Machine Learning Identification of Lymph Node Metastasis in women with early-stage epithelial ovarian cancer: a SEER population-based study

Xiu-jie Zhu
Second Affiliated Hospital of Wenzhou Medical Univ
wen-lai Fang
Second Affiliated Hospital of Wenzhou Medical Univ
chao-yi xu (yichaoxu1221@yeah.net)
Second Affiliated Hospital of Wenzhou Medical University https://orcid.org/0000-0003-1684-0007

Research

Keywords: Early-stage, Epithelial ovarian cancer, Lymph node metastasis, Machine learning, Surveillance, epidemiology, and end results

Posted Date: October 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1002546/v1

Abstract

Objective

Emerging evidence indicates that nearly 20% of patients with incompletely staged early-stage ovarian cancer develop to advanced stage because of lymph node metastases. The aim of the present study is to establish machine learning (ML) based predictive models for lymph node metastasis (LNM) in early-stage epithelial ovarian cancer (EOC).

Methods

The Surveillance, Epidemiology, and End Results (SEER) database was used to select patients diagnosed with early T classification epithelial ovarian cancer (EOC) between 2010 and 2015. The possibility of LNM was predicted by comparing the six ML algorithms. Model performance was compared in terms of accuracy, sensitivity, specificity, F1 score, and the area under the curve (AUC). The Shapley Additive Explanation (SHAP) analysis was employed to generate explanations and was presented as patient-specific visualizations.

Results

Screening of the SEER database yielded 3400 patients with early-stage epithelial ovarian cancer, and the data were divided randomly into a training set (70%), validation set (15%), and test set (15%). A grid search with 10-fold cross-validation was performed on the training set to tune the parameters. Overall, the Random Forest (RF) (accuracy of 79.8% and AUC of 0.878) was the best performing classifier and the Extreme Gradient Boosting (XGBoost) (accuracy of 77.2% and AUC of 0.857) demonstrated a similar high performance. The RF model performance was highly dependent on five top-rank features, including histology, grade, marital status, chemotherapy, and tumor size. SHAP analysis provided model-agnostic interpretation illustrating significant clinical contributors associated with risks of LNM in early-stage epithelial ovarian cancer (EOC).

Conclusions

The established ML-based prediction model for LNM in early-stage EOC is valid and valuable in improving clinical decision-making.

Background

Epithelial ovarian cancer (EOC) accounts for 90% of cancer of the ovary, fallopian tube, and peritoneum[1]. Emerging data demonstrate that nearly half of women diagnosed with ovarian cancer survive for more

than 5 years[2]. Of note, for women diagnosed with advanced stages of ovarian cancer, their 5-year survival rate was 20% compared to 89% for the early-stage disease[3]. The high mortality of ovarian cancer is because about 75% of patients are diagnosed with advanced EOC[4, 5]. Researches[5–7] also show that approximately 20-25% of patients with early-stage ovarian cancer develop stage IIIC ovarian cancer due to lymph node metastasis. These patients require timely assessment of the disease severity so that appropriate interventions are planned as early as possible to avoid reoperation and inappropriate adjuvant therapy. Therefore, exploring the risk factors may guide the prediction of the risk of LNM at early stages, which is of high clinical significance. The Surveillance, Epidemiology and End Results (SEER) program has data on cancer morbidity and mortality in the United States[8]. Methods based on machine learning (ML) are becoming more and more popular in the field of health domain with the advancement of machine learning technology[9]. In this study, we develop an ML-based model to predict the risk of LNM in early-stage EOC from the SEER program. In addition, we introduce the framework of Shapley additive explanation (SHAP)[10] values, which represent a unified approach to interpreting complex ML model predictions.

Material And Methods

Study population and variable selection.

Primary early stage EOC patients diagnosed between the years 2010 and 2015 were included using the SEER database. The records of patients diagnosed with EOC were retrieved from the SEER 18 registry database using the SEER*Stat 8.3.5 software. The specific inclusion criteria used to select patients with EOC were as follows: 1) Cases were limited stage T1 disease[11]; 2) the ICD-O-3 morphology codes ‘8020–8022, 8441–8442, 8460–8463, 9014’; ‘8470-8472, 8480-8481, 9015’; ‘8380-8383, 8570’ and ‘8290, 8310, 8313, 8443–8444’ were used to identify women with serous, mucinous, endometrioid and clear cell ovarian tumors, respectively ; 3) tumors located in primary site limited to C569. The exclusion criteria were as follows: 1)unknown grade, N stage and specific factor 1; 2) multiple primary cancers.Variables were grouped according to the actual clinical situation and previously reported cutoff values We included the following factors assessed at diagnose: age(<55 and ≥55 years old.), race (White, Black, other/unknown), marital status (married, unmarried), histology (serous, mucinous, endometrioid, clear cell), grade (wel-differentiated, moderately-differentiated, poorly differentiated, undifferentiated), laterality (unilateral, bilateral), tumor size (<5, 5–10, >10 cm), CA125 (normal, elevated),distant metastasis M0 and M1; regional lymph node N0 and N1 defined according to the American Joint Committee on Cancer seventh edition (AJCC7th), primary site surgery (yes, no), radiotherapy (yes, no/unknown) and chemotherapy (yes, no/unknown). Categorical variables were coded using one-hot-encoding, and continuous variables were transformed into z-scores. Missing data were imputed by using the non-parametric miss Forest method[12]. The ratio of positive to negative samples in the training set was 1 to 20, and this sample imbalance will have a great impact on the performance of the predictor. After data cleaning and feature processing, the Synthetic Minority Oversampling Technique (SMOTE) was used to tackle the data imbalance problem. SMOTE had been applied in machine learning applications in

