



Construction and validation of machine learning models for predicting lymph node metastasis in cutaneous malignant melanoma: a large population-based study

Ling-Feng Lan^{1#^}, Yi-Long Kai^{1#}, Xiao-Ling Xu¹, Jun-Kun Zhang¹, Guang-Bo Xu¹, Yan-Bi Dai¹, Yan Shen¹, Hua-Ya Lu^{2^}, Ben Wang^{3^}

¹Department of Otolaryngology, The First Affiliated Hospital, Zhejiang University School of Medicine, Liangzhu Branch (The First People's Hospital of Yuhang District), Hangzhou, China; ²Department of Orthopedics, Ningbo Yinzhou Second Hospital, Ningbo, China; ³Department of Dermatology, Taizhou Women and Children's Hospital of Wenzhou Medical University, Taizhou, China

Contributions: (I) Conception and design: B Wang, HY Lu, LF Lan; (II) Administrative support: YB Dai, Y Shen; (III) Provision of study materials or patients: LF Lan, YL Kai, XL Xu; (IV) Collection and assembly of data: JK Zhang, GB Xu; (V) Data analysis and interpretation: LF Lan, YL Kai, XL Xu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Ben Wang, MD. Department of Dermatology, Taizhou Women and Children's Hospital of Wenzhou Medical University, 97-115 Weier Road, Taizhou 318000, China. Email: wb15867699118@163.com; Hua-Ya Lu, BD. Department of Orthopedics, Ningbo Yinzhou Second Hospital, 998# Qianhe Bei Road, Yinzhou District, Ningbo 315100, China. Email: luhuya2018@sina.com.

Background: Lymph node status is essential for determining the prognosis of cutaneous malignant melanoma (CMM). This study aimed to develop a machine learning (ML) model for predicting lymph node metastases (LNM) in CMM.

Methods: We gathered data on 6,196 patients from the Surveillance, Epidemiology, and End Results (SEER) database, including known clinicopathologic variables, using six ML algorithms, including logistic regression (LR), support vector machine (SVM), Complement Naive Bayes (CNB), Extreme Gradient Boosting (XGBoost), RandomForest (RF), and k-nearest neighbor algorithm (kNN), to predict the presence of LNM in CMM. Subsequently, we established prediction models. The utilization of the adaptive synthetic (ADASYN) method served to address the challenge posed by imbalanced data. We assessed prediction model performance in terms of average precision (AP), sensitivity, specificity, accuracy, F1 score, precision-recall curves, calibration plots, and decision curve analysis (DCA). Furthermore, employing SHapley Additive exPlanation (SHAP) analysis resulted in the creation of visualized explanations tailored to individual patients.

Results: Among the 6,196 CMM cases, 19.9% (n=1,234) presented with LNM. The XGBoost model showed the best predictive performance when compared with the other algorithms (AP of 0.805). XGBoost showed that age and Breslow thickness were the two most important factors related to LNM.

Conclusions: The XGBoost model predicted LNM of CMM with a high level of precision. We hope that this model could assist surgeons in accurately evaluating surgical approaches and determining the extent of surgery, while also guiding the subsequent adjuvant therapies, thereby improving the prognosis of patients.

Keywords: Cutaneous malignant melanoma (CMM); lymph node metastasis (LNM); machine learning (ML); shapley additive explanation (SHAP); Surveillance, Epidemiology, and End Results (SEER)

Submitted Sep 11, 2024. Accepted for publication Jan 03, 2025. Published online Feb 18, 2025.

doi: 10.21037/tcr-24-1672

View this article at: <https://dx.doi.org/10.21037/tcr-24-1672>

[^] ORCID: Ling-Feng Lan, 0009-0008-6023-2627; Hua-Ya Lu, 0009-0004-2846-2570; Ben Wang, 0009-0006-8031-6065.

Introduction

In the United States, cutaneous malignant melanoma (CMM) ranks as the fifth most prevalent cancer, exhibiting a discernibly escalating incidence, encompassing 100,640 newly reported cases in the year 2023 (1,2). Over the next 20 years, the global incidence rates for melanoma could increase (3). The 5-year overall survival rate of CMM is 94% (4). However, lymph node metastasis (LNM) is associated with an increased risk of melanoma mortality in melanoma patients (5). For melanoma, the acknowledgment of nodal metastases, whether present or absent, has been identified as a crucially significant determinant (6,7). Relevant factors for melanoma with LNM incidence and prognosis include age, ulceration, location, and Breslow thickness (8-12). The Surveillance, Epidemiology, and End Results (SEER) Program database, which encompasses approximately 28% of the United States population, facilitates comprehensive analyses of rare cancers due to its extensive coverage (13).

Machine learning (ML) algorithms are powerful tools for processing and analyzing data, uncovering relationships, and making informed decisions (14,15). ML in healthcare data analysis has a wide range of applications (16,17). These applications help healthcare professionals make more accurate diagnoses, personalize treatments, reduce costs, and improve patient outcomes (17,18). Hence, this study utilized ML techniques to perform a comparative evaluation

of varied predictive models aimed at discerning LNM in instances of CMM. We present this article in accordance with the TRIPOD reporting checklist (available at <https://tcr.amegroups.com/article/view/10.21037/tcr-24-1672/rc>).

