaehler:** Writing – review & editing, Resources, Methodology, Formal analysis, Data curation. **Christina Berr:** Writing – review &

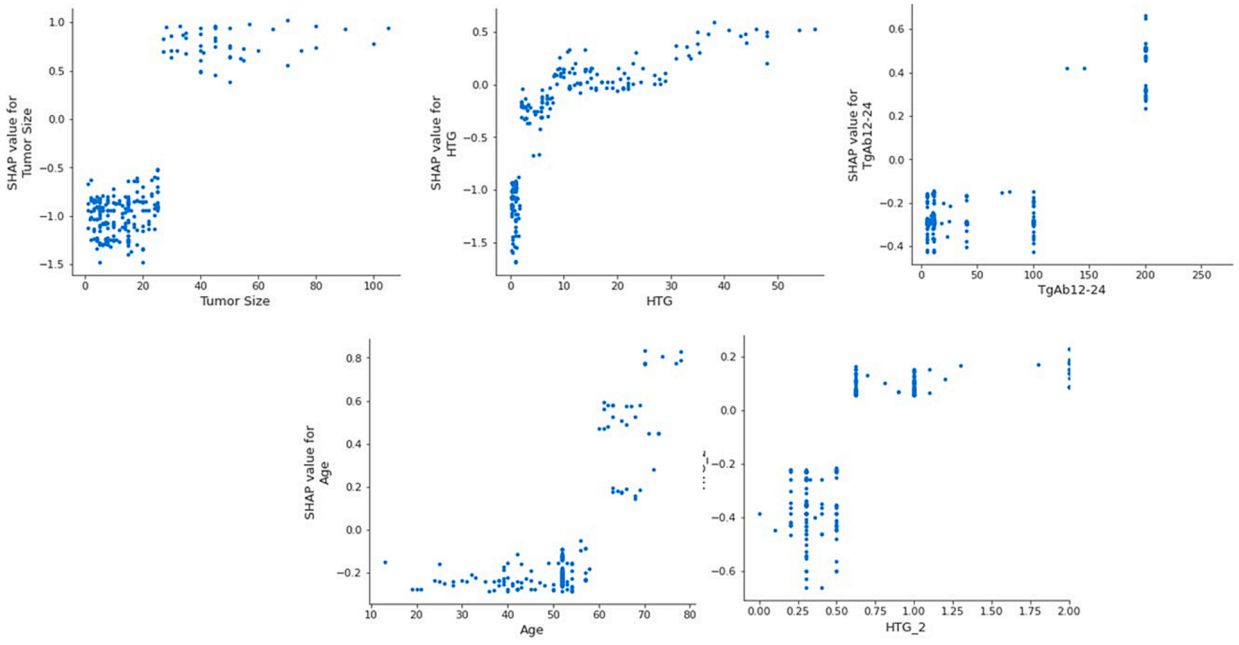

Fig. 3. Dependence Plots for individual features HTG: maximum HTG value observed within six-month period after thyroidectomy. HTG_2: HTG level corresponding to minimal TSH value observed within twelve-month period after highest measured TSH value after thyroidectomy. TgAb12-24: highest antibody level within 12 to 24 months after thyroidectomy. Each point in plot refers to one patient. On the y-axis SHAP values are displayed, on the x-axis corresponding feature values. E.g., for tumor size, SHAP values for tumor diameters smaller than 25 mm are below 0, while for tumor diameters larger than 25 mm, they are above 0. Therefore, threshold to detect a recurrence is for tumor diameter 25 mm. The y axis is differently scaled for each feature.

editing, Investigation, Conceptualization. **Ralph A. Bundschuh:** Writing – review & editing, Supervision, Resources. **Thomas Wendler:** Writing – review & editing, Project administration, Methodology, Formal analysis. **Olivia Kertels:** Writing – review & editing, Visualization, Resources, Data curation. **Johannes Tran-Gia:** Writing – review & editing, Supervision, Resources, Methodology, Formal analysis. **Christian H. Pfob:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Data curation, Conceptualization. **Constantin Lapa:** Writing – review & editing, Supervision, Project administration, Investigation, Formal analysis, Data curation, Conceptualization.

Informed consent

Written informed consent was obtained from all patients.

Ethics approval

This is an observational study. The use of the dataset for research purposes was approved by the institutional ethics committee of Ludwig-Maximilians-Universität München, Munich, Germany (approval number 22–1131).