healthcare, which can improve the performance of classifier[13]. The data were anonymous, and therefore the requirement for informed consent was waived from Institutional Review Board approval at Second Affiliated Hospital of Wenzhou Medical University.

Model construction

All patients were categorized randomly into a training (70%), validation (15%), and test set (15%). The correlation between each variable was measured by Pearson's correlation test, and results were presented in the heat map. Six ML algorithms, including logistic regression model (LR), support vector machine (SVM), multi-layer Perceptron Classifier (MLPClassifier), gaussian naive bayes (GaussianNB), Extreme Gradient Boosting (XGBoost) model, and random forest (RF) model were employed. We performed 10-fold cross-validation on the training set and the hyperparameters were tuned by grid search. The validation set was used to adjust for the model parameters, whereas the test set was used to evaluate the performance of the system. The final models were evaluated for the confusion matrix metrics of accuracy, sensitivity, specificity, F1 score, and area under the receiver operating characteristic (ROC) curve (AUC).

Model Interpretation

Feature contribution toward model prediction was evaluated by the Shapley Additive Explanations, a game-theory-based approach for elucidating feature importance for any fitted ML model. With the SHAP method, the importance of each predictor is sorted according to the SHAP value. The most important feature is that with the largest absolute SHAP value. Simultaneously, the characteristics of high SHAP values positively influence the output of the ML model and vice versa. The SHAP values were obtained in Python shap package to interpret model predictions and visualize the results. The present study refers to presentation videos and guidelines from the SHAP website (http://www.shap.ecs.soton.ac.uk/).

Statistical analysis

The SVM, LR, Gaussian NB, and RF models were implemented via the Python Sklearn package. The XGBoost was implemented using the Xgboost package. MLP Classifier was implemented using TensorFlow. All statistical analyses were performed in R (version 3.6.8, R Foundation for Statistical Computing) and Python (version 3.7, Python Software Foundation).

Results

Demographic and pathological characteristics

Of the 3400 patients with early-stage EOC enrolled in this study, 172 patients had lymph node metastasis and 3228 patients had no lymph node metastasis at primary diagnosis. All patients were completely randomized into the training set (n = 2380), the validation set (n = 510), and the test set (n = 510). Detailed clinicopathological variables are shown in Table 1. The Pearson Test was applied for correlation

analyses between the variables(Figure 1). Notably, the correlation heat map demonstrated mutually independent variables. Further, SMOTE was employed to solve the class-imbalanced problem. A 1:2 ratio of positive and negative training data sets was achieved after the multi-round testing, which effectively avoided the problem of imbalance.

Table 1
Demographic and clinicopathologic variables of the whole cohort grouped by lymph node status.

Variable	All (n=3400)	LN Metastasis (-) (n=3228)	LN Metastasis (+) (n=172)	*P* value
Age at diagnosis ,n(%)				<0.001
<50	1146(33.706)	1107(34.294)	39(22.674)	
50-60	1143(33.618)	1069(33.116)	74(43.023)	
>60	1111(32.676)	1052(32.590)	59(34.302)	
Race ,n(%)				0.149
White	2764(81.294)	2627(81.382)	137(79.651)	
Black	176(5.176)	163(5.050)	13(7.558)	
Other	460(13.529)	438(13.569)	22(12.791)	
Histology ,n(%)				<0.001
Serous	897(26.382)	786(24.349)	111(64.535)	
Mucinous	719(21.147)	712(22.057)	7(4.070)	
Endometrioid	1172(34.471)	1151(35.657)	21(12.209)	
Clear cell	612(18.000)	579(17.937)	33(19.186)	
Laterality ,n(%)				<0.001
Unilateral	3022(88.882)	2901(89.870)	121(70.349)	
Bilateral	378(11.118)	327(10.130)	51(29.651)	
M staging ,n(%)				<0.001
M0	3333(98.029)	3182(98.575)	151(87.791)	
M1	67(1.971)	46(1.425)	21(12.209)	
Marital status at diagnosis ,n(%)				<0.001
Unmarried	1489(43.794)	1414(43.804)	75(43.605)	
Married	1911(56.206)	1814(56.196)	97(56.395)	
Grade ,n(%)				<0.001
I	966(28.412)	950(29.430)	16(9.302)	
I I	997(29.324)	974(30.173)	23(13.372)	