Methods

Data source and study population

We included records from patients diagnosed with malignant melanoma from 2010 to 2018. “The International Classification of Diseases for Oncology (ICD-O-3) Hist/ behave, malignant” was used to select malignant melanoma patients. “Year of diagnosis” ranging from 2010 to 2018. “Derived AJCC N Stage 7th”, “CS site-specific factor 1”, and “CS site-specific factor 2” were also employed for patient screening.

Inclusion criteria: (I) years of diagnosis ranging from 2010 to 2018; (II) histological codes limited to 8721/3; and (III) primary site limit C440 to C449. The exclusion criteria: (I) multiple primary cancers; (II) survival time less than 1 month; (III) lack of information on race, Breslow thickness, ulceration, primary tumor site, or N stage.

Race was stratified into white, black, and other. Breslow thickness was divided into five groups: <1, 1 to <2, 2 to <3, 3 to <4, and ≥4 mm. Tumor location was divided into three groups: head and neck, trunk, and limbs. The non-parametric missForest methodology was employed for the imputation of missing data. At the same time, a heatmap was utilized to visually represent the outcomes of Pearson’s correlation test, facilitating the exploration of relationships between variables (19). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Imbalanced data processing

In the dataset under consideration, 1,234 patients (19.9%) presented with LNM, while 4,962 patients (80.1%) exhibited no LNM. The substantial class imbalance may compromise the predictive capability of classifiers, as balanced data is widely recognized to enhance prediction performance. To address this, we employed the adaptive synthetic (ADASYN) oversampling method, a well-established technique in ML for handling imbalanced data (20,21). Our implementation involved three different oversampling percentages: 200%, 250%, and 300%.

Highlight box

Key findings

- The Extreme Gradient Boosting (XGBoost) model predicting lymph node metastasis (LNM) was established in patients with cutaneous malignant melanoma (CMM).

What is known and what is new?

- The global incidence rate for melanoma is increasing. LNM has been correlated with dismal prognosis in patients suffering from CMM.
- The XGBoost model shows superior predictive efficacy.

What is the implication, and what should change now?

- An accurate machine learning-based prediction model can help clinicians during clinical decision-making.
- Further prospective studies with larger sample sizes and more detailed clinical information are warranted to improve the accuracy and applicability of our model.

Establishment and evaluation of the predictive model

In this investigation, a selection of six ML algorithms was undertaken to predict distant metastasis in CMM patients. The formulated models encompass the logistic regression (LR) model, support vector machine (SVM), Complement Naive Bayes (CNB), Extreme Gradient Boosting (XGBoost) model, RandomForest (RF), and the k-nearest neighbor algorithm (kNN). The ADASYN strategy was applied to enhance classifier performance and address the imbalanced dataset. Subsequently, the dataset will be randomly split into training and testing sets in a 7:3 ratio. We employed the K-fold cross-validation method within the training dataset to perform multiple rounds of training and validation. In K-fold cross-validation, the original dataset is divided at random into k equally sized groups. Out of these, one group is reserved for model validation, while the others serve as the training dataset (22). In our case, the parameter k has been set to 10.

The validation set served the purpose of fine-tuning model parameters, while the test set was employed to assess the model's performance. The clinical assessment of the prediction model encompassed three key quality measurements: discrimination, calibration, and clinical usefulness. Initially, discrimination was quantified through precision-recall curve analysis. Subsequently, the model's performance was evaluated using calibration plots to gauge the extent of deviation between calibration and actual events. Following this, the clinical utility was assessed via decision curve analysis (DCA), which computed net benefits across various threshold probabilities. Additionally, the evaluation of the six models included the examination of confusion matrix metrics such as average precision (AP), accuracy, sensitivity, specificity, and F1-score.

The challenges associated with interpreting results in ML are well-acknowledged. To address this issue, the SHapley Additive exPlanation (SHAP) method, proposed by Lundberg *et al.*, emerged as a game-theoretic approach providing reliable, rapid, and computationally efficient explanations for the output of any ML model (23). Crucially, the SHAP approach facilitated the prioritization of predictors based on their SHAP values, wherein higher values positively influenced the ML model's output. Conversely, lower values had a negative impact.

Statistical analysis

All statistical analyses were conducted using R (version

3.6.8) and Python (version 3.7). Categorical variables were expressed as frequency (percentage, %) and compared using the Chi-squared or Fisher's exact test. Continuous variables were expressed as median, and standard deviation (SD), and the Wilcoxon rank sum test was used to compare groups. The results were considered statistically significant when the two-sided $P < 0.05$.

Results

Patient baseline characteristics

During the study period, 6,196 patients with CMM were diagnosed from 2010 to 2018, and 19.9% of these were diagnosed with LNM disease at presentation. The detailed characteristics of patients are listed in *Table 1*.

Feature analysis

Employing Pearson correlation analysis, the interplay between each variable was assiduously examined. The resultant correlation heat map, elucidated in *Figure 1*, unveiled a weak relationship between several clinicopathological variables. Significantly discernible was the moderate correlation discerned between Breslow thickness and ulceration.

The consequential import of employing the ADASYN technique in augmenting balanced datasets stands manifest within the purview of *Table 2*. Notably, the ADASYN methodology evinced a marked enhancement in the AP values characterizing the classification models. The preeminence of the XGBoost classifier, attaining pinnacle precision values during validation set for 200% (0.678), 250% (0.724), and 300% (0.753) in validation set.

