Development and validation of an interpretable machine learning model for predicting the risk of distant metastasis in papillary thyroid cancer: a multicenter study

Fei Hou,[a,g] Yun Zhu,[c,g] Hongbo Zhao,[d,g] Haolin Cai,[a] Yinghui Wang,[a] Xiaoqi Peng,[a] Lin Lu,[e] Rongli He,[a] Yan Hou,[f] Zhenhui Li,[b,**] and Ting Chen[a,*]

[a]Department of Nuclear Medicine, Yunnan Cancer Hospital, The Third Affiliated Hospital of Kunming Medical University, Peking University Cancer Hospital Yunnan, Kunming, China
[b]Department of Radiology, Yunnan Cancer Hospital, The Third Affiliated Hospital of Kunming Medical University, Peking University Cancer Hospital Yunnan, Kunming, China
[c]Department of Radiology, The First Affiliated Hospital of Kunming Medical University, Kunming, China
[d]Laboratory Zoology Department, Kunming Medical University, Kunming, China
[e]Academy of Biomedical Engineering, Kunming Medical University, Kunming, China
[f]Internal Medicine Department, The First Affiliated Hospital of Kunming Medical University, Kunming, China

eClinicalMedicine
2024;77: 102913
Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102913

Summary

Background The survival rate of patients with distant metastasis (DM) of papillary thyroid carcinoma (PTC) is significantly reduced. It is of great significance to find an effective method for early prediction of the risk of DM for formulating individualized diagnosis and treatment plans and improving prognosis. Previous studies have significant limitations, and it is still necessary to develop new models for predicting the risk of DM of PTC. We aimed to develop and validate interpretable machine learning (ML) models for early prediction of DM in patients with PTC using a multicenter cohort.

Methods We collected data on patients with PTC who were admitted between June 2013 and May 2023. Data from 1430 patients at Yunnan Cancer Hospital (YCH) served as the training and internal validation set, while data from 434 patients at the First Affiliated Hospital of Kunming Medical University (KMU 1st AH) was used as the external test set. Nine ML methods such as random forest (RF) were used to construct the model. Model prediction performance was compared using evaluation indicators such as the area under the receiver operating characteristic curve (AUC). The SHapley Additive exPlanation (SHAP) method was used to rank the feature importance and explain the final model.

Findings Among the nine ML models, the RF model performed the best. The RF model accurately predicted the risk of DM in patients with PTC in both the internal validation of the training set [AUC: 0.913, 95% confidence interval (CI) (0.9075–0.9185)] and the external test set [AUC: 0.8996, 95% CI (0.8483–0.9509)]. The calibration curve showed high agreement between the predicted and observed risks. In the sensitivity analysis focusing on DM sites of PTC, the RF model exhibited outstanding performance in predicting "lung-only metastasis" showing high AUC, specificity, sensitivity, F1 score, and a low Brier score. SHAP analysis identified variables that contributed to the model predictions. An online calculator based on the RF model was developed and made available for clinicians at https://predictingdistantmetastasis.shinyapps.io/shiny1/. 11 variables were included in the final RF model: age of the patient with PTC, whether the tumor size is > 2 cm, whether the tumor size is ≤ 1 cm, lymphocyte (LYM) count, monocyte (MONO) count, monocyte/lymphocyte ratio (MLR), thyroglobulin (TG) level, thyroid peroxidase antibody (TPOAb) level, whether the T stage is T1/2, whether the T stage is T3/4, and whether the N stage is N0.

Interpretation On the basis of large-sample and multicenter data, we developed and validated an explainable ML model for predicting the risk of DM in patients with PTC. The model helps clinicians to identify high-risk patients early and provides a basis for individualized patient treatment plans.

*Corresponding author. Department of Nuclear Medicine, Yunnan Cancer Hospital, The Third Affiliated Hospital of Kunming Medical University, Peking University Cancer Hospital Yunnan, 519 Kunzhou Road, Kunming 650118, Yunnan, China.
**Corresponding author. Department of Radiology, Yunnan Cancer Hospital, The Third Affiliated Hospital of Kunming Medical University, Peking University Cancer Hospital Yunnan, 519 Kunzhou Road, Kunming 650118, Yunnan, China.
E-mail addresses: chenting@kmmu.edu.cn (T. Chen), lizhenhui@kmmu.edu.cn (Z. Li).
[g]Contributed equally.

Funding This work was supported by the National Natural Science Foundation of China (No. 81960426, 82360345 and 82001986), the Outstanding Youth Science Foundation of Yunnan Basic Research Project (No. 202401AY070001-316), Yunnan Province Applied and Basic Research Foundation (No. 202401AT070008), and Ten Thousand Talent Plans for Young Top-notch Talents of Yunnan Province.

Keywords: Papillary thyroid cancer; Distant metastasis; Machine learning; Interpretable; Predictive model; Multicenter study

Research in context

Evidence before this study

We searched PubMed for articles published up to April 30, 2023, using the keywords "(thyroid cancer OR thyroid carcinoma) AND (model) AND (metastasis)," without language restrictions. Our search yielded 56 articles. After excluding irrelevant studies, 17 articles were identified as relevant. Although previous studies have developed models to predict distant metastasis (DM) in papillary thyroid carcinoma (PTC), these models have significant limitations, such as limited variables, the inclusion of hard-to-obtain variables, and a lack of representativeness in the study populations, among others. Therefore, there remains a need to develop a new model predicting DM in patients with PTC.

Added value of this study

The primary objective of this study was to develop and validate an interpretable machine learning (ML) model to predict the risk of DM in patients with PTC. A total of 1864 thyroid cancer patients were included, and nine ML algorithms were used to develop prediction models for risk of DM of PTC. The Random Forest (RF) model, which included 11 features such as age, tumor size, and T stage, demonstrated the highest accuracy. In both the internal and external validation cohorts, the RF model achieved areas under the receiver operating characteristic curve (AUC) of 0.913 (95% CI 0.9075–0.9185) and 0.8996 (95% CI 0.8483–0.9509), respectively. Additionally, sensitivity analysis revealed that the RF model performed exceptionally well in predicting "lung-only metastasis."

Implications of all the available evidence

This new model is the first PTC DM risk prediction model developed and validated using a large multicenter dataset outside of the SEER database. It features readily available predictive variables, robust performance, interpretability [The SHapley Additive explanation (SHAP) approach was employed to provide both a global explanation of the model's overall functionality and a local explanation that details how specific predictions are made for individual PTC cases based on their personalized data], and ease of clinical use (with a free online platform available at: https://predictingdistantmetastasis.shinyapps.io/shiny1/), highlighting its potential for clinical translation. The goal of this new model is to assist clinicians in identifying high-risk patients and provide effective support for the development of individualized diagnostic and treatment plans, ultimately improving patient outcomes.

Introduction

Thyroid cancer is the most common malignant tumor of the endocrine system, with papillary thyroid carcinoma (PTC) accounting for more than 80% of thyroid cancer cases.[1,2] In recent years, PTC has attracted widespread concern worldwide due to its high incidence.[2–4] Although PTC has biological properties that favor a good treatment response, with a good long-term prognosis and an average 10-year overall survival rate of approximately 90%,[3] some patients face poor clinical outcomes due to the development of distant metastasis (DM).[5–9] The most common sites of DM in PTC are the lungs and bones. Once DM occurs, the overall prognosis deteriorates significantly, with the 10-year survival rate dropping to 40%.[10] The DM of most PTC are occult, posing complex treatment and management challenges for clinicians. Therefore, it is urgent to increase the understanding of the risk factors and mechanisms of DM in PTC and to find effective prediction methods to develop individualized diagnosis and treatment plans and follow-up strategies.

Although prediction models for the risk of DM in thyroid cancer have been developed in many studies, these studies have significant limitations. For example, most of these studies are based on data from the Surveillance, Epidemiology, and End Results (SEER) database. However, due to the limited variables included in the SEER database, such as the lack of laboratory data, the application of the models based on these data in individualized prediction is limited to some extent.[11–18] In some studies, the focus has been on developing models only for specific thyroid cancer subtypes or for sex-specific predictions for thyroid cancer, limiting the applicability of these models.[12,14,15] In addition, machine learning (ML) techniques have not been used in many studies, or models lack interpretability to fully capture the complex relationships among variables and to provide clinically actionable explanations.[11,12,14–16] Some

studies are based on The Cancer Genome Atlas (TCGA) database, and the variables included in the models are gene mutations associated with DM in PTC, but genetic testing for these mutations is not commonly performed in clinical practice, greatly limiting the clinical application of the models.[19,20] Finally, although some studies are not based on data from the SEER database, only the risk factors influencing DM in thyroid cancer were explored in these studies, and prediction models for clinical use were not constructed.[21–27] Therefore, new models for predicting the risk of DM in PTC still need to be developed.

In this study, we aimed to develop and validate an explainable ML model for the early and accurate prediction of the risk of DM in patients with PTC in a multicenter cohort. We used the SHapley Additive exPlanation (SHAP) method to clarify the feature importance and explain the prediction results of the model to determine the practical significance of the model for predicting DM in patients with PTC.

Methods

Study cohort

We collected the data of patients with thyroid cancer (female or male ≥14 years) who were admitted to Yunnan Cancer Hospital (YCH) and the First Affiliated Hospital of Kunming Medical University (KMU 1st AH) between June 2013 and May 2023. The inclusion criteria were as follows: (1) patients had received a pathological diagnosis of PTC and (2) patients had undergone total thyroidectomy. A total of 1447 patients from YCH and 458 patients from KMU 1st AH met the inclusion criteria. The exclusion criteria were as follows: (1) patients with other pathological thyroid cancer types such as follicular carcinoma or poorly differentiated carcinoma (n = 10); (2) patients for whom laboratory, imaging, or pathological information was absent (n = 25); and (3) patients with multiple primary cancers (n = 6). Finally, data from 1430 patients from YCH were included as the training set, and data from 434 patients from KMU 1st AH were included as the test set for external validation. This study protocol complies with the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee of Yunnan Cancer Hospital (KYCS2023-094, KYCS2024-223). This study was retrospective and all data were anonymized, so the requirement for informed consent from patients could be waived. Study protocol can be found in the Supplementary material (Supplementary pp. 22–38).