Variable	All (n=3400)	LN Metastasis (-) (n=3228)	LN Metastasis (+) (n=172)	*P* value
I I I	868(25.529)	791(24.504)	77(44.767)	
IV	569(16.735)	513(15.892)	56(32.558)	
Cancer antigen 125 ,n(%)				<0.001
Normal	871(25.618)	851(26.363)	20(11.628)	
Elevated	2529(74.382)	2377(73.637)	152(88.372)	
Tumor size ,n(%)				<0.001
<5cm	607(17.853)	578(17.906)	29(16.860)	
5-10cm	716(21.059)	674(20.880)	42(24.419)	
>10cm	2077(61.088)	1976(61.214)	101(58.721)	
Surgery on primary site ,n(%)				<0.001
No	18(0.529)	13(0.403)	5(2.907)	
Yes	3382(99.471)	3215(99.597)	167(97.093)	
Radiotherapy ,n(%)				<0.001
No	3378(99.353)	3210(99.442)	168(97.674)	
Yes	22(0.647)	18(0.558)	4(2.326)	
Chemotherapy ,n(%)				<0.001
No	1394(41.000)	1361(42.162)	33(19.186)	
Yes	2006(59.000)	1867(57.838)	139(80.814)	

Model Performance

The AUCs of all the prediction models and the evaluation metrics of the confusion matrix are shown in Table 2, whereas the 10-fold cross-validation AUCs of all the prediction models are depicted in Table 3 showed. Figure 2 shows the ROC curves for the six models in the training set and testing set. The AUCs of the XGBoost (0.896) and RF (0.943) models were better than those of other models.. RF model performed better than the XGBoost model in the training set but both models demonstrated similar performance in the test set. Further analysis revealed that the RF had the best performance in the testing set with the following results: AUC: 0.878; accuracy: 0.772; sensitivity: 0.827; specificity: 0.745; F1 score: 0.710. The RF model also showed excellent performance in the 10-fold cross-validation (average AUC = 0.878).

Table 2
Metrics of the ML models.

Models	Dataset	AUC	Accuracy	Sensitivity	Specificity	F1 score
XGB	Train	0.896	0.814	0.831	0.806	0.749
	Test	0.857	0.772	0.827	0.745	0.710
RF	Train	0.943	0.855	0.899	0.833	0.805
	Test	0.878	0.798	0.838	0.780	0.738
Gaussian NB	Train	0.779	0.739	0.706	0.757	0.643
	Test	0.782	0.743	0.725	0.754	0.653
MLP Classifier	Train	0.776	0.734	0.748	0.729	0.653
	Test	0.775	0.733	0.751	0.729	0.648
SVM	Train	0.464	0.540	0.588	0.516	NaN
	Test	0.462	0.501	0.715	0.399	NaN
LR	Train	0.782	0.722	0.783	0.692	0.652
	Test	0.774	0.715	0.782	0.684	0.649

Table 3
The k-fold cross-validation accuracy (k = 10) of all three prediction models.

Model	XGB	RF	Gaussian NB	MLP Classifier	SVM	Logistic Regression
k-fold accuracy	0.876	0.878	0,772	0.776	0.535	0.785

Model Interpretation

The SHAP algorithm provides information regarding the impact of individual predicting factors on the directionality of the output of the model. In the present investigation, the SHAP values were obtained for each predictor by proposing RF and the balance set. The top five important clinicopathological factors included histology, grade marital status chemotherapy, and tumor size (Figure 3). For example, Figure 4 illustrates how the patient is assigned her predicted risk of LNM among women with early-stage epithelial ovarian cancer. For this patient, the RF model predicted her risk as 0.24 (base value: 0.48) and. the main factors driving her risk to high values include serous histology type and grade . Married status and non-receive chemotherapy reduced the predicted risk.

Discussion

The newly established prediction model in this retrospective cohort study was able to predict LNM in 3400 EOC patients with early T classification epithelial ovarian cancer based on multiple popular ML

algorithms. Data imbalance was solved using the SMOTE to decrease the influence of the class-imbalanced problem. In view of the results, the RF demonstrated better performance compared to classification algorithms. Simultaneously, the RF model was trained with the baseline clinicopathological factors, including age, race, histology, laterality, M staging, N staging, marital status, grade, cancer antigen 125, tumor size, surgery radiotherapy, and chemotherapy. The model output was evaluated using the SHAP method, which guided the understanding of the mechanism by which a single factor the model output.