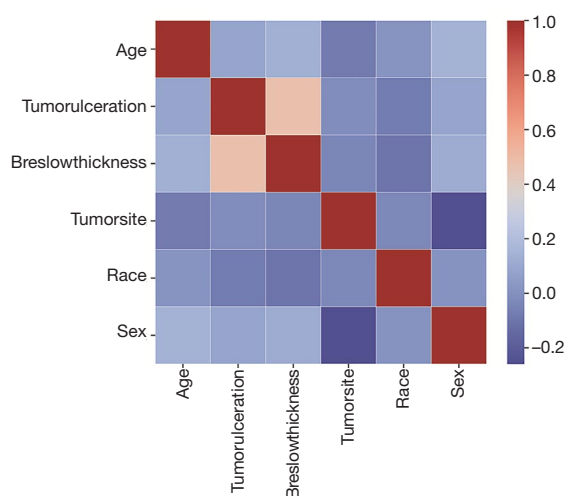
Model development and evaluation

In the RandomForest algorithm, the AP value of the training set is significantly greater than that of the validation set (*Figure 2*). RandomForest algorithm is highly likely to exhibit overfitting, while the XGBoost algorithm may have relatively good stability. The XGBoost algorithm yielded the best prediction performance (*Figure 2*). Subsequent evaluation through precision-recall curves, calibration plots, and DCA for the validation set further fortifies the scrutiny applied to the prediction model's efficacy. The calibration plots of the validation set manifest outstanding concordance between predictive probabilities and the observed risk of

Table 1 Demographic and clinicopathologic variables of the whole cohort grouped by metastasis status

Characteristics	All (n=6,196)	Non-lymph node metastasis (n=4,962)	Lymph node metastasis (n=1,234)	P
Age (years)	62.000 [51.000, 73.000]	63.000 [51.000, 73.000]	60.000 [48.000, 69.000]	<0.001
Tumor ulceration				<0.001
No	4,619 (74.548)	4,010 (80.814)	609 (49.352)	
Yes	1,577 (25.452)	952 (19.186)	625 (50.648)	
Breslow thickness (mm)				<0.001
<1	2,663 (42.979)	2,531 (51.008)	132 (10.697)	
1 to <2	1,467 (23.677)	1,202 (24.224)	265 (21.475)	
2 to <3	686 (11.072)	482 (9.714)	204 (16.532)	
3 to <4	413 (6.666)	265 (5.341)	148 (11.994)	
≥4	967 (15.607)	482 (9.714)	485 (39.303)	
Tumor site				<0.001
Trunk	1,442 (23.273)	1,197 (24.123)	245 (19.854)	
Limbs	2,060 (33.247)	1,600 (32.245)	367 (29.741)	
Head and neck	2,694 (43.480)	2,165 (43.632)	622 (50.405)	
Race				<0.001
Black	25 (0.403)	11 (0.222)	14 (1.135)	
Other	75 (1.210)	51 (1.028)	24 (1.945)	
White	6,096 (98.386)	4,900 (98.751)	1,196 (96.921)	
Sex				<0.001
Female	2,356 (38.025)	1,956 (39.420)	400 (32.415)	
Male	3,840 (61.976)	3,006 (60.580)	834 (67.585)	

Data are presented as median [IQR] or n (%). IQR, interquartile range.

**Figure 1** Correlation between factors. The depth of color indicates the magnitude of correlation.**Table 2** Number of instances increased by the ADASYN technique

Percentage of ADASYN increase	Non-lymph node metastasis	Lymph node metastasis
0%	4,962 (80.10)	1,234 (19.90)
200%	4,962 (57.27)	3,702 (42.73)
250%	4,962 (53.46)	4,319 (46.54)
300%	4,962 (50.13)	4,936 (49.87)

Data are presented as n (%). Three different oversampling percentages: 200%, 250%, and 300%. ADASYN, adaptive synthetic.

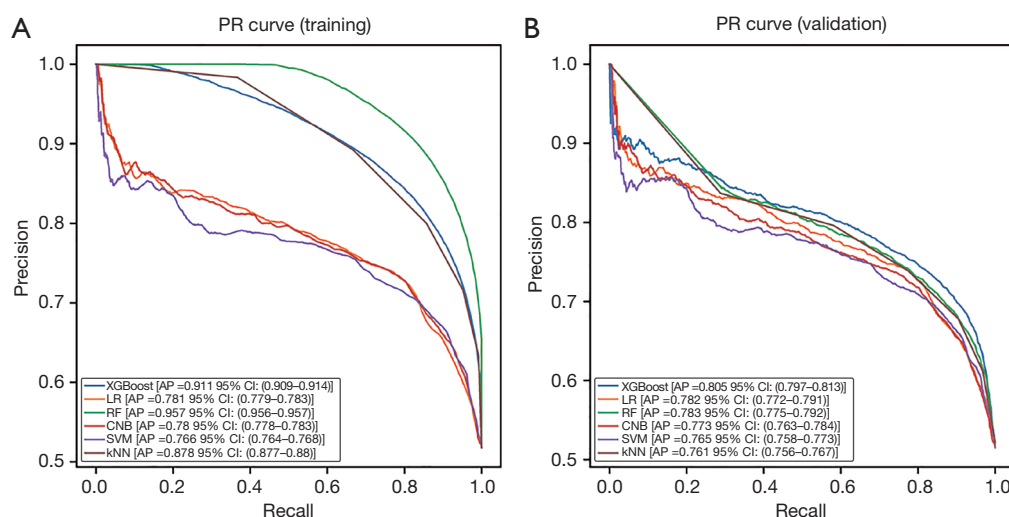


Figure 2 Evaluation of the prediction models for lymph node metastases in cutaneous malignant melanoma for the training set (A), and validation set (B). The average precision-recall curves, indicating the trade-off between precision and recall. PR, precision-recall; AP, average precision; XGBoost, Extreme Gradient Boosting; LR, logistic regression; RF, RandomForest; CNB, Complement Naive Bayes; SVM, support vector machine; kNN, the k-nearest neighbor algorithm; CI, confidence interval.

