Article

Not peer-reviewed version

Machine Learning Algorithm for Predicting Distant Metastasis of T1 and T2 Gallbladder Cancer Based on Seer Database

Zhentian Guo , Zongming Zhang * , Limin Liu , Yue Zhao , Zhuo Liu , Chong Zhang , Hui Qi , Jinqiu Feng , Peijie Yao , Haiming Yuan

Posted Date: 22 July 2024

doi: 10.20944/preprints202407.1736.v1

Keywords: machine learning; SEER; gallbladder cancer; distant metastasis

Article

Machine Learning Algorithm for Predicting Distant Metastasis of T1 and T2 Gallbladder Cancer Based on SEER Database

Zhentian Guo [1,2], Zongming Zhang [1,2,*], Limin Liu [1,2], Yue Zhao [1,2], Zhuo Liu [1,2], Chong Zhang [1,2], Hui Qi [1,2], Jinqiu Feng [2], Peijie Yao [2] and Haiming Yuan [1]

[1] Department of General Surgery, Beijing Electric Power Hospital, State Grid Corporation of China, Capital Medical University, Beijing 100073, China; zhentian990305@163.com (Z.G.); zhangzongming @mail.tsinghua. edu.cn (Z.Z.); liulimin15@sina.com (L.L); zhaoyue806@sina.com (Y.Z.); lz0002@163.com (Z.L.); zhangchong2118@163.com (C.Z.); 407021026@qq.com (H.Q.); 18811777978@163.com (J.F.); ypj5209@163.com (P.Y.); haimingyuan@aliyun.com (H.Y.)

[2] Key Laboratory of Geriatrics (Hepatobiliary Diseases) of China General Technology Group, Beijing 100073, China

* Correspondence: zhangzongming@mail.tsinghua.edu.cn

Abstract: (1) Background: We aim to construct a machine learning (ML) algorithm to predict the risk of distant metastasis (DM) of T1 and T2gallbladder cancer (GBC); (2) Demographic and clinical pathological data of T1 and T2 GBC patients were extracted from the National Institutes of Health (NIH)'s Surveillance, Epidemiology, and End Results (SEER) database between 2004 and 2015 to develop seven ML algorithm models. Models were evaluated based on accuracy, precision, recall rate, F1- score, and area under the receiver operating characteristic curve (AUC); (3) Results:A total of 4371 patients were included in the study, of whom 764 (17.4%) developed DM. Multivariate logistic regression showed that age, histology, tumor size, T and N stages were independent factors in GBC with DM. A novel nomogram was established to predict distant metastasis in early T stage GBC patients. Evaluation indicators of the best model Random Forest (RF) were as follows: accuracy (0.828), recall rate (0.862), precision (0.811), F1- score (0.836), and AUC value (0.913); (4) Conclusions: The RF model constructed in this study could accurately predict distant metastasis in GBC patients, which may provide clinicians with more personalized clinical decision-making recommendations.

Keywords: machine learning; SEER; gallbladder cancer; distant metastasis

1. Introduction

Gallbladder cancer (GBC), as a common malignant tumor in the biliary system, has the characteristics of concealed onset, rapid progress, early metastasis, and poor prognosis. Its incidence rate is closely related to gallstones and chronic cholecystitis[1, 2]. Due to the high malignancy and lack of specific symptoms and signs in the early stages of gallbladder cancer, distant metastasis often occurs when the disease is detected. The 5-year survival rates of GBC patients in T3 and T4 stages are 32.4% and 3.5%, respectively[3,4]. At present, there is still a lack of early diagnostic methods with good specificity and sensitivity for gallbladder cancer, and most of the clinically discovered GBC are in the middle and late stages[5]. Studies have shown that the incidence of lymph node and distant metastasis in GBC patients ranges from 17.9% to 64.5%, and the most common metastatic organs are the liver, lungs, and peritoneum[6–8].Among GBC patients, the prognosis of patients with distant metastasis is worse than those without distant metastasis, and the one-year survival rate of GBC patients with distant metastasis is 20% -50%[7,9]. Research has shown that distnt metastasis is an important predictive factor for the survival of GBC patients[10]. Early assessment of the risk of distant metastasis is crucial for early intervention and improving the prognosis of GBC patients in T1 and T2