Clinical features and data processing

The data for variables evaluated in this study were obtained from the patients' hospitalization electronic medical records (EMRs), including basic patient information, TNM stage (American Joint Committee on Cancer (AJCC) 8th edition), laboratory indicators (within one month before surgery), and postoperative pathological information. The basic patient information obtained included age, sex, and body mass index (BMI). Regarding TNM stage, T stage and N stage information was obtained. The tumor pathological information obtained included benign thyroid diseases, multifocality, invasion of adjacent tissues, and tumor size. Laboratory indicators obtained included white blood cell (WBC) count, red blood cell (RBC) count, platelet (PLT) count, hemoglobin (HGB) level, lymphocyte (LYM) count, monocyte (MONO) count, neutrophil (NE) count, eosinophil (EOS) count, basophil (BASO) count, glucose (GLU) level, alkaline phosphatase (ALP) level, thyroglobulin (TG) level, thyroid-stimulating hormone (TSH) level, thyroglobulin antibody (TGAb) level, and thyroid peroxidase antibody (TPOAb) level. Based on data preprocessing, the following inflammation-related factors were included: the PLT-to-LYM ratio (PLR), MONO-to-LYM ratio (MLR), EOS-to-LYM ratio (ELR), BASO-to-LYM ratio (BLR), NE-to-LYM ratio (NLR) and the systemic immune-inflammation index (SII; where SII = PLT count × NE count/LYM count). The above predictive factors were all derived from objective data in the EMRs.

Assessment of study outcomes

The clinical diagnostic criteria for DM in PTC included the following: (1) abnormalities found on an iodine-131 whole-body scan; (2) elevated serum TG levels, with supporting evidence from computed tomography (CT), single photon emission CT (SPECT)/CT, or positron emission tomography (PET)/CT; or (3) pathological confirmation of the disease as metastatic PTC after needle biopsy or surgical resection.[28,29] The imaging results were independently evaluated by two senior radiologists, who were blinded to any information about the predictors. In cases of disagreement, a third radiologist, also blinded, was consulted, and the final decision was reached through consensus.

Data preprocessing

The data outliers for continuous variables were checked using box plots. A data outlier was defined as a value higher than the upper quartile plus 1.5 times the interquartile range or lower than the lower quartile minus 1.5 times the interquartile range. Each data outlier was replaced by one of the two limits to make it closer to the distribution of the main data (Supplementary Fig. S1). Missing variables in the training and test sets are shown in Supplementary Fig. S2. The 'mice' package was used for multiple imputation of missing variables (Supplementary pp.3, Supplementary Tables S1 and S2). Additionally, we conducted a sensitivity analysis of various imputation methods for missing values (Supplementary Fig. S3 and Tables S3 and S4 and pp.5). All laboratory indicators retained their continuity and were not classified. The

categorical predictors were predetermined before model construction. Data for categorical variables were processed using the one-hot encoding method. To avoid data leakage, the above data preprocessing was performed separately on the training set and test set.

Selection of variables

In this study, we used the recursive feature elimination (RFE) method for variable selection to improve model prediction performance and increase model stability. RFE is a mainstream feature selection method for ML in which unimportant features are removed to eventually obtain the best feature combination, thus achieving optimal model performance.[30,31] Throughout the RFE process, we used 10 rounds of 10-fold cross-validation to evaluate the model performance, ensuring the robustness of the variable selection process and the generalization ability of the model.

Model development and validation

Nine ML models, including logistic regression (LR), decision tree (DT), random forest (RF), K-nearest neighbor (KNN), support vector machine (SVM), naive bayes (NB), extreme gradient boosting (XGB), stochastic gradient boosting (SGBT), and neural network (NNET) were used to predict the risk of DM in PTC. To optimize the prediction models, the final hyperparameters for each model were obtained on the optimal feature subset based on 10 rounds of 10-fold cross-validation combined with the default hyperparameter grid search of the "caret" package (Supplementary Table S5). Finally, the models were refitted on the training set with the optimal feature subset and the final hyperparameters (based on 10 rounds of 10-fold internal cross-validation).

Model performance comparison

The reliability of the models was evaluated with several commonly used evaluation indicators, including the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, F1 score, and Brier score. We used the Hosmer–Lemeshow test to assess the consistency between the model-predicted probabilities and the observed outcomes. Logarithmic Loss (Log–Loss) was used to calculate the difference between the actual labels and the predicted probabilities to measure the accuracy of the predictions. In addition, calibration curves were used to reflect the match between the predicted probabilities and the actual results. The DeLong test was employed to determine whether there was a significant difference in the AUC values of different models. Decision curve analysis (DCA) was used to evaluate the net benefit of the models at different thresholds. We selected the best prediction model on the basis of the performance of the above evaluation indicators in the training set and the test set.

Model explanation

Explaining ML models is challenging. The SHAP method is a game theory-based technique that ranks the importance of input features and explains the results of the prediction model, overcoming the "black box" problem. The SHAP method provides local and global explanations by calculating the contribution of each feature to the prediction results, thus increasing the model transparency and interpretability.[32]

Network calculator

To facilitate model application in a clinical setting, the final prediction model was integrated into a Shiny application-based web platform. When the values of the relevant features in the final model are provided, this application returns the probability of DM in patients with PTC.

Statistics

These ML models were developed using R version 4.3.1 and the "caret" (Version: 6.0.94) package. "caret" is a comprehensive package that provides a unified interface for various ML algorithms. The models were constructed by using the train function and the corresponding method parameters, i.e., LR (method = "glm"), DT (method = "rpart"), RF (method = "ranger"), SVM (method = "svmRadial"), KNN (method = "knn"), NB (method = "naive_bayes"), XGB (method = "xgbTree"), SGBT (method = "gbm"), and NNET (method = "nnet"). The discrimination performance was evaluated using the ROC curve analysis, and the AUC and its bias-corrected 95% confidence interval (CI) using 1000-fold bootstrap were reported.[33] The Brier score (ranging from 0 to 1) was used to calculate the difference between the estimated risk and the observed risk, with a value closer to 0 indicating better calibration, thus assessing model calibration. In addition, the calibration performance of the clinical prediction models was evaluated by the Hosmer–Lemeshow test, with P-values higher than 0.05 usually indicating a good fit between the model the actual data.[34] The performance of two prediction models was evaluated by comparing their AUC values using DeLong's test.[35] Integrated discrimination improvement (IDI) and net reclassification improvement (NRI) were used to evaluate the prediction performance improvement of a new model over the baseline model.[36] DCA was conducted to show the net benefit of using a model at different thresholds to assess the clinical value of the model.[37]

Continuous variables with normal distribution are presented as mean ± standard deviation and were compared with the t-test. Continuous variables with skewed distributions are presented as medians with interquartile ranges and compared with the Mann–Whitney U test or the Kruskal–Wallis H test. Categorical variables are presented as numbers with percentages and compared with the chi-square test. Independent

risk factors for the entire cohort (training set + test set) were determined by univariable and multivariable logistic regression analyses. The dose–response relationship between independent risk factors and DM was evaluated using a restricted cubic spline (RCS) function. A two-tailed P-value < 0.05 was considered statistically significant.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

Baseline clinical information

Among the 1430 patients from YCH (the training set), 207 patients (14%) developed DM, including 157 with lung metastasis, 23 with bone metastasis, 26 with bone and lung metastasis, and 1 with bone and liver metastasis. Among the 434 patients from KMU 1st AH (the test set), 47 patients (11%) developed DM, including 42 with lung metastasis, 3 with lung and bone metastasis, 1 with bone metastasis, and 1 with bone-brain metastasis. The demographic and clinicopathological characteristics of all patients are shown in Table 1. The distributions of data for most variables were comparable between the training set and the test set (most P-values were higher than 0.05). In addition, there were significant differences in multiple clinicopathological features between the group without DM (M0) and the group with DM (M1). For example, M1 patients were more likely to be male (30% vs. 22%), older (48.5 vs. 44 years), have a larger tumor diameter (>2 cm: 46% vs. 7.5%), and have a higher MONO count (0.39 vs. 0.31) than M0 patients. In addition, the PLR and MLR in the M1 group were significantly higher than those in the M0 group (PLR: 136.03 vs. 120.11, MLR: 0.21 vs. 0.15), while the TPOAb level in the M1 group was significantly lower than that in the M0 group (5.83 vs. 13.72), with P-values all <0.001. Details of the study design are shown in Fig. 1.

Independent risk factors

We explored the independent risk factors for DM in patients with PTC based on the entire cohort. Using univariate logistic regression analysis, 22 potential risk factors associated with DM (P < 0.05) were identified. After multivariate logistic regression analysis, 11 factors independently associated with the risk of DM in PTC were finally identified, namely, age, BMI, benign thyroid disease, tumor size, RBC count, MONO count, PLR, TG level, TPOAb level, T stage, and N stage (Supplementary Table S6).

Dose-response relationship

According to the results of the multivariate logistic regression analysis, we further explored the associations of age, BMI, RBC count, MONO count, PLR, TG level, and TPOAb level with DM in PTC. RCS is a commonly used method for exploring nonlinear associations between independent and dependent variables.[38] Before analyzing the dose–response relationship, we adjusted for confounding factors and performed nonlinear tests. The dose–response curves (Fig. 2) showed nonlinear relationships of age, TG level, and TPOAb level with DM in PTC (P < 0.05 for overall, P < 0.05 for nonlinear). The risk of DM in PTC increased rapidly when age was <24.87 or >50.77 years, as well as when the TG level was >1.39 ng/mL. The risk of DM in PTC decreased rapidly when the TPOAb level was >0.96 IU/mL. BMI had a linear relationship with DM in patients with PTC (P < 0.05 for overall, P > 0.05 for nonlinear), with a risk threshold of 18.73. In addition, MONO count, PLR, and RBC count had no significant overall or nonlinear associations with DM (P > 0.05 for overall, P > 0.05 for nonlinear).

Selection of predictor variables

We used the RFE strategy for feature selection. The optimal feature subset for each ML model was determined using the RFE method. The RFE variable selection process for each ML model is visualized in Supplementary Fig. S4. The importance score for each variable was calculated and displayed in a bar plot (Supplementary Fig. S5).

Model development, performance comparison and sensitivity analysis

We carried out ten rounds of 10-fold internal cross-validation to construct nine ML models. The results showed that the XGB model performed the best in terms of AUC (0.916 ± 0.028, 95% CI: 0.9106–0.9214) and specificity (0.973 ± 0.016, 95% CI: 0.9698–0.9762), followed by RF (AUC = 0.913 ± 0.028, 95% CI: 0.9075–0.9185; specificity = 0.983 ± 0.014, 95% CI: 0.9802–0.9858) and LR (AUC = 0.908 ± 0.026, 95% CI: 0.9029–0.9131; specificity = 0.955 ± 0.021, 95% CI: 0.9509–0.9591) (Fig. 3).