Furthermore, all the variables in the present model were related to LNM in women with early-stage EOC. It was generally accepted that adjuvant chemotherapy was usually used for distant control. However, not all patients could benefit from adjuvant chemotherapy, especially as some patients got experience worse results after treatment[14–17]. The present investigation reported an association of adjuvant chemotherapy with increased risk of developing nodal metastasis in early-stage EOC. For patients with early-stage ovarian cancer confined to the ovary, surgical treatment alone could be curative, whereas adjuvant chemotherapy does not seem to confer any survival benefit[18, 19]. Elsewhere, patients with serous adenocarcinoma were found to exhibit a higher probability of developing lymph node metastasis compared to those with mucinous and other histological types of ovarian cancer[20, 21]. Similar results were confirmed in the present work; notably, 12.4% (111/897) of patients with serous histology had positive lymph nodes compared to only 2.4% (61/2503) among those with the non-serous disease. Additionally, the incidence of lymph node positivity was 1.7% among patients with low-grade tumors. Suzuki et al[22]. found that mucinous tumors and low-grade stage tumors were rarely possible to have nodal dissemination based on risk factors. Consistent results were reported in the present study in which the histological types and grade were considered to be the most significant predictors in the established model. Previous studies[23, 24] supported our findings that larger tumors are related to lymph nodes metastasis due to radioresistance caused by the biological aggressiveness of cancer clones, an inadequate blood supply, tumor hypoxia, and other adverse factors. In the present analysis, unmarried patients were found to exhibit a significantly higher risk of lymph node metastasis. This observation may be ascribed to the fact that married patients are characterized by less pain and anxiety compared to unmarried patients because partners can share emotional burdens and provide appropriate social support[25, 26].

These observations were considered in constructing an RF model with optimized hyperparameters in the SEER dataset. The RF model is a supervised, non-parametric method of classification on the basis of the random forest algorithm and the random forest packets of thousands of classification trees or regression trees[27]. Leveraging SHAP-based analysis, the present study generated a visualization format that allowed for the interpretation of patient-associated risks based on clinical variables. By highlighting significant clinical variables that contribute to risk predictions, such visualization can guide practitioners in the preemptive and early identification of key factors.

There are several shortcomings to the present study. First, the retrospective nature of the study implies unavoidable selection bias. Second, several variables were not accounted for in the models, including targeted endocrine, length of surgery, specific chemotherapy plans, smoking status and lifetime ovulation, which potentially could affect the model performance. Third, the established model warrants further validation in a larger external queue and optimization to ascertain its predictive performance in a specific population.

Conclusion

In conclusion, the established ML model is an RF-based predictor for lymph nodes metastasis in patients with early-stage EOC. The ML-based model holds great promise for accurate prediction of LNM in early-stage EOC, therefore, it will provide an easy, accurate, inexpensive, and time-effective method in clinical practice.

Abbreviations

LNM: lymph node metastasis; EOC: epithelial ovarian cancer; AJCC7th: American Joint Committee on Cancer seventh edition (AJCC7th); SMOTE:Synthetic Minority Oversampling Technique; AUC:area under the curve; LR: logistic regression model; SVM: support vector machine; MLPClassifier: multi-layer Perceptron Classifier; GaussianNB: gaussian naive bayes; XGBoost: Extreme Gradient Boosting; ML: machine learning; SHAP: Shapley additive explanation; SEER: Surveillance epidemiology and end Results

Declarations

Acknowledgements

This work is supported by Extreme Smart Analysis platform (https://www.xsmartanalysis.com/)

Ethics approval and consent to participate

All procedures in studies involving human participants were performed in accordance with the ethical standards of the institutional review board of Second Affiliated Hospital of Wenzhou Medical University Ethics Committee basing on the 1964 Helsinki declaration and its later amendments.

Authors' contribution

WL-F and XJ-Z provided contribution to (i) data analysis and interpretation and (iii)manuscript drafting and critical revising. CY-X contributed to (i) data analysis and interpretation and (ii) critical manuscript revising for important intellectual content. All authors have approved the final version and submission of this manuscript.

Funding

This research received no external funding.

Availability of data and materials

The dataset generated and analyzed during the current study is available in the Surveillance Epidemiology and End Results (SEER) Database repository [https://seer.cancer.gov/].

Consent for publication

Not applicable.

Competing interests

The authors declare no conflicts of interest in preparing this article

Author Details

1: Department of Obstetrics and Gynecology, Second Affiliated Hospital of Wenzhou Medical University,109 Xueyuan Xi Road, Wenzhou 325027, Zhejiang, China.

2: Department of Orthopedics, Second Affiliated Hospital of Wenzhou Medical University,109 Xueyuan Xi Road, Wenzhou 325027, Zhejiang, China.