 doi:10.20944/preprints202407.1736.v1

stages of gallbladder cancer.Although Nomogram is currently the most commonly used clinical prediction model, machine learning algorithms are increasingly being applied to construct clinical models for their practicality, innovation, and accuracy[11]. Machine learning algorithms have broad prospects in utilizing complex and massive clinical data for disease diagnosis and outcome prediction. Previous studies have shown that machine learning has more advantages than traditional big data clinical prediction research methods[12].

Therefore, this study aims to establish a machine learning prediction model to predict the occurrence of distant metastasis in GBC patients. This study can provide clinicians with more personalized clinical decisions, improve patient prognosis through early intervention, and effectively enhance patient quality of life.

2. Materials and Methods

2.1. Data Sources and Study Population

Data for this study were acquired from the SEER public databases, utilizing SEER*stat 8.4.2 software for data extraction. Our study focused on patients diagnosed with GBC in the United States between 2004 and 2015. and we chose patients using the procedure depicted in Figure 1.The criteria for including data in this study include: 1) The 6 edition of the AJCC TNM staging system was used as the basis for staging the cases included in the study; 2) Clear histological diagnosis; 3) For a single tumor.

Exclusion criteria include 1) missing or incomplete data, including T staging, M staging, etc. Variables included age, sex (male or female), race (White, Black, and others), year of diagnosis, Hispanic, histology (adenocarcinoma and others), tumor size, marital status, T stage, N stage, grade, and DM.Distant metastasis means that the tumor invades at least one or more target organs such as the liver, lung,peritoneum, and so on.As the SEER database contains public data, informed consent from relevant patients for using the SEER database for research purposes was not required, nor was ethical approval. The National Cancer Institute, USA (reference number 19238-Nov2021) approved our request for access to the SEER data.

2.2. Screening for Risk Factors and Model Construction

Statistical analysis was conducted using SPSS software (version 26.0; IBM Corporation). Construct a nomogram prediction model for DM using R 4.3.2 and draw a calibration curve. All patients were randomly divided into training set and test set at 8:2. The categorical variable was expressed in numbers and percentages, and the Chi-squared test, Fisher's exact test, and Mann-Whitney U test were used for inter-group comparison.We establish a logistic regression model based on the results of univariate and multivariate logistic regression analysis and display them in the form of a nomogram. A nomogram is a graphical representation that converts mathematical formulas into geometric expressions and explains the interactions between predicted variables, Mainly used in logistic regression models and COX proportional risk models[13]. The receiver operating characteristic (ROC) curve was plotted and analyzed based on the results. An area under the ROC curve (AUC) greater than 0.5 was considered meaningful. All computed p values were two-sided, and statistical significance was accepted at <0.05.

Use Python software(version 3.9.12, Python Software Foundation).Include all variables in the ML model, and a prediction model is built. In the SEER database, there are fewer cases of distant metastasis in T1 and T2 gallbladder cancer patients, the original dataset is imbalanced. We use under-sampling and oversampling techniques to process the raw data and use correlation matrices to analyze the changes in the sampled data. The technically processed data (oversampled and undersampled data) were randomly divided into a training set (80%) and a test set (20%). After sampling, the correlation between variables becomes clearer, as shown in Figure 2. The training set uses seven common machine learning algorithms,including random forest (RF), decision Tree (DT) support vector machine (SVM), naive Bayes (NB), k nearest neighbor (KNN), eXtreme gradient boosting (XGBoost), and gradient boosting machine (GBM).Model evaluation is mainly based on

 doi:10.20944/preprints202407.1736.v1

accuracy, precision, recall, F1 score, and AUC value, and the model with the highest ROC value and F1 score is the optimal model.