In the training set, the RF model performed the best in terms of discrimination and calibration, with an AUC of 0.9999 (95% CI: 0.9997–1) and a calibration score of 0.004 (95% CI: 0.001–0.009) (Fig. 4A and C). The Hosmer–Lemeshow test showed that the RF model fit well (P > 0.05) (Supplementary Fig. S7). The Log–Loss value of the RF model is 0.1046, which is the lowest among all models, indicating that the RF model fits the training data well and has strong predictive accuracy (Supplementary Fig. S7). The RF model had the highest accuracy (0.9811), precision (1.0000), sensitivity (0.8696), specificity (1.000), NPV (0.9784), and F1 score (0.9302) (Fig. 4E and Supplementary Table S8). The DeLong test showed that the AUC of the RF model was significantly different from that of all other models (P < 0.05), followed by the XGB model (Supplementary Table S9). NRI and

Characteristic	Overall, N = 1864	M0, N = 1610	M1, N = 254	P-value[a]	Training set, N = 1430	Test set, N = 434	P-value[a]
M stage, n (%)							0.05
M0	1610 (86%)				1223 (86%)	387 (89%)	
M1	254 (14%)				207 (14%)	47 (11%)	
Sex, n (%)				<0.01			0.06
Female	1432 (77%)	1254 (78%)	178 (70%)		1084 (76%)	348 (80%)	
Male	432 (23%)	356 (22%)	76 (30%)		346 (24%)	86 (20%)	
Benign thyroid lesions, n (%)				0.03			0.40
Yes	1014 (54%)	860 (53%)	154 (61%)		771 (54%)	243 (56%)	
No	850 (46%)	750 (47%)	100 (39%)		659 (46%)	191 (44%)	
Multifocal, n (%)				0.30			<0.01
Yes	1009 (54%)	864 (54%)	145 (57%)		716 (50%)	293 (68%)	
No	855 (46%)	746 (46%)	109 (43%)		714 (50%)	141 (32%)	
Infiltrated the adjacent tissue, n (%)				<0.01			>0.90
Yes	380 (20%)	256 (16%)	124 (49%)		292 (20%)	88 (20%)	
No	1848 (80%)				1138 (80%)	346 (80%)	
Tumor size, n (%)				<0.01			0.03
>1 and ≤ 2	385 (21%)	303 (19%)	82 (32%)		277 (19%)	108 (25%)	
>2	238 (13%)	120 (7.5%)	118 (46%)		179 (13%)	59 (14%)	
≤1	1241 (67%)	1187 (74%)	54 (21%)		974 (68%)	267 (62%)	
T stage, n (%)				<0.01			0.07
T1/2	1459 (78%)	1344 (83%)	115 (45%)		1133 (79%)	326 (75%)	
T3/4	405 (22%)	266 (17%)	139 (55%)		297 (21%)	108 (25%)	
N stage, n (%)				<0.01			<0.01
N0	1052 (56%)	994 (62%)	58 (23%)		831 (58%)	221 (51%)	
N1	812 (44%)	616 (38%)	196 (77%)		599 (42%)	213 (49%)	
Age, Median (IQR)	45.00 (37.00, 52.00)	44.00 (37.00, 51.00)	48.50 (37.25, 59.00)	<0.01	45.00 (37.00, 52.00)	45.00 (38.00, 53.00)	0.20
BMI, Median (IQR)	23.51 (21.30, 25.89)	23.63 (21.36, 25.91)	22.94 (20.76, 25.56)	<0.01	23.51 (21.23, 25.89)	23.45 (21.40, 25.91)	0.70
WBC, Median (IQR)	6.05 (5.08, 7.18)	6.07 (5.08, 7.14)	6.02 (5.09, 7.45)	0.70	6.14 (5.13, 7.29)	5.86 (4.87, 6.79)	<0.01
RBC, Median (IQR)	4.90 (4.61, 5.30)	4.92 (4.66, 5.30)	4.86 (4.50, 5.19)	<0.01	4.96 (4.70, 5.30)	4.83 (4.52, 5.21)	<0.01
PLT, Median (IQR)	243.00 (207.00, 289.00)	243.00 (207.00, 287.00)	248.50 (208.25, 303.75)	0.20	244.00 (206.00, 290.00)	243.00 (209.25, 287.75)	>0.90
HGB, Median (IQR)	145.00 (137.00, 156.00)	145.00 (137.00, 156.00)	144.00 (132.25, 155.00)	0.03	146.00 (137.00, 156.00)	143.00 (136.00, 154.00)	0.04
LYM, Median (IQR)	2.01 (1.64, 2.44)	2.03 (1.67, 2.45)	1.81 (1.40, 2.34)	<0.01	2.00 (1.64, 2.44)	2.05 (1.63, 2.39)	>0.90
MONO, Median (IQR)	0.32 (0.25, 0.41)	0.31 (0.24, 0.39)	0.39 (0.30, 0.51)	<0.01	0.31 (0.24, 0.39)	0.35 (0.28, 0.44)	<0.01
NE, Median (IQR)	3.45 (2.73, 4.37)	3.44 (2.73, 4.34)	3.49 (2.70, 4.57)	0.30	3.55 (2.81, 4.46)	3.19 (2.56, 4.02)	<0.01
EOS, Median (IQR)	0.11 (0.06, 0.19)	0.11 (0.06, 0.18)	0.12 (0.07, 0.19)	0.60	0.11 (0.06, 0.19)	0.11 (0.06, 0.18)	0.40
BASO, Median (IQR)	0.02 (0.01, 0.04)	0.02 (0.01, 0.04)	0.02 (0.01, 0.03)	0.09	0.02 (0.02, 0.03)	0.02 (0.00, 0.04)	0.01
PLR, Median (IQR)	121.73 (96.06, 154.06)	120.11 (95.33, 150.63)	136.03 (103.72, 179.57)	<0.01	122.14 (96.05, 154.83)	120.60 (96.18, 153.41)	0.80
MLR, Median (IQR)	0.16 (0.12, 0.21)	0.15 (0.12, 0.20)	0.21 (0.16, 0.26)	<0.01	0.15 (0.12, 0.20)	0.18 (0.14, 0.23)	<0.01
ELR, Median (IQR)	0.06 (0.03, 0.09)	0.05 (0.03, 0.09)	0.06 (0.04, 0.11)	<0.01	0.06 (0.03, 0.09)	0.05 (0.03, 0.09)	0.50
BLR, Median (IQR)	0.01 (0.01, 0.02)	0.01 (0.01, 0.02)	0.01 (0.01, 0.02)	0.90	0.01 (0.01, 0.02)	0.01 (0.00, 0.02)	0.03
NLR, Median (IQR)	1.72 (1.32, 2.26)	1.69 (1.30, 2.22)	1.86 (1.49, 2.57)	<0.01	1.76 (1.37, 2.31)	1.58 (1.22, 2.06)	<0.01
SII, Median (IQR)	413.92 (299.80, 578.42)	405.23 (295.31, 565.19)	459.42 (343.14, 699.43)	<0.01	422.57 (308.04, 599.29)	380.58 (275.52, 508.75)	<0.01
GLU, Median (IQR)	4.90 (4.51, 5.31)	4.90 (4.50, 5.30)	4.94 (4.54, 5.40)	0.12	4.88 (4.51, 5.33)	4.92 (4.50, 5.29)	0.40
ALP, Median (IQR)	70.00 (58.00, 85.00)	69.00 (57.30, 84.00)	73.40 (60.00, 88.00)	0.02	70.00 (58.00, 85.00)	68.55 (57.05, 83.18)	0.13
TG, Median (IQR)	16.39 (6.68, 46.29)	13.89 (5.87, 35.15)	83.48 (20.90, 86.93)	<0.01	15.14 (6.40, 42.67)	21.97 (9.33, 60.84)	<0.01
TSH, Median (IQR)	2.64 (1.71, 4.11)	2.67 (1.75, 4.11)	2.48 (1.41, 4.05)	0.02	2.72 (1.76, 4.24)	2.39 (1.59, 3.61)	<0.01
TGAb, Median (IQR)	15.03 (10.00, 57.97)	15.00 (10.00, 62.20)	15.89 (10.92, 42.86)	0.40	15.29 (10.00, 58.33)	15.00 (10.54, 51.55)	0.40
TPOAb, Median (IQR)	12.95 (4.95, 31.49)	13.72 (6.06, 33.12)	5.83 (2.00, 18.00)	<0.01	11.86 (3.77, 27.98)	17.54 (10.00, 38.94)	<0.01

BMI: body mass index; WBC: white blood cell; RBC: red blood cell; PLT: platelet; HGB: hemoglobin; LYM: lymphocyte; MONO: monocyte; NE: neutrophil; EOS: eosinophil; BASO: basophil; GLU: glucose; ALP: alkaline phosphatase; TG: thyroglobulin; TSH: thyroid stimulating hormone; TGAb: thyroglobulin antibody; TPOAb: thyroid peroxidase antibody; PLR: PLT-to-LYM ratio; MLR: MONO-to-LYM ratio; ELR: EOS-to-LYM ratio; BLR: BASO-to-LYM ratio; NLR: NE-to-LYM ratio; SII: the systemic immune-inflammation index = PLT count × NE count/LYM count. [a]Pearson's Chi-squared test; Wilcoxon rank sum test.

Table 1: **Comparison of demographic characteristics and clinical characteristics between distant metastasis (DM) and Non-DM patients, and between training and test sets.**