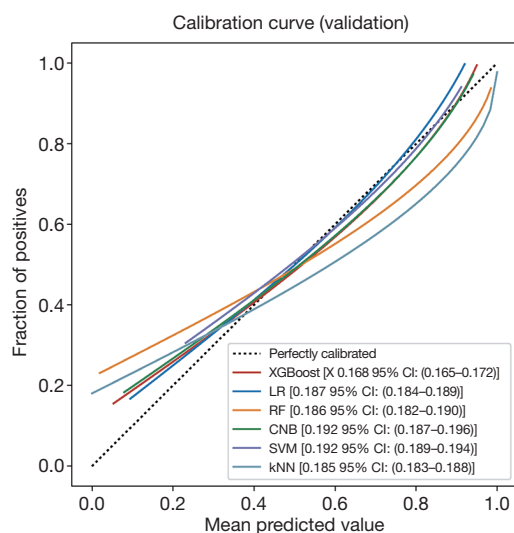


Figure 3 Examples of calibration plots (Brier Score) for predicting lymph node metastases with various models: XGBoost, LR, RF, CNB, SVM, and kNN. The 45° straight line on each graph represents the perfect match between the observed (y-axis) and predicted (x-axis) survival probabilities. A closer distance between two curves indicates greater accuracy. XGBoost, Extreme Gradient Boosting; LR, logistic regression; RF, RandomForest; CNB, Complement Naive Bayes; SVM, support vector machine; kNN, the k-nearest neighbor algorithm; CI, confidence interval.

distant metastasis within the XGBoost model (Figure 3). Subsequently, in our study, DCA was constructed for the six models (Figure 4), revealing that each model exhibited a discernible net clinical benefit compared to an approach of universal treatment or no treatment. Notably, the XGBoost model consistently demonstrated the utmost net benefit across the entirety of threshold probabilities.

Table 3 presents the evaluation measures of the confusion matrices for all prediction models, while Table 4 displays the k-fold cross-validation accuracies (k=10) associated with each model. The results suggest that the XGBoost model demonstrated the highest accuracy in k-fold cross-validation compared to the other models examined. Additionally, the predictive model utilizing the XGBoost model exhibited superior performance, as illustrated in Figure 5.

Model interpretation

Figure 6 elucidates the SHAP summary plot of the predictive model, delineating six features sorted by their impact on metastatic status. Higher SHAP values delineate augmented risks of distant metastasis, with a color spectrum from red signifying heightened feature values, purple representing proximation to the overall average, and blue

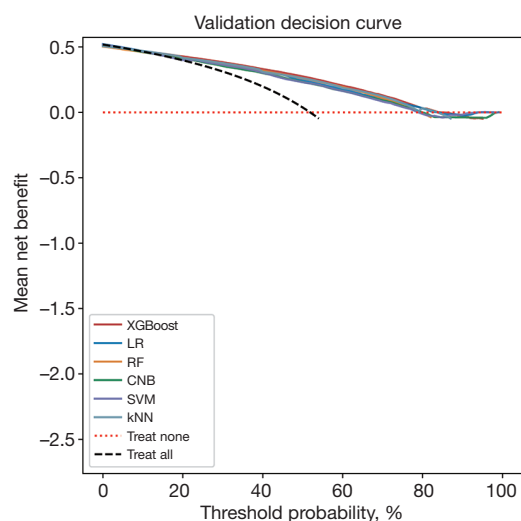


Figure 4 Decision curves of various models: XGBoost, LR, RF, CNB, SVM, and kNN. XGBoost, Extreme Gradient Boosting; LR, logistic regression; SVM, support vector machine; CNB, Complement Naive Bayes; RF, RandomForest; kNN, the k-nearest neighbor algorithm.

Table 4 The k-fold cross-validation accuracies (k=10) of all six prediction models

Model	k-fold accuracy
XGBoost	0.754 (0.748–0.760)
LR	0.742 (0.736–0.748)
RF	0.744 (0.739–0.749)
CNB	0.734 (0.727–0.741)
SVM	0.727 (0.719–0.734)
kNN	0.703 (0.698–0.708)

Data are presented as the estimated value with its 95% confidence interval. XGBoost, Extreme Gradient Boosting; LR, logistic regression; SVM, support vector machine; CNB, Complement Naive Bayes; RF, RandomForest; kNN, the k-nearest neighbor algorithm.