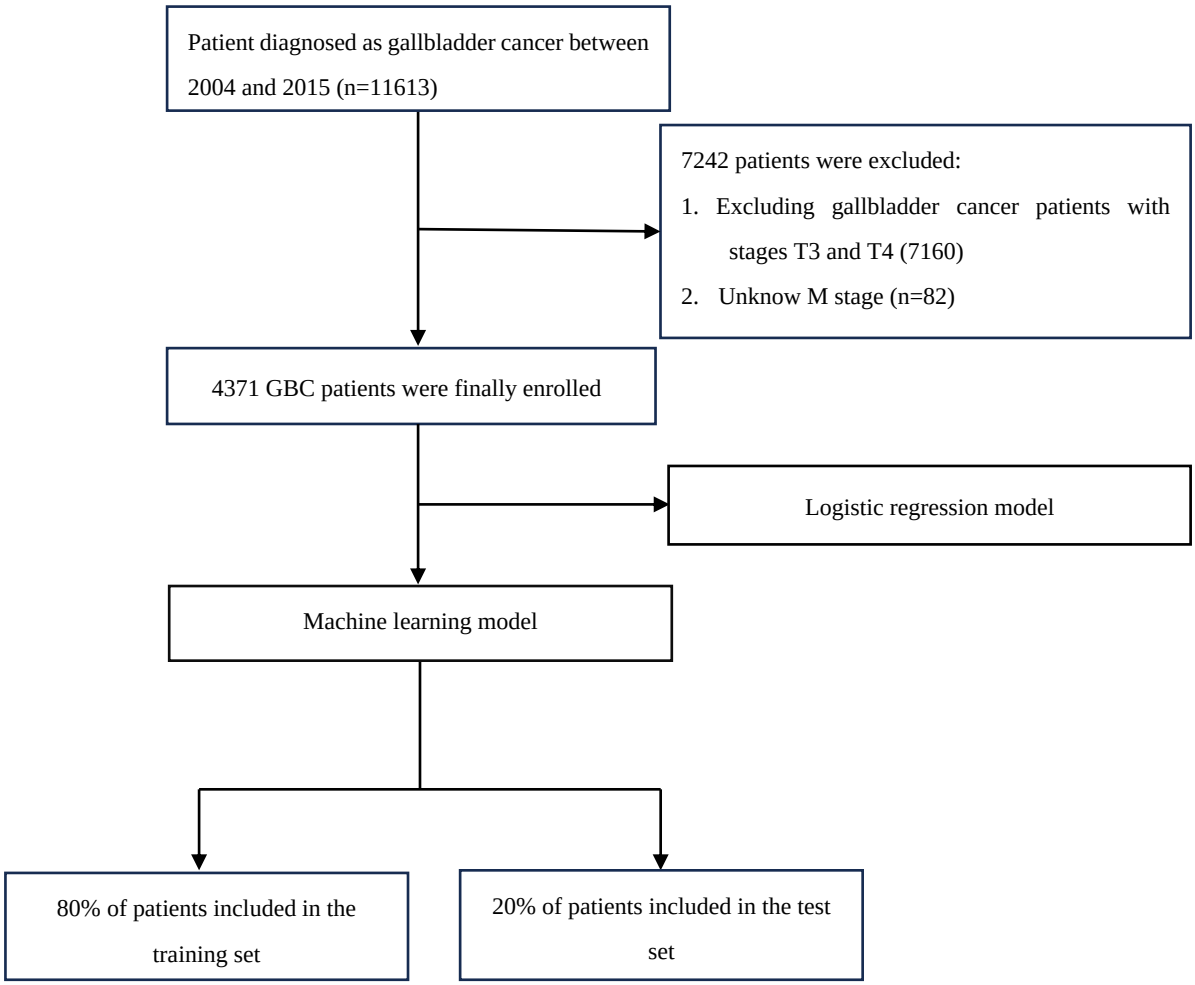
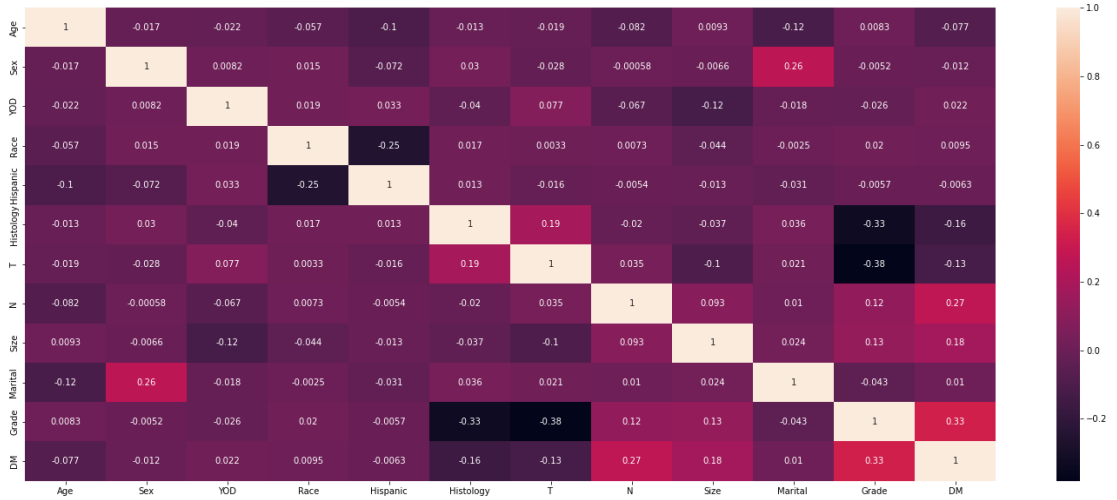