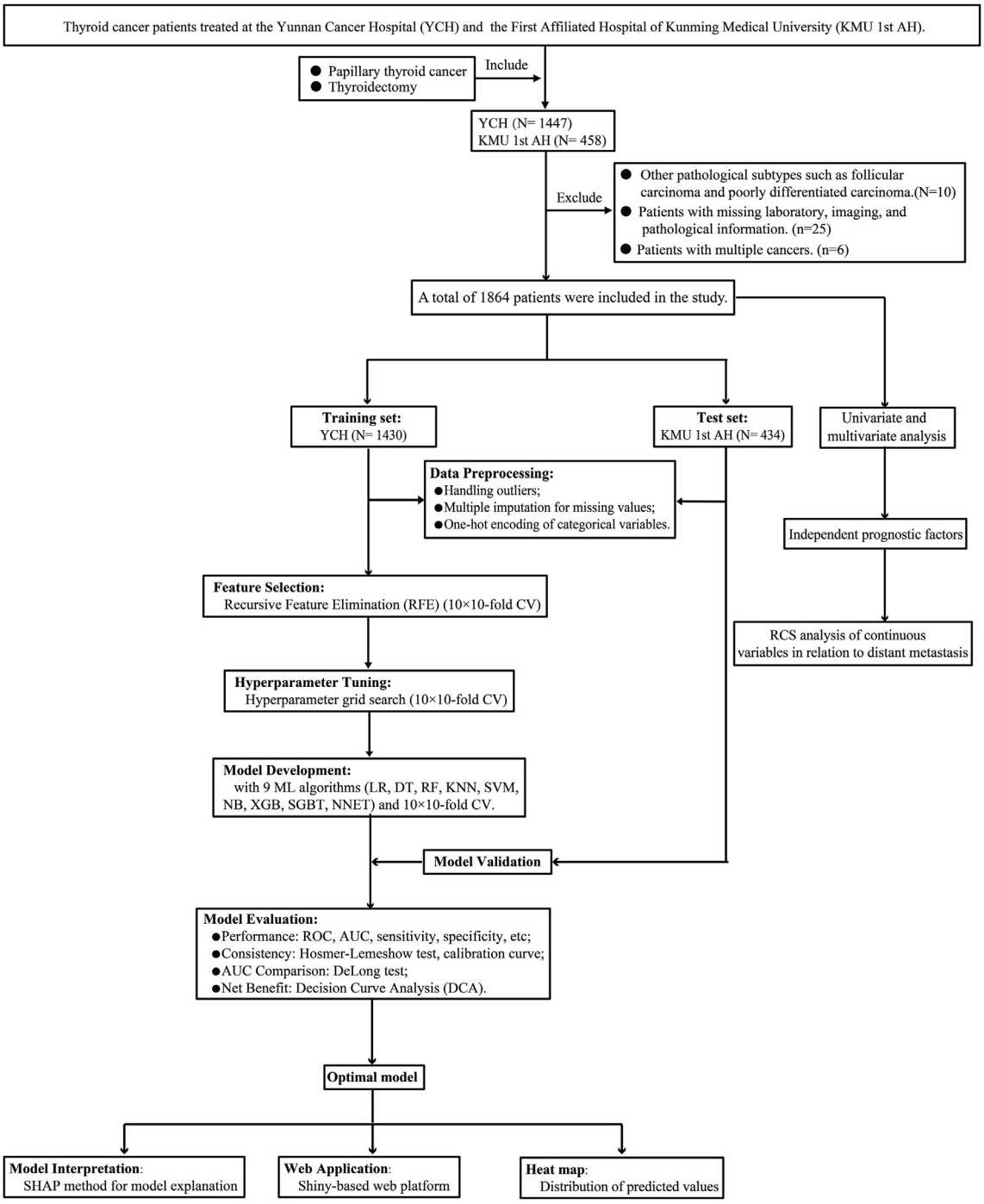

Fig. 1: Flow chart of the study design. YCH: Yunnan Cancer Hospital; KMU 1st AH: First Affiliated Hospital of Kunming Medical University; RCS: restricted cubic splines; CV: cross-validation; LR: logistic regression; DT: decision tree; RF: random forest; SVM: support vector machine; KNN: K-nearest neighbors; NB: naive bayes; XGB: extreme gradient boosting; SGBT: stochastic gradient boosting; NNET: neural network; ROC: receiver operating characteristic; AUC: area under curve; SHAP: SHapley Additive explanation.

IDI analysis showed that the RF model performed the best in terms of reclassification and overall discriminatory ability, followed by the XGB model (Supplementary Table S10). DCA showed that the RF model performed the best across the entire threshold range (0–1.0), followed by the XGB model (Fig. 4G).

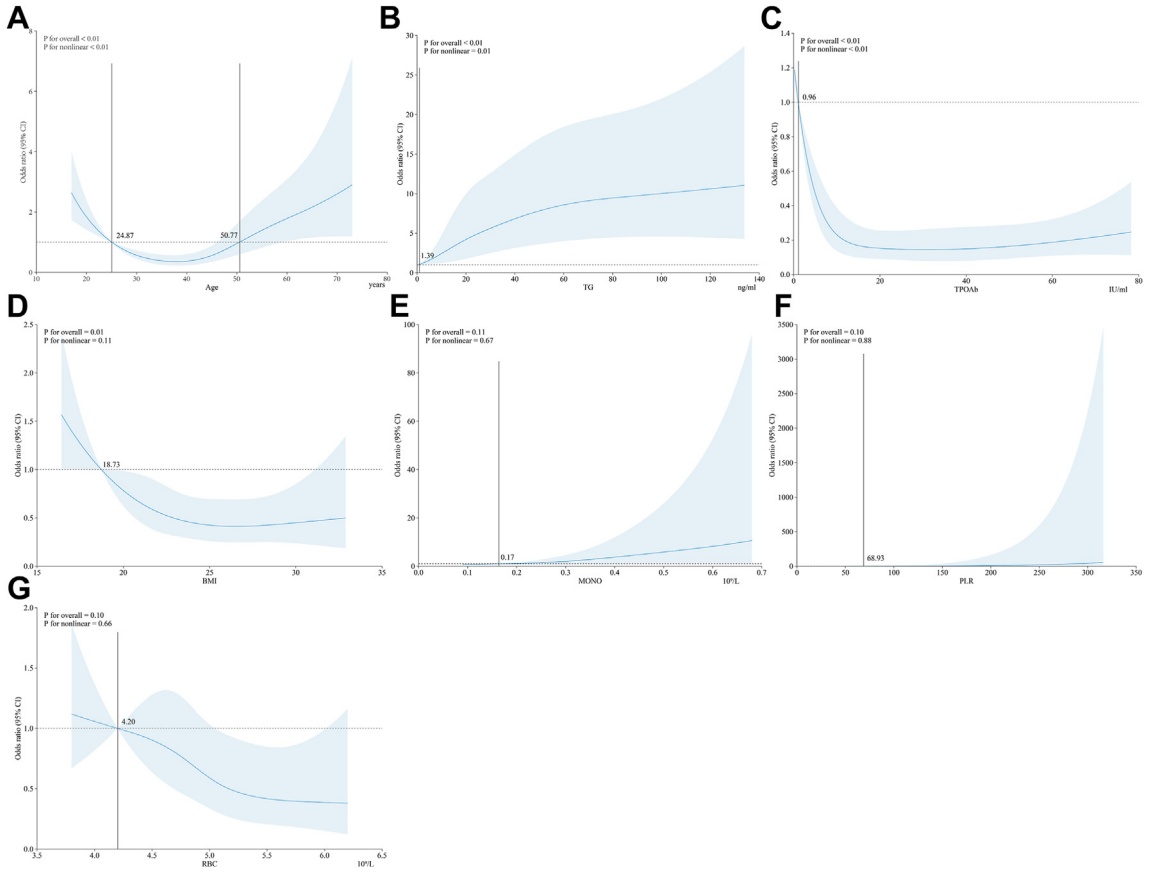

Fig. 2: Restricted cubic spline (RCS) plots for different continuous variables. These plots illustrate the nonlinear relationships between each continuous variable and distant metastasis (DM). The specific continuous variables include age (A), triglycerides (TG) (B), thyroid peroxidase antibody (TPOAb) (C), body mass index (BMI) (D), monocytes (MONO) (E), platelet-to-lymphocyte ratio (PLR) (F), and red blood cells (RBC) (G). Each variable's overall and nonlinear relationships are accompanied by P-values, indicating the significance and nonlinearity of the associations.

In the test set, the RF model had the best performance, with an AUC of 0.8996 (95% CI: 0.8483–0.9509) and a calibration degree of 0.057 (95% CI: 0.036–0.081) (Fig. 4B and D). The Hosmer–Lemeshow test revealed that the LR, RF, SVM, KNN, XGB, SGBT, and NNET models fit well (P > 0.05) (Supplementary Fig. S7). The Log–Loss value of the RF model is 0.2243, which is also the lowest among all models. This indicates that the RF model has good generalization ability, maintaining stable performance on unseen data (Supplementary Table S7). The RF model had the highest accuracy (0.9332), precision (0.8214), specificity (0.9871), and F1 score (0.6133) (Fig. 4F and Supplementary Table S8). The DeLong test showed that the AUC of the RF model was higher than that of several other models, and the difference was statistically significant (P < 0.05) (Supplementary Table S11). NRI and IDI analysis revealed that the RF and XGB models performed excellently in the external test set. In particular, the NRI and IDI values of the RF and XGB models were significantly higher than those of the DT and KNN models (P < 0.05), suggesting that the RF and XGB models performed the best in terms of reclassification and overall discriminatory ability in the external test set (Supplementary Table S12). DCA showed that the RF model performed the best across the entire threshold range (0–0.8), with the XGB model being the second best (Fig. 4H).

In the sensitivity analysis for predicting distant metastasis sites in PTC, the RF, SGBT, and XGB models performed the best, particularly in predicting 'lung-only metastasis', where they demonstrated high sensitivity, specificity, F1 score, AUC, and low Brier score. However, when predicting 'other metastases', the sensitivity, F1 score, and AUC of all models decreased significantly, showing poorer performance (Supplementary Fig. S6 and Table S13). Overall, the RF, SGBT, and XGB models exhibited strong overall performance across different types of metastasis predictions.

In summary, the RF model performed the best in both the training set and the test set and is thus

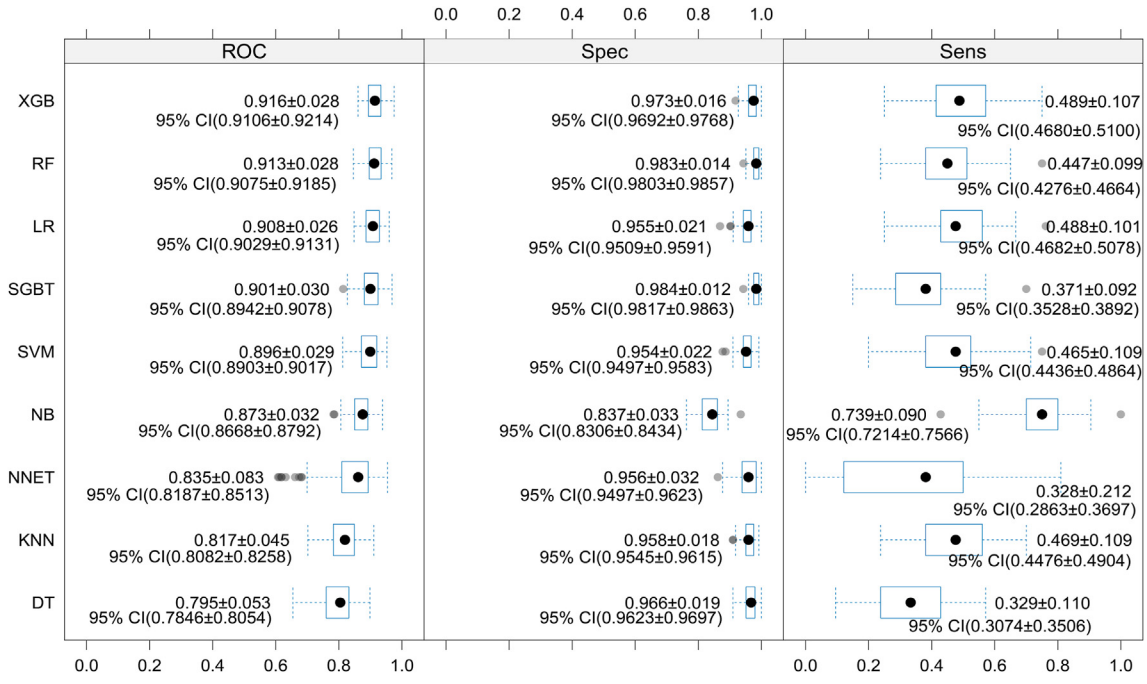

Fig. 3: The internal validation results of nine machine learning (ML) models. It displays the receiver operating characteristic (ROC) curves, sensitivity (Sens), and specificity (Spec) of the nine ML models in internal validation. These models include decision tree (DT), k-nearest neighbors (KNN), neural network (NNET), naive Bayes (NB), support vector machine (SVM), stochastic gradient boosting (SGBT), logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB).

recommended as the preferred model for the prediction of the risk of DM in PTC, followed by the XGB model.