Correspondence to: Dr. Yi-chao Xu, Department of Obstetrics and Gynecology, Second Affiliated Hospital of Wenzhou Medical University,109 Xueyuan Xi Road, Wenzhou 325027, Zhejiang, China.Mail: yichaoxu1221@yeah.net

References

1. Suh DH, Park J-Y, Lee J-Y, Kim B-G, Lim MC, Kim J-W, et al. The clinical value of surgeons' efforts of preventing intraoperative tumor rupture in stage I clear cell carcinoma of the ovary: a Korean multicenter study. Gynecol Oncol. 2015;137(3):412–7.
2. Brantnerova R, Manchanda R, Colombo N. The European Society of Gynaecological Oncology: Update on Objectives and Educational and Research Activities. American Society of Clinical Oncology Educational Book. 2012;32(1):335–8.
3. Survival rates for ovarian cancer,by stage. American Cancer Society. 21 March 2013; http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-survival-rates.
4. Drakes ML, Mehrotra S, Aldulescu M, Potkul RK, Liu Y, Grisoli A, et al. Stratification of ovarian tumor pathology by expression of programmed cell death-1 (PD-1) and PD-ligand-1 (PD-L1) in ovarian cancer. Journal of ovarian research. 2018;11(1):1–11.
5. Kleppe M, Wang T, Van Gorp T, Slangen B, Kruse AJ, Kruitwagen R. Lymph node metastasis in stages I and II ovarian cancer: a review. Gynecol Oncol. 2011;123(3):610–4.

6. Harter P, Gnauert K, Hils R, Lehmann T, Fisseler-Eckhoff A, Traut A, et al. Pattern and clinical predictors of lymph node metastases in epithelial ovarian cancer. International Journal of Gynecologic Cancer. 2007;17(6).
7. Smits A, Bryant A, Lopes AD, Galaal K. Lymph node dissection (lymphadenectomy) for presumed early stage epithelial ovarian cancer. The Cochrane Database of Systematic Reviews. 2015;2015(1).
8. Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME. The surveillance, epidemiology and end results (SEER) program and pathology: towards strengthening the critical relationship. Am J Surg Pathol. 2016;40(12):e94.
9. Gore JC. Artificial intelligence in medical imaging. Elsevier; 2020.
10. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems; 2017.
11. Kremer K, Carlson M, Lee J, Lococo S, Miller D, Lea J. Prognostic factors associated with overall survival in presumed early stage, high-grade serous ovarian cancer: an analysis of the SEER cancer database. Gynecol Oncol. 2021;162:246.
12. Xue B, Li D, Lu C, King CR, Wildes T, Avidan MS, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. JAMA network open. 2021;4(3):e212240-e.
13. Hsieh M-H, Sun L-M, Lin C-L, Hsieh M-J, Hsu CY, Kao C-H. The performance of different artificial intelligence models in predicting breast cancer among individuals having type 2 diabetes mellitus. Cancers. 2019;11(11):1751.
14. Liu Z, Meng X, Zhang H, Li Z, Liu J, Sun K, et al. Predicting distant metastasis and chemotherapy benefit in locally advanced rectal cancer. Nature communications. 2020;11(1):1–11.
15. Glimelius B, Tiret E, Cervantes A, Arnold D. Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2013;24:vi81-vi8.
16. Ledermann J, Raja F, Fotopoulou C, Gonzalez-Martin A, Colombo N, Sessa C. Newly diagnosed and relapsed epithelial ovarian carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2013;24:vi24–32.
17. Maas M, Nelemans PJ, Valentini V, Crane CH, Capirci C, Rödel C, et al. Adjuvant chemotherapy in rectal cancer: defining subgroups who may benefit after neoadjuvant chemoradiation and resection: a pooled analysis of 3,313 patients. International journal of cancer. 2015;137(1):212–20.
18. Collaborators ICON. International Collaborative Ovarian Neoplasm trial 1: a randomized trial of adjuvant chemotherapy in women with early-stage ovarian cancer. J Natl Cancer Inst. 2003;95(2):125–32.
19. Winter-Roach BA, Kitchener HC, Lawrie TA. Adjuvant (post-surgery) chemotherapy for early stage epithelial ovarian cancer. Cochrane Database of Systematic Reviews. 2012(3).
20. Onda T, Yoshikawa H, Yokota H, Yasugi T, Taketani Y. Assessment of metastases to aortic and pelvic lymph nodes in epithelial ovarian carcinoma: a proposal for essential sites for lymph node biopsy. Cancer: Interdisciplinary International Journal of the American Cancer Society. 1996;78(4):803–8.