Table 3 Evaluation of the performance of classification models on imbalance dataset using ADASYN technique in validation set

Model	ADASYN	Precision	Accuracy	Sensitivity	Specificity	F1-score
XGBoost	200%	0.678 (0.667–0.689)	0.748 (0.740–0.756)	0.801 (0.778–0.823)	0.714 (0.686–0.742)	0.734 (0.724–0.744)
	250%	0.724 (0.713–0.735)	0.752 (0.745–0.759)	0.805 (0.784–0.826)	0.715 (0.699–0.731)	0.762 (0.750–0.774)
	300%	0.753 (0.743–0.764)	0.754 (0.748–0.760)	0.824 (0.800–0.848)	0.697 (0.678–0.715)	0.787 (0.775–0.798)
LR	200%	0.643 (0.633–0.652)	0.726 (0.718–0.733)	0.774 (0.753–0.794)	0.701 (0.689–0.714)	0.702 (0.688–0.716)
	250%	0.695 (0.685–0.706)	0.733 (0.727–0.739)	0.754 (0.741–0.766)	0.722 (0.714–0.730)	0.723 (0.716–0.731)
	300%	0.738 (0.731–0.746)	0.742 (0.736–0.748)	0.785 (0.774–0.796)	0.701 (0.691–0.710)	0.761 (0.754–0.768)
RF	200%	0.676 (0.664–0.687)	0.745 (0.741–0.750)	0.762 (0.741–0.784)	0.742 (0.724–0.761)	0.716 (0.703–0.730)
	250%	0.713 (0.698–0.728)	0.738 (0.728–0.747)	0.797 (0.771–0.823)	0.692 (0.657–0.726)	0.752 (0.740–0.764)
	300%	0.744 (0.737–0.750)	0.744 (0.739–0.749)	0.804 (0.776–0.832)	0.693 (0.661–0.725)	0.772 (0.760–0.785)
CNB	200%	0.647 (0.633–0.660)	0.728 (0.720–0.737)	0.777 (0.761–0.792)	0.708 (0.695–0.721)	0.705 (0.694–0.717)
	250%	0.692 (0.681–0.704)	0.728 (0.725–0.732)	0.779 (0.768–0.789)	0.688 (0.673–0.703)	0.733 (0.726–0.740)
	300%	0.724 (0.714–0.733)	0.734 (0.727–0.741)	0.785 (0.772–0.797)	0.688 (0.668–0.708)	0.753 (0.744–0.762)
SVM	200%	0.650 (0.641–0.660)	0.733 (0.727–0.738)	0.792 (0.778–0.806)	0.696 (0.684–0.707)	0.714 (0.705–0.723)
	250%	0.665 (0.658–0.673)	0.718 (0.714–0.722)	0.808 (0.797–0.820)	0.645 (0.634–0.655)	0.730 (0.721–0.738)
	300%	0.703 (0.695–0.711)	0.727 (0.719–0.734)	0.817 (0.785–0.850)	0.638 (0.605–0.670)	0.755 (0.740–0.771)
kNN	200%	0.759 (0.747–0.770)	0.713 (0.706–0.719)	0.753 (0.705–0.801)	0.722 (0.674–0.770)	0.754 (0.730–0.779)
	250%	0.776 (0.765–0.787)	0.706 (0.700–0.712)	0.761 (0.732–0.790)	0.714 (0.685–0.742)	0.768 (0.755–0.780)
	300%	0.796 (0.789–0.803)	0.703 (0.698–0.708)	0.759 (0.727–0.791)	0.717 (0.691–0.743)	0.776 (0.758–0.795)

Data are presented as the estimated value with its 95% confidence interval. ADASYN, adaptive synthetic; XGBoost, Extreme Gradient Boosting; LR, logistic regression; SVM, support vector machine; CNB, Complement Naive Bayes; RF, RandomForest; kNN, the k-nearest neighbor algorithm.

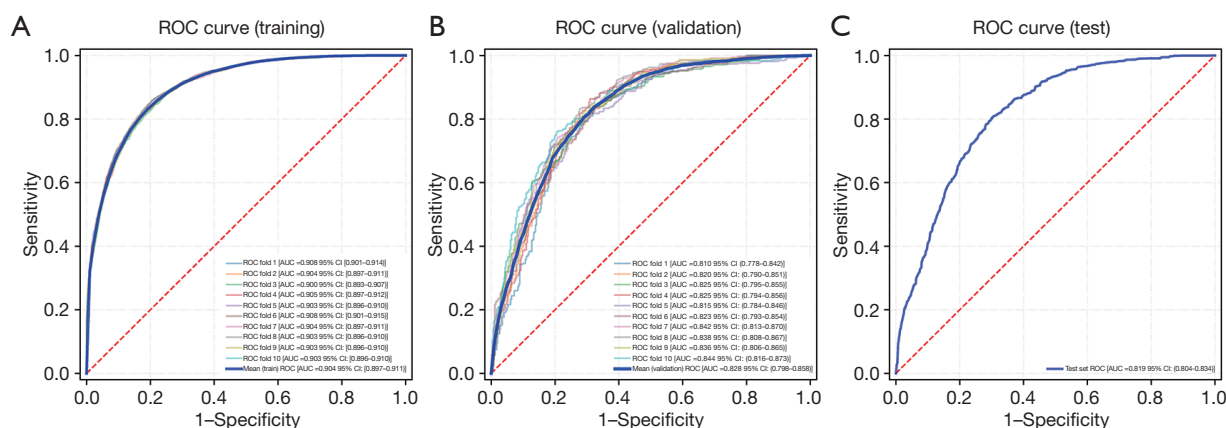


Figure 5 ROC curves of Extreme Gradient Boosting for the training (A), validation (B), and test (C) set. ROC, receiver operating characteristic; AUC, area under the curve; CI, confidence interval.