Heatmap analysis of RF model variables

A heatmap was created to show the performance of the RF model in predicting DM in PTC. Different colors were used to show the distributions of the actual values of various predictive factors (such as age, tumor size, and LYM count) among different patients, as well as the model-predicted DM probability and the actual outcomes (Fig. 5). Heatmap analysis revealed that the model could effectively distinguish between high-risk

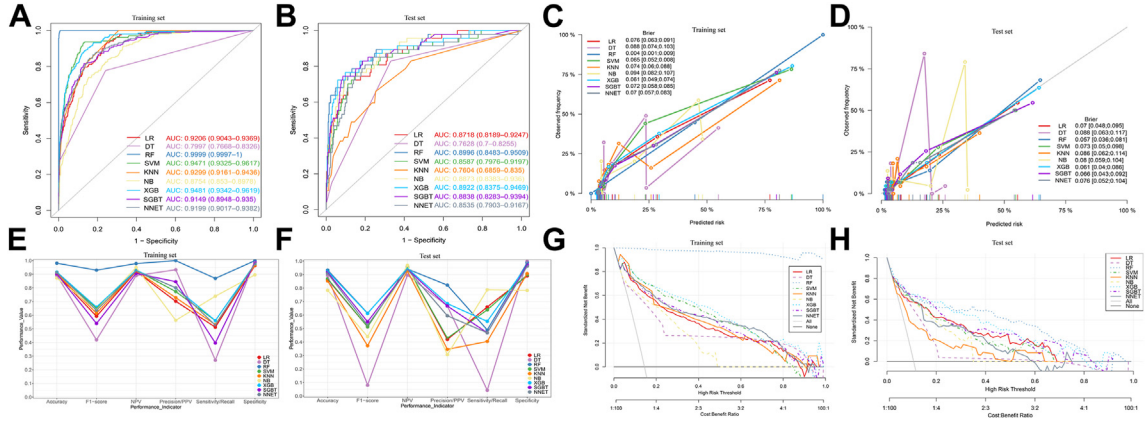

Fig. 4: Performance of machine learning (ML) models predicting distant metastasis (DM) in patients with papillary thyroid cancer (PTC) in the training and test sets. ROC curve analysis (A, B), calibration curve analysis (C, D), parallel line graph of the evaluation metrics for each model (E, F), and DCA curves for each model (G, H) predicting DM in PTC patients using nine ML algorithms in the training and test sets. Abbreviations: LR: logistic regression; DT: decision tree; RF: random forest; SVM: support vector machine; KNN: K-nearest neighbors; NB: naive bayes; XGB: extreme gradient boosting; SGBT: stochastic gradient boosting; NNET: neural network.

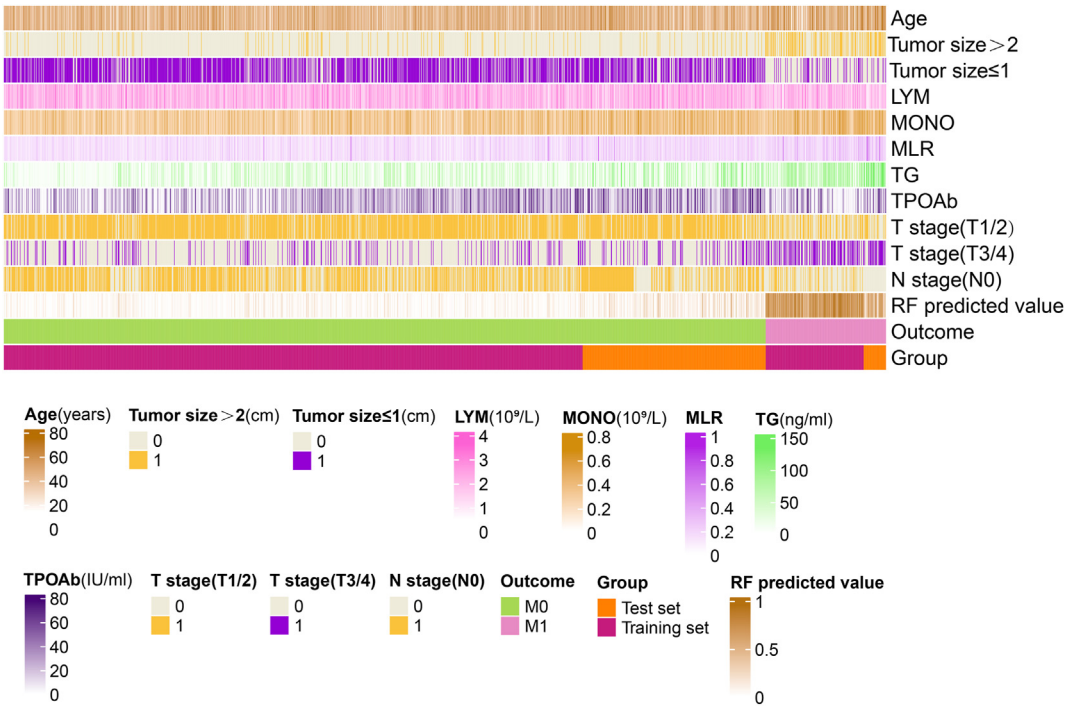

Fig. 5: Variable heatmap of the random forest (RF) model predicting distant metastasis (DM) in papillary thyroid cancer (PTC). Each row represents a variable, and each column represents a sample, with colors indicating the variable's values. The color legend represents the different value ranges of each variable, including age, tumor size (Tumor_size ≤ 1 cm, Tumor_size > 2 cm), lymphocyte count (LYM), monocyte count (MONO), monocyte-to-lymphocyte ratio (MLR), thyroglobulin (TG), thyroid peroxidase antibody (TPOAb), T stage [T_stage(T1/2), T_stage(T3/4)], N stage [N_stage(N0)], RF model predicted probability (RF), actual distant metastasis outcome (Outcome), and dataset grouping (Group). Color gradients indicate the value ranges of variables, and categorical variables are represented by different colors for categories.

and low-risk patients, with consistent performance in the training set and the test set, indicating that the model had good prediction accuracy and generalization ability and therefore may be helpful for the early clinical identification of high-risk patients with PTC.

Model explanation

Since it is difficult for clinicians to accept prediction models that are not directly explainable or not explainable at all, we used the SHAP method to explain the output of the final model by calculating the contribution of each variable to the prediction. This explainable method provides two types of explanations: a global explanation of the model at the feature level and a local explanation at the individual level. The global explanation describes the overall functionality of the model. As shown in the SHAP summary bar plot (Fig. 6A), the contribution of features to the model was evaluated using the mean SHAP values, which were displayed in descending order: tumor size, TG level, TPOAb level, MLR and age were the five most important features in the prediction model. In addition, the SHAP summary dot plot (Fig. 6B) visually shows the direction and strength of the influence of each feature on the model prediction: features such as a large tumor size, high TG level, advanced age, and late tumor stages (T3 and T4) significantly increased the risk of DM. In addition, the SHAP dependence plot helps to understand how individual features affect the output of the prediction model. Fig. 6D compares the actual values and the SHAP values of these 11 features, where features with SHAP values greater than zero correspond to positive predictions in the model; that is, these features indicate a higher risk of DM. For example, patients with PTC with a TG level >45.56 ng/mL, an MLR >0.1893, a MONO count $>0.45 \times 10^9$/L, and age <25 years or >50 years had SHAP values greater than zero, pushing the decision toward the "DM" category. In addition, a TPOAb level <4.05 U/mL and LYM count $<1.32 \times 10^9$/L both lead to the classification of "DM". Finally, when the tumor size >2 cm, T stage = T3/4, and N stage≠N0, the classification is also "DM".

Additionally, local explanation helps us understand the decision-making mechanism of the model by calculating and displaying the contribution of each feature to the prediction result of individual samples.