21. Lang J. Lymph node metastasis in stage I ovarian carcinoma. Chin Med J. 1994;107(9):643–7.
22. Suzuki M, Ohwada M, Yamada T, Kohno T, Sekiguchi I, Sato I. Lymph node metastasis in stage I epithelial ovarian cancer. Gynecol Oncol. 2000;79(2):305–8.
23. Lartigau E, Ridant AML, Lambin P, Weeger P, Martin L, Sigal R, et al. Oxygenation of head and neck tumors. Cancer. 1993;71(7):2319–25.
24. Janssen H, Haustermans K, Balm A, Begg A. Hypoxia in head and neck cancer: how much, how important? Head & Neck. Journal for the Sciences Specialties of the Head Neck. 2005;27(7):622–38.
25. Goldzweig G, Andritsch E, Hubert A, Brenner B, Walach N, Perry S, et al. Psychological distress among male patients and male spouses: what do oncologists need to know? Annals of oncology. 2010;21(4):877–83.
26. Aizer AA, Chen M-H, McCarthy EP, Mendu ML, Koo S, Wilhite TJ, et al. Marital status and survival in patients with cancer. Journal of clinical oncology. 2013;31(31):3869.
27. Wang J, Zhang T, Yang L, Yang G. Comprehensive genomic analysis of microenvironment phenotypes in ovarian cancer. PeerJ. 2020;8:e10255.