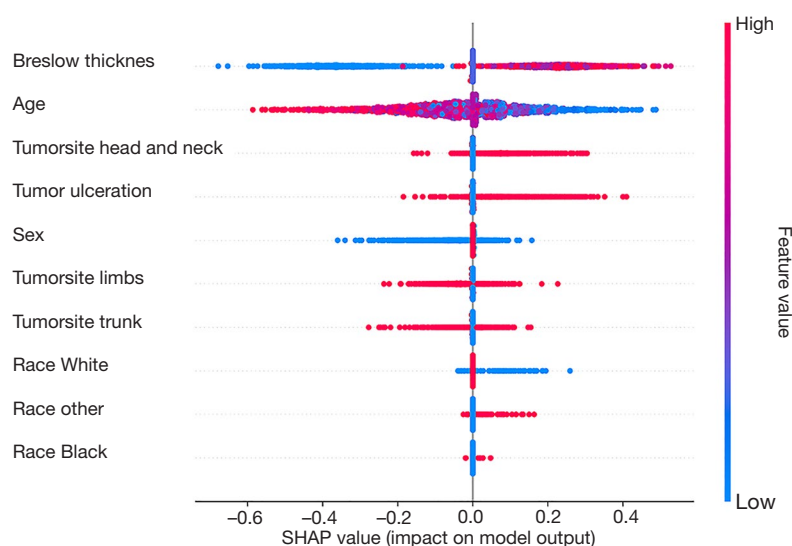


Figure 6 Summary plots for SHAP values. For each feature, one point corresponds to a single patient. A point's position along the x-axis represents the impact that feature had on the model's output for that specific patient. The redder the color indicates that the value is greater, and the bluer the color indicates that the value is smaller. Features are arranged along the y-axis based on their importance, which is given by the mean of their absolute Shapley values. The higher the feature is positioned in the plot, the more important it is for the model. SHAP, SHapley Additive exPlanation.

connoting diminished feature values.

Discussion

In this investigation, 19.9% of patients with CMM presented with metastatic lymph node disease at diagnosis. When melanoma patients develop LNM, the 5-year survival rate significantly drops from over 99% for localized disease

to 74% for regional disease (4). This indicates that LNM is a significant factor for poor prognosis in melanoma and greatly affects the survival rate of patients (24).

Sentinel lymph node biopsy is a widely accepted prognostic investigation in the management of malignant melanoma. Under the prevailing selection criteria, approximately 80–85% of patients undergoing sentinel lymph node biopsy exhibit negative findings upon

pathological evaluation of the sentinel lymph node (25). The existing evidence indicates that patients with negative findings in the sentinel lymph node do not experience survival benefits from the procedure. The sentinel lymph node is regarded as one of the most important prognostic indicators for assessing survival in individuals diagnosed with melanoma (25). Consequently, it is imperative to discern the risk factors correlated with regional nodal metastasis in individuals with melanoma and thus contribute to early detection, as well as prognosis assessment.

The ML model utilized clinical data, encompassing variables such as age, gender, race, Breslow thickness, ulceration, tumor site, and tumor location. Notably, the XGBoost model demonstrated superior performance compared to all other models in both the training and validation sets. As a result, the XGBoost model outperformed all other models in training and validation sets.

Within the current investigation, SHAP analysis revealed that the two most informative features in the model were Breslow thickness and age. The measurement of Breslow thickness has been recognized as a crucial factor in forecasting the likelihood of LNM in melanoma patients (12,26), and it is intimately associated with the risk of mortality (27). Lesions with a Breslow thickness of less than 0.76 mm rarely tend to metastasize (28). Our study also showed that only 4.96% of patients with a thickness of less than 1 mm had LNM. Our results also demonstrated that LNM rates varied by Breslow thickness, with only 132 (10.7%) of 1,234 positive LNM patients having Breslow thickness <1 mm.

Cavanaugh-Hussey *et al.* demonstrated a positive correlation between advancing age and a heightened incidence of melanoma-related mortality while concurrently observing a lower incidence of LNM (29). Isaksson *et al.* also showed that younger age is identified as a significant risk factor for sentinel lymph node positivity (30). It is consistent with the findings of our study. Some authors found that there are age-related changes in immunologic surveillance, including decreased lymphatic flow to nodes or nodal involution (31-33).

Several studies have confirmed that the notion that there is a significant difference between male and female patients in terms of the incidence rate of melanoma, the performance of primary tumors, and mortality (34-37). Direct regional LNM is more likely to occur in males (35,38). Hormone or immune factors might be the reason for the differences between the sexes.

Studies have shown that the anatomical location of the primary tumor has a significant impact on the risk of LNM in melanoma (12,39). Many recent studies confirmed the protective effect of the location of the primary tumor on the risk of LNM. Melanoma in the head and neck region is typically more prone to LNM than melanoma located in other anatomical sites. This may be attributed to the rich and complex network of capillaries and lymphatic drainage systems in the head and neck, which can promote the growth of ulcers, and tumor metastasis (12,40,41).

Previous studies have shown a specific correlation between the occurrence of ulcers and LNM, and sentinel lymph node biopsy is recommended (42,43). However, some studies have shown that the impact of ulcers on LNM is not statistically significant (30).

When a tumor breaches the surface of the skin and forms an ulcer, it indicates that the tumor has penetrated the skin's protective barrier. It becomes exposed to the external environment, and the ulcerated area is typically affected by inflammation and infection. This localized inflammation attracts immune cells and other inflammatory mediators, thereby providing more opportunities for cancer cells to enter the lymphatic system. Furthermore, since the ulcer disrupts the integrity of normal skin tissue, it may create new pathways around the malignant melanoma, making it easier for cancer cells to enter the lymphatic vessels (44,45).