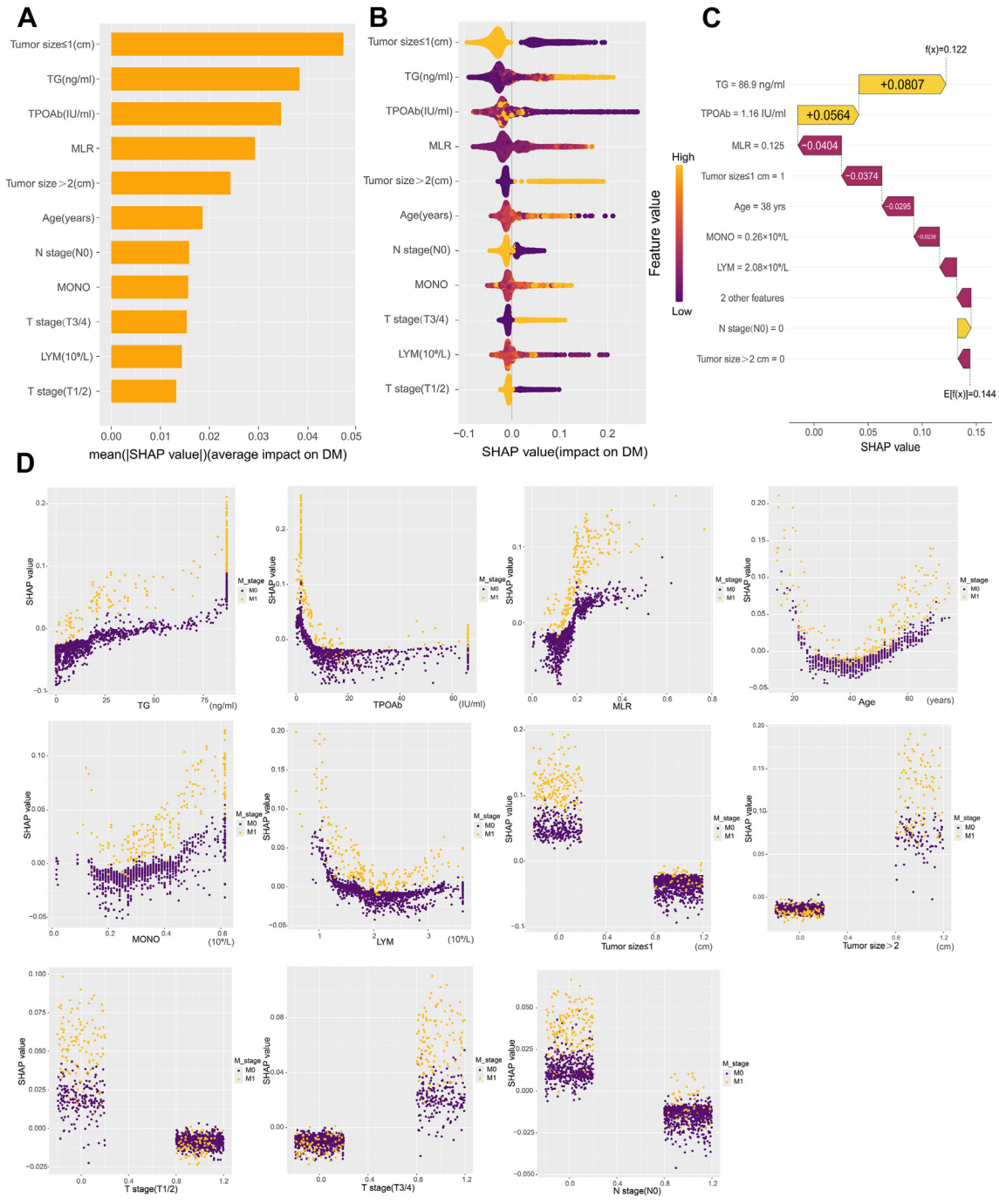

Fig. 6: Global and local model explanation by the SHapley Additive exPlanation (SHAP) method. (A) SHAP summary bar plot. This plot evaluates the contribution of each feature to the model using mean SHAP values, displayed in descending order. (B) SHAP summary dot plot. The probability of developing distant metastasis (DM) increases with the SHAP values of the features. Each dot represents a patient's SHAP value for a given feature, with orange indicating higher feature values and purple indicating lower values. Dots are stacked vertically to show density. (C) SHAP waterfall plot. This plot shows the contribution of each feature to the prediction result of the third patient using the random forest (RF) model. Orange bars indicate features that contribute positively to the prediction, while purple bars indicate negative contributions. Feature values are shown alongside their SHAP values, highlighting key features such as thyroglobulin (TG, +0.0807), thyroid peroxidase antibody (TPOAb, +0.0564), monocyte to lymphocyte ratio (MLR, −0.0404), and tumor size ≤1 cm (−0.0374). The overall contribution is 0.122, with a baseline contribution of 0.144. (D) SHAP dependence plot. Each dependence plot shows how a single feature affects the model's output, with each point representing a patient. For example, age values either below 25 or above 50 push the decision towards the "DM" class. SHAP values are on the y-axis, and actual feature values are on the x-axis. Features with SHAP values above zero push the decision towards the "DM" class.

The SHAP waterfall plot (Fig. 6C) shows the contribution of each feature to the model prediction of DM in for the third patient with PTC. The specific value of each feature and its corresponding SHAP value in the plot indicate the positive and negative impact of the feature on the prediction result. A TG level of 86.9 and TPOAb level of 1.16 made significant positive contributions of +0.0807 and + 0.0564 to the prediction result, respectively, while an MLR of 0.125 and a tumor size ≤1 cm made significant negative contributions of –0.0404 and –0.0374 to the prediction results, respectively. Other features, such as age, MONO count, LYM count, lymph node stage, and tumor size >2 cm, also had varying degrees of impact. By accumulating SHAP values, the waterfall plot visually demonstrates the formation process of the prediction result for a specific patient, helping us to understand in depth the decision-making mechanism of the model.

Implementation of the web calculator

As shown in Fig. 7, the final prediction model was integrated into a web application for use in clinical settings. By inputting the actual values of the 11 features required for the model, the application can automatically predict the risk of DM in individual patients with PTC. The web application can be accessed online at the following link: https://predictingdistantmetastasis.shinyapps.io/shiny1/.

Discussion

To our knowledge, this is the first ML model based on large-sample multicenter data from a source other than the SEER database. A total of nine ML models for the prediction and analysis of the risk of DM in PTC were studied and compared. We identified a set of predictive risk factors and used various ML algorithms along with clinical and laboratory data to construct a DM risk prediction model for patients with PTC.

To date, studies on predicting the risk of DM in thyroid cancer have been reported. However, most of these studies are based on the mining of data from the SEER database,[11–18] which may not reflect the real situation of the Chinese patient population, and use a single method (traditional logistic regression) for model building, which may not be capable of handling complex

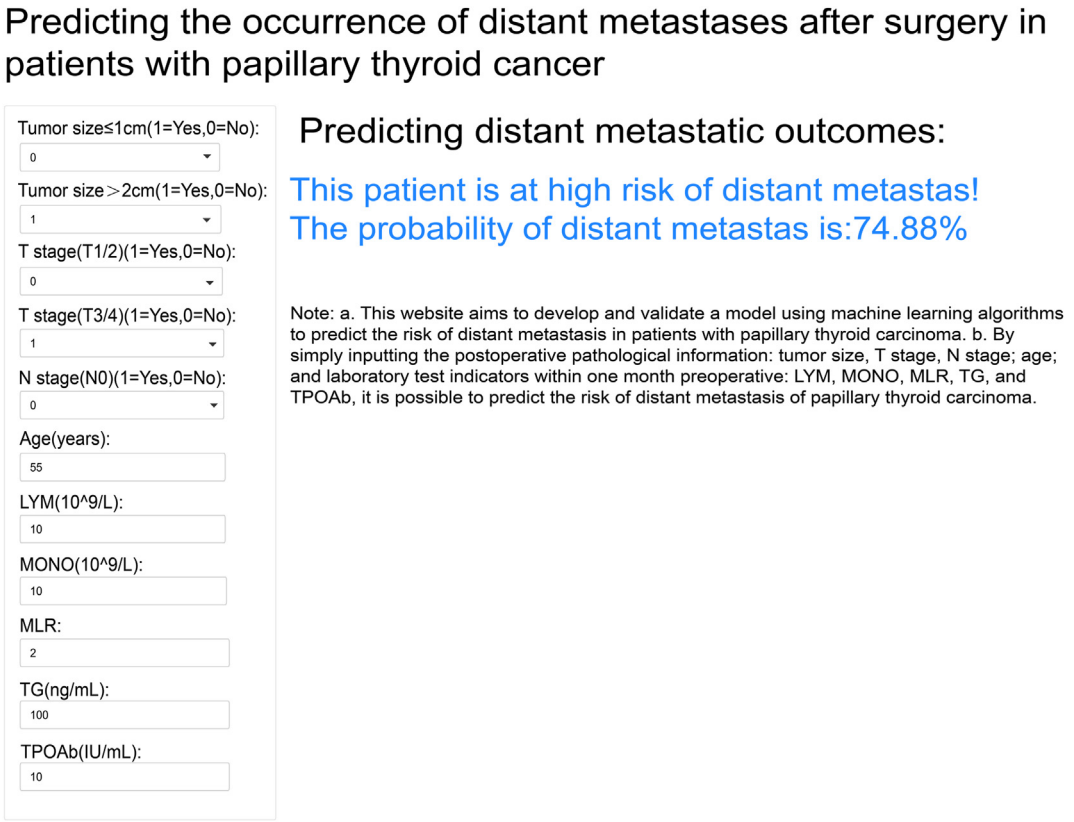

Fig. 7: The web-based calculator for predicting distant metastasis (DM) in papillary thyroid cancer (PTC) using this model. By simply inputting the postoperative pathological information: tumor size, T stage, N stage; age; and laboratory test indicators within one month preoperative: lymphocyte (LYM), monocyte (MONO), monocyte/lymphocyte ratio (MLR), thyroglobulin (TG), and thyroid peroxidase antibody (TPOAb), it is possible to predict the risk of DM of PTC.

relationships and hence affect prediction performance.[39] For clinicians and researchers, clinical EMR data are relatively objective, accurate, and easy to access. Combining EMR data with complex ML algorithms can facilitate the development of clinical prediction models.[40] Among the nine ML models, the RF model had the highest AUC, with good accuracy, calibration, and net benefit. The RF model also exhibited the best performance in the independent external validation set. The RF algorithm makes predictions by integrating multiple decision trees in combination with a voting mechanism, which can increase the prediction accuracy and stability of the model. The RF algorithm can capture nonlinear relationships in data, making this algorithm suitable for complex clinical data. In addition, by constructing and averaging many decision trees, the risk of the overfitting of a single decision tree is reduced.[41] Multiple studies have demonstrated that the RF method is very valuable for prediction models in the medical field.[42–44] In the present study, we employed the RF algorithm to develop a final model containing 11 features. These features can be easily acquired and evaluated during the hospitalization of patients with PTC, making this model a promising tool for effectively predicting the risk of DM in patients with PTC.

Due to the lack of unified guidelines or consensuses guiding the selection of features for prediction models, it is currently unclear how many features should be included in the model. Although more features may provide richer information for the prediction model, too many features may limit the clinical application of the model, while the inclusion of noncausal features may reduce prediction accuracy. To assist feature selection, we employed an RFE method. Our final model is a simple and convenient ML prediction model that can be easily used for clinical decision-making for patients with PTC.

Multivariate logistic regression analysis revealed that age, BMI, benign thyroid disease, tumor size, RBC count, MONO count, PLR, TG level, TPOAb level, T stage, and N stage were factors independently associated with the risk for DM in patients with PTC. Most studies have shown that a high BMI is associated with tumor progression.[45,46] However, some studies have suggested that a high BMI may have a protective effect against some types of cancer, and this effect may be associated with the higher nutrient reserves and muscle mass in patients with obese.[47] Laura reported that a higher BMI was associated with a lower risk of breast cancer in premenopausal women, possibly because a higher BMI affects hormone levels in the body, especially estrogen and insulin levels, and these changes provide a protective effect on premenopausal women.[48] Our study found that BMI is a protective factor for DM in PTC; there is a linear relationship between BMI and DM, and the risk of DM is reduced when BMI is > 18.73. It is well known that the risk of DM of thyroid cancer is related to tumor size, patient age, lymph node involvement, and tumor stage.[25,49] Notably, we detected a U-shaped association between age and DM in patients with PTC, and the risk of DM in PTC increased rapidly for patients aged <24.87 years or >50.77 years. Similarly, the study by Huang et al. included 111 patients with PTC, and the RCS curve also showed a U-shaped pattern between age and DM, with significantly higher incidences of DM in patients ≤21 years and >55 years old than that in patients 22–55 years old.[50] The multivariate logistic regression analysis in our study reached similar conclusions. In addition, our study revealed that patients with benign thyroid diseases were prone to DM. Cari et al. also reported that the risk of regional metastasis and DM in PTC increases after the diagnosis of hyperthyroidism and thyroiditis.[51] Huang et al.[52] studied the metabolic features of benign thyroid nodules and PTC and found that the metabolic features of the two overlapped, suggesting that benign thyroid diseases may affect the risk of metastasis in thyroid cancer. Our study found that RBC count was a protective factor against DM in PTC. Although previous studies have almost never reported a direct association between the RBC count and DM in PTC, the study by Jin et al.[53] investigated the impact of anemia during chemotherapy on the outcomes of patients with advanced epithelial ovarian cancer, finding that anemia is correlated with poor progression-free survival and overall survival, indirectly suggesting that anemia may promote tumor metastasis. Therefore, we speculate that patients with low RBC counts are in an anemic state that might promote the development of DM in patients with PTC. Our study revealed that the MONO count was an independent factor associated with the risk for DM in PTC. Currently, no study has reported the relationship between the MONO count and DM in PTC, but some studies have found that a low pretreatment LMR is associated with advanced clinicopathological features and poor outcomes in patients with pancreatic cancer, making it a prognostic predictor. Our study found that the PLR was also an independent factor associated with the risk for DM in PTC. Cao et al. reported that the higher the PLR is, the worse the clinicopathological features of PTC are, with larger tumor diameters, a higher N1 stage, and a higher TG level.[54] However, Elena's meta-analysis, which included 7599 patients with differentiated thyroid cancer (DTC), showed no association between the PLR and disease-free survival (DFS) in patients.[55] In addition, the preoperative TG level helps to predict the initial DM in patients with DTC. The study by Kim et al., which included 4735 patients with DTC, found that the preoperative TG level helps to predict the initial DM in patients with DTC, with a median preoperative TG level of 328.4 ng/mL in the initial DM group and 10.0 ng/mL in the non-DM group.[56] Similarly, a study from South Korea showed that preoperative TG measurement may help predict cervical lymph node metastasis. Our study also found