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C.D. Seib, J.A. Sosa, Evolving Understanding of the Epidemiology of Thyroid Cancer, Endocrinol Metab Clin North Am 48 (2019) 23–35.
[2] G. Pellegriti, F. Frasca, C. Regalbuto, et al., Worldwide increasing incidence of thyroid cancer: update on epidemiology and risk factors, J Cancer Epidemiol 2013 (2013) 965212.
[3] I.D. Hay, G.B. Thompson, C.S. Grant, et al., Papillary thyroid carcinoma managed at the Mayo Clinic during six decades (1940-1999): temporal trends in initial therapy and long-term outcome in 2444 consecutively treated patients, World J Surg 26 (2002) 879–885.
[4] I.J. Nixon, L.Y. Wang, F.L. Palmer, et al., The impact of nodal status on outcome in older patients with papillary thyroid cancer, Surgery 156 (2014) 137–146.
[5] L. Giovanella, P.M. Clark, L. Chiovato, et al., Thyroglobulin measurement using highly sensitive assays in patients with differentiated thyroid cancer: a clinical position paper, Eur J Endocrinol 171 (2014) R33–R46.
[6] K. Jo, M.H. Kim, J. Ha, et al., Prognostic value of preoperative anti-thyroglobulin antibody in differentiated thyroid cancer, Clin Endocrinol (oxf) 87 (2017) 292–299.
[7] C.H. Shin, J.L. Roh, D.E. Song, et al., Prognostic value of tumor size and minimal extrathyroidal extension in papillary thyroid carcinoma, Am J Surg 220 (2020) 925–931.
[8] C. Elmaraghi, M. Shaaban, C. Reda, Prognostic value of postoperative stimulated thyroglobulin in differentiated thyroid cancer, Ann Endocrinol (paris) 84 (2023) 32–36.
[9] Z. Zhang, L. Huang, J. Li, et al., Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: a novel machine learning survival predictive system, BMC Bioinf. 23 (2022) 124.
[10] S. Moazemi, A. Erle, Z. Khurshid, et al., Decision-support for treatment with (177) Lu-PSMA: machine learning predicts response with high accuracy based on PSMA-PET/CT and clinical parameters, Ann Transl Med 9 (2021) 818.
[11] Y. Li, X. Wu, P. Yang, et al., Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis, Genomics Proteomics Bioinformatics 20 (2022) 850–866.
[12] Jones OT, Matin RN, van der Schaar M, et al: Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. Lancet Digit Health 4:e466-e476, 2022.
[13] F.H. Petzschner, Practical challenges for precision medicine, Science 383 (2024) 149–150.
[14] A. Zhang, L. Xing, J. Zou, et al., Shifting machine learning for healthcare from development to deployment and from models to data, Nat Biomed Eng 6 (2022) 1330–1345.
[15] K.N. Vokinger, S. Feuerriegel, A.S. Kesselheim, Mitigating bias in machine learning for medicine, Commun Med (lond) 1 (2021) 25.
[16] S. Kundu, AI in medicine must be explainable, Nat Med 27 (2021) 1328.

[17] CHEN T, GUESTRIN C: XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge, Discovery and Data Mining:785-794, 2016.
[18] E.M.L. Ruiz, T. Niu, M. Zerfaoui, et al., A novel gene panel for prediction of lymph-node metastasis and recurrence in patients with thyroid cancer, Surgery 167 (2020) 73–79.
[19] J. Kil, K.G. Kim, Y.J. Kim, et al., Deep Learning in Thyroid Ultrasonography to Predict Tumor Recurrence in Thyroid Cancers, Taehan Yongsang Uihakhoe Chi 81 (2020) 1164–1174.
[20] S.Y. Kim, Y.I. Kim, H.J. Kim, et al., New approach of prediction of recurrence in thyroid cancer patients using machine learning, Medicine (Baltimore) 100 (2021) e27493.
[21] F.A. Verburg, U. Mader, I. Grelle, et al., Only a Rapid Complete Biochemical Remission After 131I-Therapy is Associated with an Unimpaired Life Expectancy in Differentiated Thyroid Cancer, Horm Metab Res 49 (2017) 860–868.
[22] A. Vrachimis, B. Riemann, U. Mader, et al., Endogenous TSH levels at the time of (131)I ablation do not influence ablation success, recurrence-free survival or differentiated thyroid cancer-related mortality, Eur J Nucl Med Mol Imaging 43 (2016) 224–231.
[23] B.R. Haugen, E.K. Alexander, K.C. Bible, et al., 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer, Thyroid 26 (2016) 1–133.
[24] M.A. Adam, J. Pura, P. Goffredo, et al., Presence and Number of Lymph Node Metastases Are Associated With Compromised Survival for Patients Younger Than Age 45 Years With Papillary Thyroid Cancer, J Clin Oncol 33 (2015) 2370–2375.
[25] Y. Ito, A. Miyauchi, M. Fujishima, et al., Prognostic significance of patient age in papillary thyroid carcinoma with no high-risk features, Endocr J 69 (2022) 1131–1136.
[26] Amin MB, Greene FL, Edge SB, et al: The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J Clin 67:93-99, 2017.
[27] HTTPS://WWW.THYROID.ORG/PATIENT-THYROID-INFORMATION/CT-FOR-PATIENTS/JANUARY-2022/VOL-15-ISSUE-1-P-13-14/: AMERICAN THYROID ASSOCIATION, Accessed Dec. 18, 2023, pp 13-14.
[28] C. Spencer, I. Petrovic, S. Fatemi, Current thyroglobulin autoantibody (TgAb) assays often fail to detect interfering TgAb that can result in the reporting of falsely low/undetectable serum Tg IMA values for patients with differentiated thyroid cancer, J Clin Endocrinol Metab 96 (2011) 1283–1291.
[29] S. Sennan, A. Potti, C. Gogilamudi, et al: THYROID DISEASE PREDICTION USING XGBOOST ALGORITHMS. J. MOB. MULTIMED 18, 2022.
[30] E. Stenwig, G. Salvi, P.S. Rossi, et al., Comparative analysis of explainable machine learning prediction models for hospital mortality, BMC Med Res Methodol 22 (2022) 53.
[31] N. Yin, S.I. Sherman, Y. Pak, et al., The De Novo Detection of Anti-Thyroglobulin Antibodies and Differentiated Thyroid Cancer Recurrence, Thyroid 30 (2020) 1490–1495.