Traditional regression models are often more interpretable, as the coefficients of variables have a clear meaning in terms of their impact on the outcome. Prior endeavors aimed at enhancing CMM prognostic prediction have predominantly relied on parametric regression models, a prevalent choice in clinical studies owing to their simplicity and interpretability. Traditional regression methods, such as linear regression, LR, and Poisson regression, are based on mathematical equations that model the relationship between independent and dependent variables. They make certain assumptions about the data distribution, linearity, and functional form. However, ML algorithms, including regression-based models like linear regression within ML, are data-driven and focus on learning patterns and relationships from data without strict assumptions about linearity or distribution. ML models, including regression models like support vector regression or RF regression and XGBoost, are more flexible and can handle complex relationships and non-linear patterns without requiring strict assumptions. Within both the training and validation sets, our ML model, utilizing XGBoost, exhibited a notable predictive performance for

assessing the risk of LNM in patients with CMM, with AP values of 0.911 and 0.805, respectively.

There were several limitations. First, the SEER database is a retrospective cohort; there are inevitably missing data that result in a reduced sample size. Prospective studies should be performed to validate our findings further. Second, although the nomogram was built using a large cohort and was validated internally, the predictive model should be validated in another database. Finally, our study did not include genetic markers because these data were not collected in the SEER.

Conclusions

ML using large databases can play an essential role in establishing treatment strategies. The XGBoost model exhibited heightened predictive efficacy for the prediction of LNM compared to other models. We hope that this model could assist surgeons in accurately evaluating surgical approaches and determining the extent of surgery, while also guiding the subsequent adjuvant therapies, thereby improving the prognosis of patients.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at <https://tcr.amegroups.com/article/view/10.21037/tcr-24-1672/rc>

Peer Review File: Available at <https://tcr.amegroups.com/article/view/10.21037/tcr-24-1672/prf>

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at <https://tcr.amegroups.com/article/view/10.21037/tcr-24-1672/coif>). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as

revised in 2013).

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Melanoma of the Skin Statistics American Cancer Society—Cancer Facts and Statistics. American Cancer Society 2023. Available online: www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html
2. Siegel RL, Miller KD, Wagle NS, et al. Cancer statistics, 2023. *CA Cancer J Clin* 2023;73:17-48.
3. Skin cancer World Cancer Research Fund International. Available online: <https://www.wcrf.org/dietandcancer/skin-cancer/>
4. SEER*Explorer. An interactive website for SEER cancer statistics Surveillance Research Program, National Cancer Institute. 2023. Available online: <https://seer.cancer.gov/explorer/>
5. Faries MB, Han D, Reintgen M, et al. Lymph node metastasis in melanoma: a debate on the significance of nodal metastases, conditional survival analysis and clinical trials. *Clin Exp Metastasis* 2018;35:431-42.
6. Song X, Zhao Z, Barber B, et al. Overall survival in patients with metastatic melanoma. *Curr Med Res Opin* 2015;31:987-91.
7. Han D, van Akkooi ACJ, Straker RJ 3rd, et al. Current management of melanoma patients with nodal metastases. *Clin Exp Metastasis* 2022;39:181-99.
8. Stassen RC, Maas CCHM, van der Veldt AAM, et al. Development and validation of a novel model to predict recurrence-free survival and melanoma-specific survival after sentinel lymph node biopsy in patients with melanoma: an international, retrospective, multicentre analysis. *Lancet Oncol* 2024;25:509-17.
9. Roulin D, Matter M, Bady P, et al. Prognostic value of sentinel node biopsy in 327 prospective melanoma patients from a single institution. *Eur J Surg Oncol* 2008;34:673-9.
10. Song JY, Ryu YJ, Lee HK, et al. Risk factors for sentinel lymph node metastasis in Korean acral and non-acral