that the preoperative TG level was a factor associated with the risk for DM in PTC, showing a nonlinear relationship with DM, and that the risk of DM increased significantly when the TG level was >1.39 ng/mL. For TPOAb, the study by Shen et al. included 1126 patients with PTC, and multivariate logistic regression analysis revealed that TPOAb positivity is a protective factor against DM in PTC, with an odds ratio (OR) of 0.403 (95% CI 0.216–0.622, P < 0.001). In addition, subgroup analysis showed that combined positivity for TGAb and TPOAb is associated with fewer distant metastatic diseases.[57] Interestingly, TPOAb is also associated with DM in breast cancer. A study from Germany found that TPOAb positivity is associated with a significantly reduced incidence of DM in breast cancer, and that the TPOAb level is inversely proportional to the levels of conventional tumor markers CA-15-3 and CEA.[58] Similarly, our study found that the preoperative TPOAb level was a protective factor against DM in PTC and that it had a nonlinear relationship with DM; the risk of DM was significantly reduced when the TPOAb level was >0.96 U/mL.

Because of one-hot encoding, our final model actually included nine variables, namely, tumor size, TG level, TPOAb level, MLR, age, N stage, MONO count, T stage, and LYM count. The SHAP summary dot plot showed that the associations of tumor size, TG level, TPOAb level, age, N stage, MONO count, and T stage with the outcome of DM were consistent with the results of aforementioned multivariate analysis. In addition, the SHAP summary dot plot indicated that patients with higher LYM counts were less likely to have DM, while patients with higher MLRs were more prone to develop DM. A study by Cínthia et al. validated our results, finding that the MLR, NLR, and PLR are higher in patients with DM than in those without, and that the threshold of MLR for the diagnosis of DM in DTC is 0.21 (sensitivity = 80%, specificity = 45.2%, and accuracy = 57.9%).[59]

In this study, we employed multiple goodness-of-fit tests to evaluate the predictive performance of the ML models, including the Hosmer–Lemeshow (HL) test, AUC-ROC, and Calibration Curve. Although the HL test is a commonly used calibration tool for logistic regression models, its effectiveness may be limited in both large and small sample sizes.[34] Therefore, we introduced the Log–Loss test as a supplement. Unlike the HL test, Log–Loss does not rely on data grouping and provides a continuous error measurement, making it more sensitive in reflecting the accuracy of predicted probabilities. The results of the Log–Loss test were consistent with those of the AUC-ROC and Calibration Curve, further supporting the predictive performance of the models. Thus, by combining the HL test with Log–Loss, we were able to more comprehensively assess the calibration and discriminative ability of the models. It should also be noted that in this study, the training and test sets were sourced from two different medical institutions, which may account for the observed performance differences. The AUC of the RF model on the training set was exceptionally high (0.9999), suggesting potential overfitting, where the model may have learned not only the underlying patterns but also the noise in the training data. In contrast, the AUC from 10 rounds of 10-fold cross-validation (0.913 ± 0.028) provides a more realistic estimate of the model's generalizability. The AUC of 0.8996 on the independent test set, while slightly lower, still indicates excellent predictive performance, as an AUC close to 0.9 is considered very strong in clinical predictions. This demonstrates the model's robustness across different datasets. To further evaluate the model's stability and generalizability, we plan to include additional external validation cohorts in future research.

ML techniques are often referred to as "black boxes", making their prediction processes nearly impossible to explain.[60] This lack of transparency may cause clinicians to hesitate to use these techniques because they are unwilling to make medical decisions based on opaque information. However, a major advantage of the present study is that we used the SHAP method to reasonably elucidate the "black box" of the ML model. The SHAP method clarifies the functionality of the model by providing global and local explanations, detailing how the personalized input data are used to make specific predictions for individual patients. In addition, by using the convenient tools of the Shinyapps platform, we integrated out prediction model into a user-friendly online prediction platform for both doctors and patients. Another strength of the present study is the comparison of the prediction performance of different ML models for the risk of DM in PTC. The evaluation of the performance of the models in the external validation set and the comparison of the models also showed that the RF model had good predictive value for DM in PTC. Another strength of the present study is that the predictive factors included in the model were all routine items obtained for patients during hospitalization and were easy to obtain, providing feasibility for the promotion and application of the model in clinical practice.

We acknowledge several limitations of the present study. First, this was a retrospective study. Although we had strict inclusion and exclusion criteria, it was difficult to completely avoid bias in the research results. Second, the model was constructed based on data from Chinese patients. Although the performance of the model was externally validated, there is still no basis for the generalizability of the model for use in patients from other populations. This study did not measure participants' socioeconomic status or other structural factors (such as access to healthcare), which may influence the prognosis of patients with thyroid cancer. Future research should consider incorporating socioeconomic background and related variables to further enhance the generalizability of the model. Third, even though "big

data" are needed for creating a prediction model, there is currently no criterion for determining an appropriate sample size. Nevertheless, the good performance of the model in internal cross-validation and external validation indicated that the sample size in this study was appropriate and provided sufficient support for exploring DM in patients with PTC. Fourth, considering the generalizability of the model, the variables included in this study were all routine items that are easy to obtain in clinical practice. Some novel molecular markers associated with the DM in thyroid cancer, such as BRAF V600E and TERT mutations, were not included in the present study.[61,62] Fifth, the model's performance in predicting "other metastases" was lower, likely due to the smaller sample size of patients with this type of metastasis, leading to data imbalance and affecting the model's sensitivity and accuracy. In the future, increasing the sample size or applying data balancing techniques could improve the predictive performance. Additionally, this study did not incorporate imaging data, despite the critical role that imaging examinations play in the diagnosis and staging of thyroid cancer. The inclusion of imaging features, whether traditional radiomic features or deep learning-based features, could potentially enhance the accuracy and clinical applicability of the model. Despite these limitations, the impressive performance of our final prediction model is not overshadowed.

In conclusion, we successfully developed an explainable ML model to predict the risk of DM in patients with PTC based on clinical data easily extracted from EMRs. The final RF model exhibited excellent prediction ability in both internal and external validations. In the future, further prospective randomized controlled studies are needed to clarify whether individualized diagnosis and treatment plans and follow-up strategies developed based on the final prediction model can effectively improve the outcomes of patients with a high risk of DM in PTC.

Contributors

T. C. was responsible for the conceptualization of the paper. F.H., Y.Z., H.C., Y.W., X.P., R.H., Y.H., Z.L., T.C. accessed and verified the data. The investigation was conducted by all authors except Z.L. Methodology was developed by F.H. and T.C. Software was developed by F.H. and H.Z. Visualization was done by F.H. The original draft was written by F.H. Supervision was provided by Z.L. and T.C., who also reviewed and edited the paper. Z.L. and T.C. were responsible for the Funding acquisition and Project administration of the research. All authors agree to be fully accountable for ensuring the integrity and accuracy of the work and have read and approved the final manuscript. The corresponding author had full access to all data in the study and assumed final responsibility for the decision to submit the manuscript for publication.

Data sharing statement

The data analyzed and the codes used during the current study are available from the corresponding author on reasonable request.

Declaration of interests

We declare no competing interests.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant No. 81960426, Ting Chen; grant No. 82360345 and 82001986, Zhenhui Li), the Outstanding Youth Science Foundation of Yunnan Basic Research Project (grant No. 202401AY070001-316, Zhenhui Li), Yunnan Province Applied and Basic Research Foundation (grant No. 202401AT070008, Ting Chen), and Ten Thousand Talent Plans for Young Top-notch Talents of Yunnan Province (Ting Chen). The sponsors had no role in the design of the study, data collection, data analysis, data interpretation, report writing, or the decision to submit the report for publication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102913.