Figures

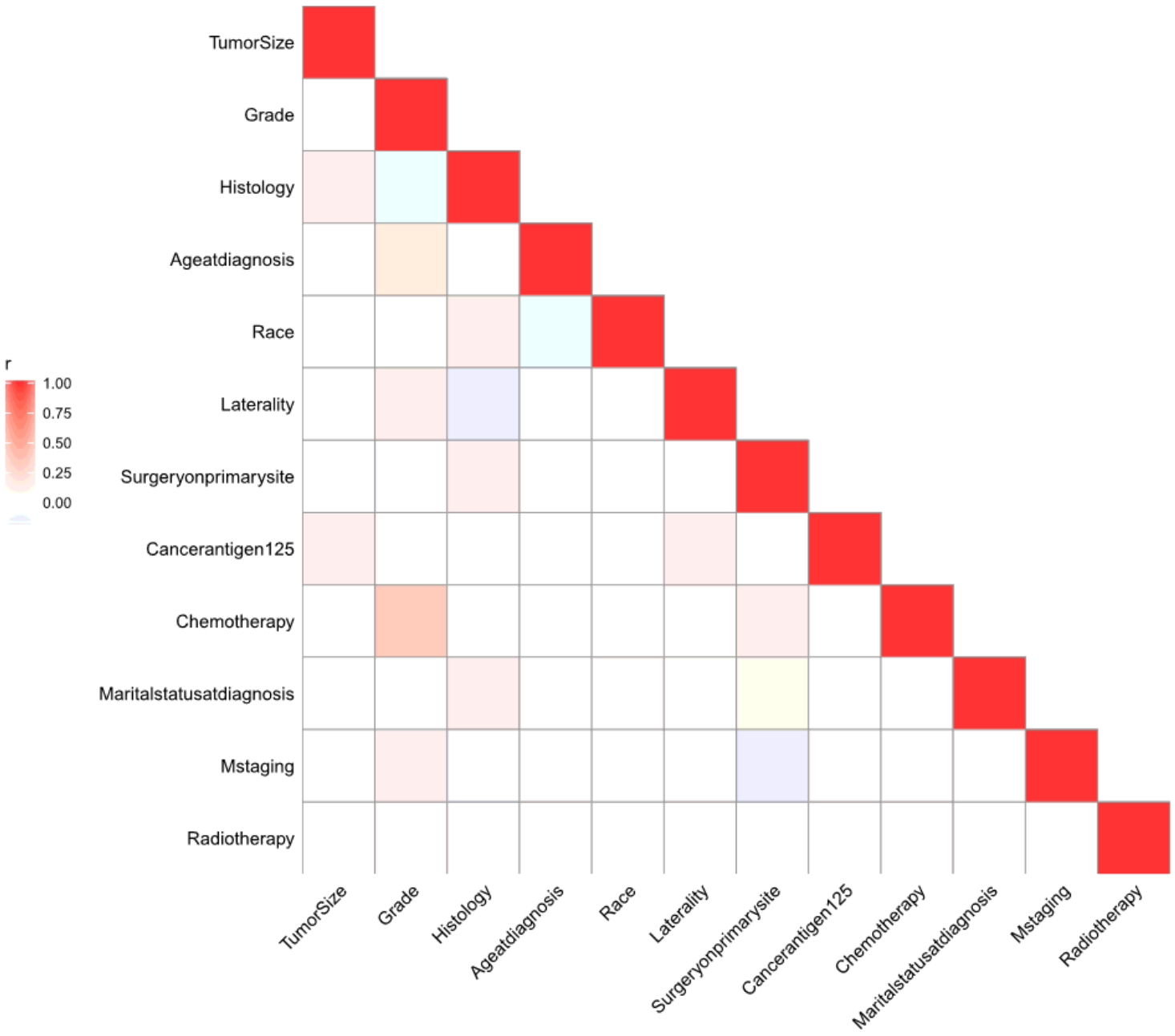

Figure 1

Correlation between factors

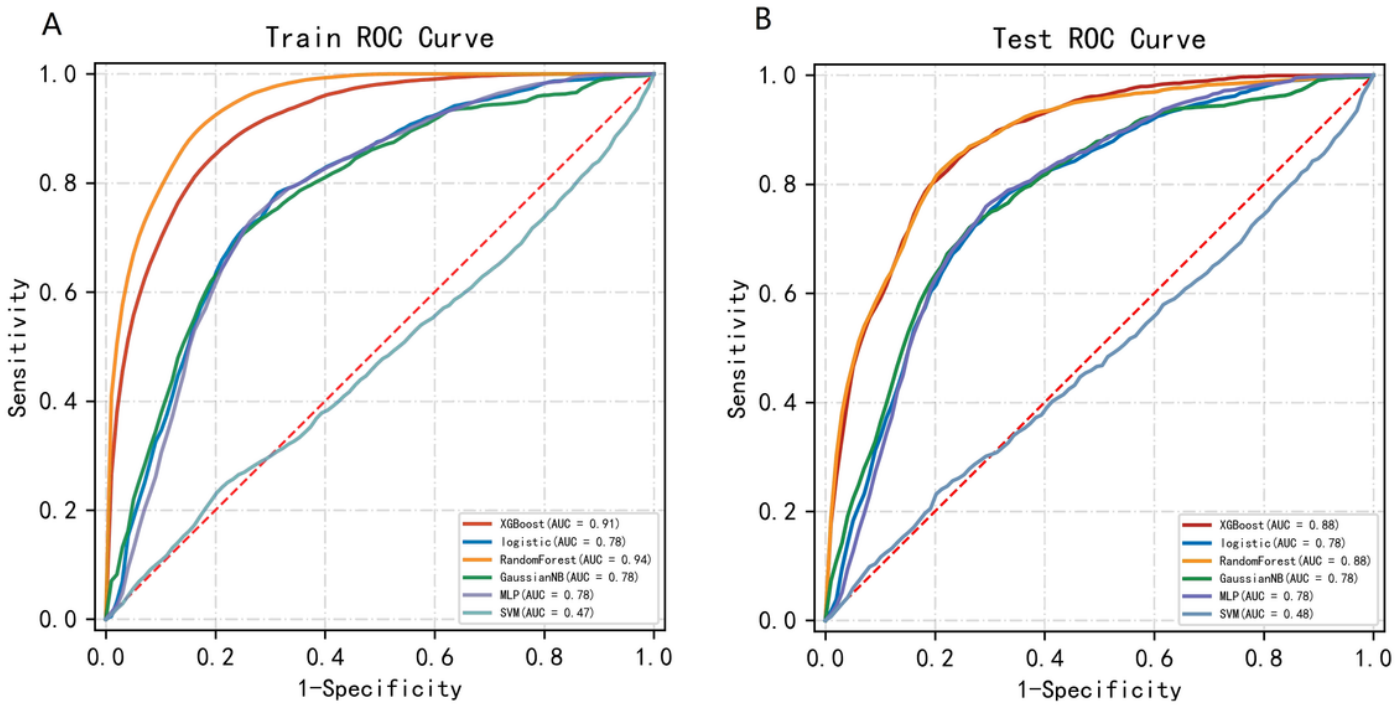

Figure 2

Different machine learning algorithms predict the LNM in the training and testing group.