- melanoma patients. *Pigment Cell Melanoma Res* 2024;37:332-42.
11. Wu PC, Chen YC, Chen HM, et al. Prognostic factors and population-based analysis of melanoma with sentinel lymph node biopsy. *Sci Rep* 2021;11:20524.
 12. Han AY, John MAS. Predictors of Nodal Metastasis in Cutaneous Head and Neck Cancers. *Curr Oncol Rep* 2022;24:1145-52.
 13. Tang X, Zhou X, Li Y, et al. A Novel Nomogram and Risk Classification System Predicting the Cancer-Specific Survival of Patients with Initially Diagnosed Metastatic Esophageal Cancer: A SEER-Based Study. *Ann Surg Oncol* 2019;26:321-8.
 14. Mirza B, Wang W, Wang J, et al. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes (Basel)* 2019;10:87.
 15. Oliveira AL. Biotechnology, Big Data and Artificial Intelligence. *Biotechnol J* 2019;14:e1800613.
 16. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019;20:e262-73.
 17. Feng JW, Ye J, Qi GF, et al. LASSO-based machine learning models for the prediction of central lymph node metastasis in clinically negative patients with papillary thyroid carcinoma. *Front Endocrinol (Lausanne)* 2022;13:1030045.
 18. Bai BL, Wu ZY, Weng SJ, et al. Application of interpretable machine learning algorithms to predict distant metastasis in osteosarcoma. *Cancer Med* 2023;12:5025-34.
 19. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;28:112-8.
 20. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE*; 2008.
 21. Zhu J, Pu S, He J, et al. Processing imbalanced medical data at the data level with assisted-reproduction data as an example. *BioData Min* 2024;17:29.
 22. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* 2015;48:2839-46.
 23. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. 2017. Available online: <https://doi.org/10.48550/arXiv.1705.07874>
 24. Survival Rates for Melanoma Skin Cancer. American Cancer Society 2023. Available online: [https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-](https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html)
 25. Fayne RA, Macedo FI, Rodgers SE, et al. Evolving management of positive regional lymph nodes in melanoma: Past, present and future directions. *Oncol Rev* 2019;13:433.
 26. Piris A, Mihm MC Jr, Duncan LM. AJCC melanoma staging update: impact on dermatopathology practice and patient management. *J Cutan Pathol* 2011;38:394-400.
 27. Ghebrial M, Wang Q, Zhang R, et al. Predictors of primary cutaneous melanoma stage at diagnosis: observations from Alberta's Tomorrow Project. *Ann Cancer Epidemiol* 2024;8:1.
 28. Rashed H, Flatman K, Bamford M, et al. Breslow density is a novel prognostic feature in cutaneous malignant melanoma. *Histopathology* 2017;70:264-72.
 29. Cavanaugh-Hussey MW, Mu EW, Kang S, et al. Older Age is Associated with a Higher Incidence of Melanoma Death but a Lower Incidence of Sentinel Lymph Node Metastasis in the SEER Databases (2003-2011). *Ann Surg Oncol* 2015;22:2120-6.
 30. Isaksson K, Nielsen K, Mikiver R, et al. Sentinel lymph node biopsy in patients with thin melanomas: Frequency and predictors of metastasis based on analysis of two large international cohorts. *J Surg Oncol* 2018;118:599-605.
 31. Conway WC, Faries MB, Nicholl MB, et al. Age-related lymphatic dysfunction in melanoma patients. *Ann Surg Oncol* 2009;16:1548-52.
 32. Wang TW, Nakanishi M. Immune surveillance of senescence: potential application to age-related diseases. *Trends Cell Biol* 2024;S0962-8924(24)00121-1.
 33. Nogalska A, Eerdeng J, Akre S, et al. Age-associated imbalance in immune cell regeneration varies across individuals and arises from a distinct subset of stem cells. *Cell Mol Immunol* 2024. [Epub ahead of print]. doi: 10.1038/s41423-024-01225-y.
 34. Lasithiotakis K, Leiter U, Meier F, et al. Age and gender are significant independent predictors of survival in primary cutaneous melanoma. *Cancer* 2008;112:1795-804.
 35. Chhabra Y, Fane ME, Pramod S, et al. Sex-dependent effects in the aged melanoma tumor microenvironment influence invasion and resistance to targeted therapy. *Cell* 2024;187:6016-6034.e25.
 36. Olsen CM, Pandeya N, Miranda-Filho A, et al. Does Sex Matter? Temporal Analyses of Melanoma Trends among Men and Women Suggest Etiologic Heterogeneity. *J Invest Dermatol* 2024;S0022-202X(24)01500-8.
 37. Olsen CM, Thompson JF, Pandeya N, et al. Evaluation

- of Sex-Specific Incidence of Melanoma. *JAMA Dermatol* 2020;156:553-60.
38. Joosse A, de Vries E, Eckel R, et al. Gender differences in melanoma survival: female patients have a decreased risk of metastasis. *J Invest Dermatol* 2011;131:719-26.
 39. Yalamanchi P, Brant JA, Chen J, et al. Clinicopathologic Factors Predictive of Occult Lymph Node Involvement in Cutaneous Head and Neck Melanoma. *Otolaryngol Head Neck Surg* 2018;158:489-96.
 40. Lachiewicz AM, Berwick M, Wiggins CL, et al. Survival differences between patients with scalp or neck melanoma and those with melanoma of other sites in the Surveillance, Epidemiology, and End Results (SEER) program. *Arch Dermatol* 2008;144:515-21.
 41. Saaiq M, Zalaudek I, Rao B, et al. A brief synopsis on scalp melanoma. *Dermatol Ther* 2020;33:e13795.
 42. Vița O, Jurescu A, Văduva A, et al. Invasive Cutaneous Melanoma: Evaluating the Prognostic Significance of Some Parameters Associated with Lymph Node Metastases. *Medicina (Kaunas)* 2023;59:1241.
 43. Voinea S, Sandru A, Gherghe M, et al. Peculiarities of lymphatic drainage in cutaneous malignant melanoma: clinical experience in 75 cases. *Chirurgia (Bucur)* 2014;109:26-33.
 44. Davies J, Muralidhar S, Randerson-Moor J, et al. Ulcerated melanoma: Systems biology evidence of inflammatory imbalance towards pro-tumourigenicity. *Pigment Cell Melanoma Res* 2022;35:252-67.
 45. In 't Hout FE, Haydu LE, Murali R, et al. Prognostic importance of the extent of ulceration in patients with clinically localized cutaneous melanoma. *Ann Surg* 2012;255:1165-70.

Cite this article as: Lan LF, Kai YL, Xu XL, Zhang JK, Xu GB, Dai YB, Shen Y, Lu HY, Wang B. Construction and validation of machine learning models for predicting lymph node metastasis in cutaneous malignant melanoma: a large population-based study. *Transl Cancer Res* 2025;14(2):706-716. doi: 10.21037/tcr-24-1672