References

1 Boucai L, Zafereo M, Cabanillas ME. Thyroid cancer: a review. *JAMA*. 2024;331(5):425–435.
2 Chen DW, Lang BHH, McLeod DSA, Newbold K, Haymart MR. Thyroid cancer. *Lancet*. 2023;401(10387):1531–1544.
3 Pizzato M, Li M, Vignat J, et al. The epidemiological landscape of thyroid cancer worldwide: GLOBOCAN estimates for incidence and mortality rates in 2020. *Lancet Diabetes Endocrinol*. 2022;10(4):264–272.
4 Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China, 2016. *J Natl Cancer Center*. 2022;2(1):1–9.
5 Yang L, Shen W, Sakamoto N. Population-based study evaluating and predicting the probability of death resulting from thyroid cancer and other causes among patients with thyroid cancer. *J Clin Oncol*. 2013;31(4):468–474.
6 Wang LY, Palmer FL, Nixon IJ, et al. Multi-organ distant metastases confer worse disease-specific survival in differentiated thyroid cancer. *Thyroid*. 2014;24(11):1594–1599.
7 Hirsch D, Levy S, Tsvetov G, et al. Long-term outcomes and prognostic factors in patients with differentiated thyroid cancer and distant metastases. *Endocr Pract*. 2017;23(10):1193–1200.
8 Gomes-Lima CJ, Wu D, Rao SN, et al. Brain metastases from differentiated thyroid carcinoma: prevalence, current therapies, and outcomes. *J Endocr Soc*. 2019;3(2):359–371.
9 Sabet A, Binse I, Dogan S, et al. Distinguishing synchronous from metachronous manifestation of distant metastases: a prognostic feature in differentiated thyroid carcinoma. *Eur J Nucl Med Mol Imaging*. 2017;44(2):190–195.
10 Durante C, Haddy N, Baudin E, et al. Long-term outcome of 444 patients with distant metastases from papillary and follicular thyroid carcinoma: benefits and limits of radioiodine therapy. *J Clin Endocrinol Metab*. 2006;91(8):2892–2899.
11 Wang W, Liu J, Xu X, Huo L, Wang X, Gu J. A high-quality nomogram for predicting lung metastasis in newly diagnosed stage IV thyroid cancer: a population-based study. *Technol Cancer Res Treat*. 2023;22:15330338231167807.
12 Wang W, Shen C, Yang Z. Nomogram individually predicts the risk for distant metastasis and prognosis value in female differentiated thyroid cancer patients: a SEER-based study. *Front Oncol*. 2022;12:800639.
13 Bi J, Zhang H. Nomogram predicts risk and prognostic factors for lung metastasis of anaplastic thyroid carcinoma: a retrospective study in the Surveillance Epidemiology and End Results (SEER) database. *Transl Cancer Res*. 2023;12(12):3547–3564.
14 Liu W, Wang S, Ye Z, Xu P, Xia X, Guo M. Prediction of lung metastases in thyroid cancer using machine learning based on SEER database. *Cancer Med*. 2022;11(12):2503–2515.
15 Kuang HF, Lu WL. Predictive factors for lung metastasis in pediatric differentiated thyroid cancer: a clinical prediction study. *J Pediatr Endocrinol Metab*. 2024;37(3):250–259.
16 Li Y, Gao X, Guo T, Liu J. Development and validation of a nomogram for risk of pulmonary metastasis in non-papillary thyroid carcinoma: a SEER-based study. *Medicine (Baltimore)*. 2023;102(32):e34581.
17 Tong Y, Hu C, Huang Z, Fan Z, Zhu L, Song Y. Novel nomogram to predict risk of bone metastasis in newly diagnosed thyroid carcinoma: a population-based study. *BMC Cancer*. 2020;20(1):1055.
18 Liu WC, Li ZQ, Luo ZW, Liao WJ, Liu ZL, Liu JM. Machine learning for the prediction of bone metastasis in patients with

newly diagnosed thyroid cancer. *Cancer Med*. 2021;10(8):2802–2811.
19 Wang W, Shen C, Zhao Y, Sun B, Bai N, Li X. Identification and validation of potential novel biomarkers to predict distant metastasis in differentiated thyroid cancer. *Ann Transl Med*. 2021;9(13):1053.
20 Lan X, Bao H, Ge X, et al. Genomic landscape of metastatic papillary thyroid carcinoma and novel biomarkers for predicting distant metastasis. *Cancer Sci*. 2020;111(6):2163–2173.
21 Cheng X, Zhou Y, Xu S, et al. Risk-stratified distant metastatic thyroid cancer with clinicopathological factors and BRAF/TERT promoter mutations. *Exp Clin Endocrinol Diabetes*. 2023;131(11):577–582.
22 Huang H, Xu S, Wang X, Liu S, Liu J. Patient age is significantly related to distant metastasis of papillary thyroid microcarcinoma. *Front Endocrinol (Lausanne)*. 2021;12:748238.
23 Liu JB, Baugh KA, Ramonell KM, et al. Molecular testing predicts incomplete response to initial therapy in differentiated thyroid carcinoma without lateral neck or distant metastasis at presentation: retrospective cohort study. *Thyroid*. 2023;33(6):705–714.
24 Kim H, Shin JH, Hahn SY, et al. Prediction of follicular thyroid carcinoma associated with distant metastasis in the preoperative and postoperative model. *Head Neck*. 2019;41(8):2507–2513.
25 Khan U, Al Afif A, Aldaihani A, et al. Patient and tumor factors contributing to distant metastasis in well-differentiated thyroid cancer: a retrospective cohort study. *J Otolaryngol Head Neck Surg*. 2020;49(1):78.
26 Parvathareddy SK, Siraj AK, Iqbal K, et al. TERT promoter mutations are an independent predictor of distant metastasis in middle eastern papillary thyroid microcarcinoma. *Front Endocrinol*. 2022;13:808298.
27 Wijewardene A, Gill AJ, Gild M, et al. A retrospective cohort study with validation of predictors of differentiated thyroid cancer outcomes. *Thyroid*. 2022;32(10):1201–1210.
28 Chinese Society of Clinical Oncology (CSCO). Diagnosis and treatment guidelines for persistent/recurrent and metastatic differentiated thyroid cancer 2018 (English version). *Chin J Cancer Res*. 2019;31(1):99–116.
29 Haugen BR, Alexander EK, Bible KC, 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*. 2016;26(1):1–133.
30 Escanilla NS, Hellerstein L, Kleiman R, Kuang Z, Shull JD, Page D. Recursive feature elimination by sensitivity testing. *Proc Int Conf Mach Learn Appl*. 2018;2018:40–47.
31 Ravishankar H, Madhavan R, Mullick R, Shetty T, Marinelli L, Joel SE. Recursive feature elimination for biomarker discovery in resting-state functional connectivity. *Annu Int Conf IEEE Eng Med Biol Soc*. 2016;2016:4071–4074.
32 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st conference on neural information processing systems (NIPS 2017); Long Beach, CA, USA*. 2017.
33 Pepe M, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves. *STATA J*. 2009;9(1):1.
34 Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35(9):2052–2056.
35 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
36 Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. *Epidemiology*. 2014;25(1):114–121.
37 Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138.
38 Smith AD, Crippa A, Woodcock J, Brage S. Physical activity and incident type 2 diabetes mellitus: a systematic review and dose-response meta-analysis of prospective cohort studies. *Diabetologia*. 2016;59(12):2527–2545.
39 Coughlin SS, Kapuku G. Commentary: cancer survivorship and subclinical myocardial damage. *Am J Epidemiol*. 2022;191(3):367–368.
40 Goecks J, Jalili V, Heiser LM, Gray JW. How machine learning will transform biomedicine. *Cell*. 2020;181(1):92–101.
41 Denisko D, Hoffman MM. Classification and interaction in random forests. *Proc Natl Acad Sci USA*. 2018;115(8):1690–1692.
42 Moehring RW, Phelan M, Lofgren E, et al. Development of a machine learning model using electronic health record data to identify antibiotic use among hospitalized patients. *JAMA Netw Open*. 2021;4(3):e213460.
43 Goldstein BA, Cerullo M, Krishnamoorthy V, et al. Development and performance of a clinical decision support tool to inform resource utilization for elective operations. *JAMA Netw Open*. 2020;3(11):e2023547.
44 Zhang X, Yue P, Zhang J, et al. A novel machine learning model and a public online prediction platform for prediction of post-ERCP-cholecystitis (PEC). *EClinicalMedicine*. 2022;48:101431.
45 Pazaitou-Panayiotou K, Polyzos SA, Mantzoros CS. Obesity and thyroid cancer: epidemiologic associations and underlying mechanisms. *Obes Rev*. 2013;14(12):1006–1022.
46 Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet*. 2008;371(9612):569–578.
47 Strulov Shachar S, Williams GR. The obesity paradox in cancer-moving beyond BMI. *Cancer Epidemiol Biomarkers Prev*. 2017;26(1):13–16.
48 Garcia-Estevez L, Cortes J, Perez S, Calvo I, Gallegos I, Moreno-Bueno G. Obesity and breast cancer: a paradoxical and controversial relationship influenced by menopausal status. *Front Oncol*. 2021;11:705911.
49 Mao J, Zhang Q, Zhang H, Zheng K, Wang R, Wang G. Risk factors for lymph node metastasis in papillary thyroid carcinoma: a systematic review and meta-analysis. *Front Endocrinol (Lausanne)*. 2020;11:265.
50 Huang H, Ni S, Liu W, Wang X, Liu S. The U-Shaped association between age and distant metastasis in patients with papillary thyroid carcinoma. *Endocrine*. 2024;85(1):258–266.
51 Kitahara CM, D KRF, Jorgensen JOL, Cronin-Fenton D, Sorensen HT. Benign thyroid diseases and risk of thyroid cancer: a nationwide cohort study. *J Clin Endocrinol Metab*. 2018;103(6):2216–2224.
52 Huang FQ, Li J, Jiang L, et al. Serum-plasma matched metabolomics for comprehensive characterization of benign thyroid nodule and papillary thyroid carcinoma. *Int J Cancer*. 2019;144(4):868–876.
53 Kim JH, Lee JM, Ryu KS, et al. The prognostic impact of duration of anemia during chemotherapy in advanced epithelial ovarian cancer. *Oncologist*. 2011;16(8):1154–1161.
54 Cao J, He X, Li X, et al. The potential association of peripheral inflammatory biomarkers in patients with papillary thyroid cancer before radioiodine therapy to clinical outcomes. *Front Endocrinol*. 2023;14:1253394.
55 Russo E, Guizzardi M, Canali L, et al. Preoperative systemic inflammatory markers as prognostic factors in differentiated thyroid cancer: a systematic review and meta-analysis. *Rev Endocr Metab Disord*. 2023;24(6):1205–1216.
56 Kim H, Kim YN, Kim HI, et al. Preoperative serum thyroglobulin predicts initial distant metastasis in patients with differentiated thyroid cancer. *Sci Rep*. 2017;7(1):16955.
57 Shen CT, Zhang XY, Qiu ZL, et al. Thyroid autoimmune antibodies in patients with papillary thyroid carcinoma: a double-edged sword? *Endocrine*. 2017;58(1):176–183.
58 Farahati J, Roggenbuck D, Gilman E, et al. Anti-thyroid peroxidase antibodies are associated with the absence of distant metastases in patients with newly diagnosed breast cancer. *Clin Chem Lab Med*. 2012;50(4):709–714.
59 Riguetto CM, Barreto IS, Maia FFR, Assumpcao L, Zantut-Wittmann DE. Usefulness of pre-thyroidectomy neutrophil-lymphocyte, platelet-lymphocyte, and monocyte-lymphocyte ratios for discriminating lymph node and distant metastases in differentiated thyroid cancer. *Clinics (Sao Paulo)*. 2021;76:e3022.
60 The Lancet Respiratory Medicine. Opening the black box of machine learning. *Lancet Respir Med*. 2018;6(11):801.
61 Pan W, Tian Y, Zheng Q, et al. Oncogenic BRAF noncanonically promotes tumor metastasis by mediating VASP phosphorylation and filopodia formation. *Oncogene*. 2023;42(43):3194–3205.
62 Melo M, Gaspar da Rocha A, Batista R, et al. TERT, BRAF, and NRAS in primary thyroid cancer and metastatic disease. *J Clin Endocrinol Metab*. 2017;102(6):1898–1907.