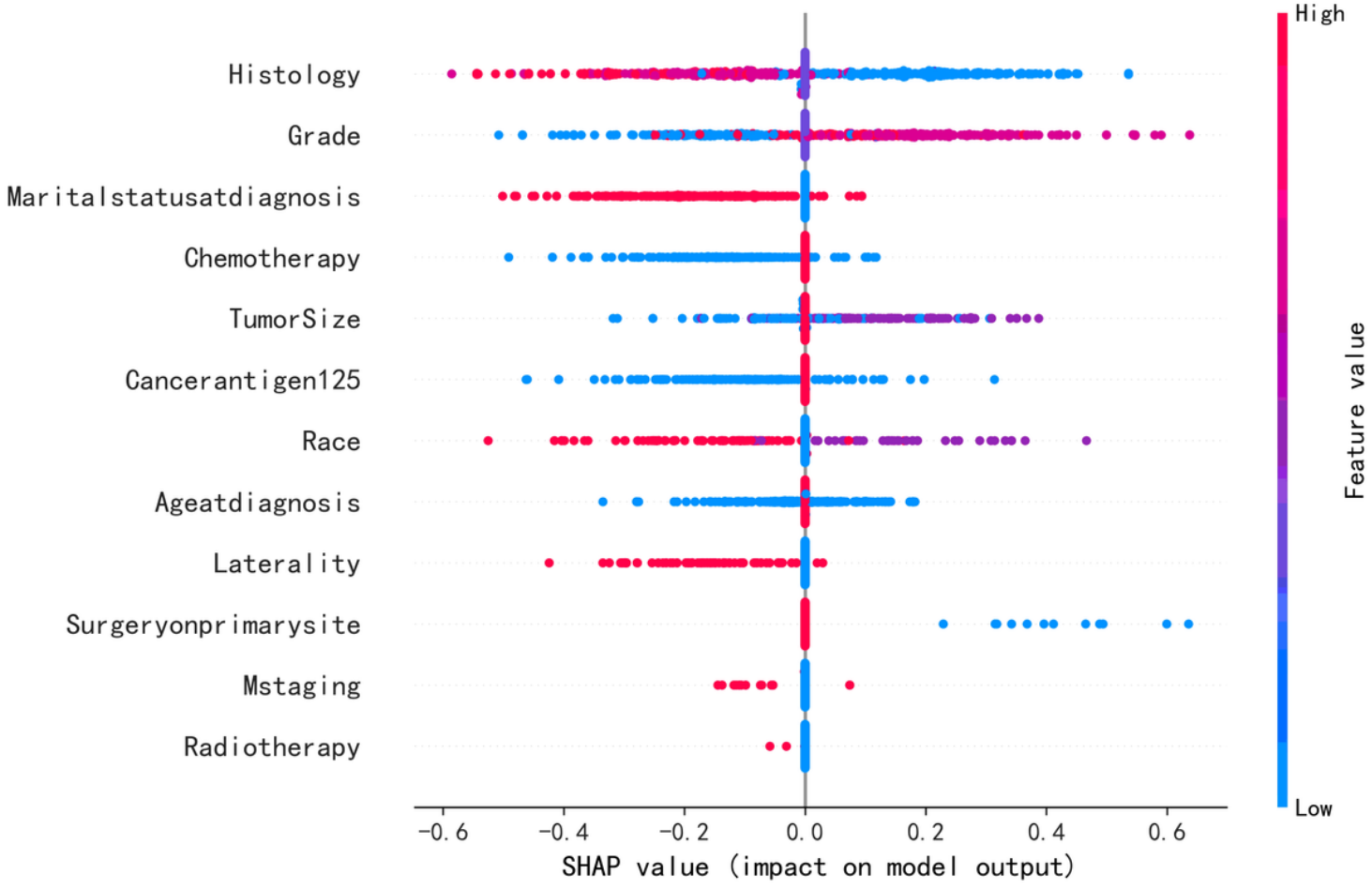

Figure 3

Summary plots for SHAP values. For each feature, one point corresponds to a single patient. A point's position along the x axis (i.e., the actual SHAP value) represents the impact that feature had on the model's output for that specific patient. Mathematically, this corresponds to the (logarithm of the) mortality risk relative across patients (i.e., a patient with a higher SHAP value has a higher mortality risk relative to a patient with a lower SHAP value). 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.

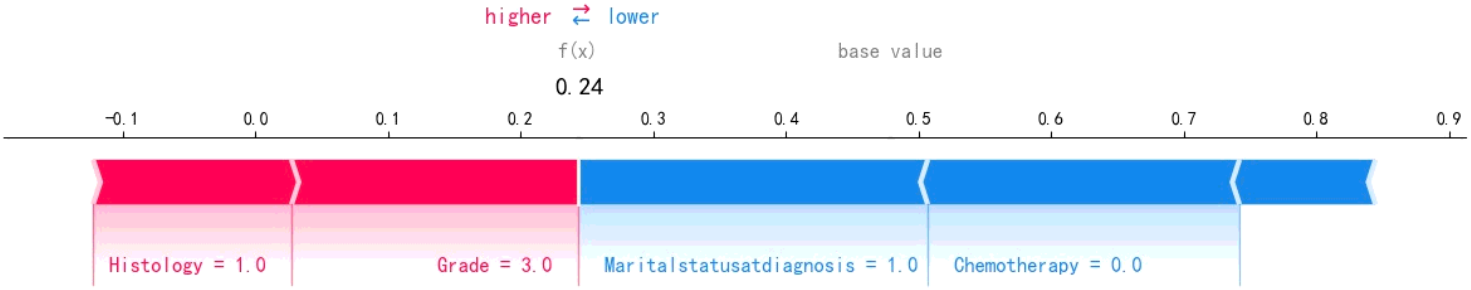

Figure 4

Explained risk for individual. The contributing variables are arranged in the horizontal line, sorted by the absolute value of their impact. The output value is the predicted risk of lymph node metastasis. The base value means the expected value of model C, over the training dataset.