Figure 1. The flow diagram of the selection process for the study.



(**a**)

 doi:10.20944/preprints202407.1736.v1

(**b**)

Figure 2. Correlation heatmaps of patient characteristics feature in different datasets (a):Over-sampling data. (b): Under-sampling data.

3. Results

3.1. Analysis of Patient Information

This study included a total of 4371 patients diagnosed with T1 and T2 gallbladder cancer, Among them, 764 patients had distant metastasis, while the other 3607 patients did not have distant metastasis. The majority of patients in this study were elderly (≥70 years old, 56.9%), female (70.3%), and white (76.5%). There were significant differences in age, histology, tumor size, T stage, N stage, and grade among patients with DM ($p<0.05$), and there were no significant differences in other data. The baseline data characteristics and survival data of all patients are shown in Table 1.

In this study, we used univariate and multivariate logistic regression to screen for clinical factors that affect distant metastasis. Age, history, tumor size, T stage, N stage, and grade are all risk factors for distant metastasis in T1 and T2 gallbladder cancer patients in univariate and multivariate logistic regression (Table 2). Based on the results of multivariate LR analysis, an LR model was constructed with AUC=0.755 (95%: 0.734-0.776) in the test set and AUC=0.738 (95%: 0.693-0.783) in the training set (Figure 3). Figure 4 shows the calibration curves of the model in both the test and training sets. The calibration curves show that the predicted probability curve is roughly similar to the predicted actual value, indicating that the predicted model is consistent with the actual model and has good calibration readability. Figure 5A is the nomogram of GBC distant metastasis, which clearly shows the impact of each risk factor on the outcome variable. From the DCA of the distant metastasis nomogram (Figure 5B), it can be seen that within the threshold probability range of 1% -40%, the net benefit (NB) of the model's decision curve is higher than the net benefit of the two invalid lines.

Table 1. Demographics and clinical characteristics of the gallbladder cancerpatients in T1 and T2.

Characteristic	Without DM (N=3607)	With DM (N=764)	p-value
Age(year)			<0.001
<70	1508 (41.8%)	374 (49.0%)	
≥70	2099 (58.2%)	390 (51.0%)	
Gender			0.181
Female	2523 (69.9%)	553 (72.4%)	
Male	1084 (30.1%)	211 (27.6%)	
Race			0.599
white	2770 (76.8%)	578 (75.7%)	

 doi:10.20944/preprints202407.1736.v1

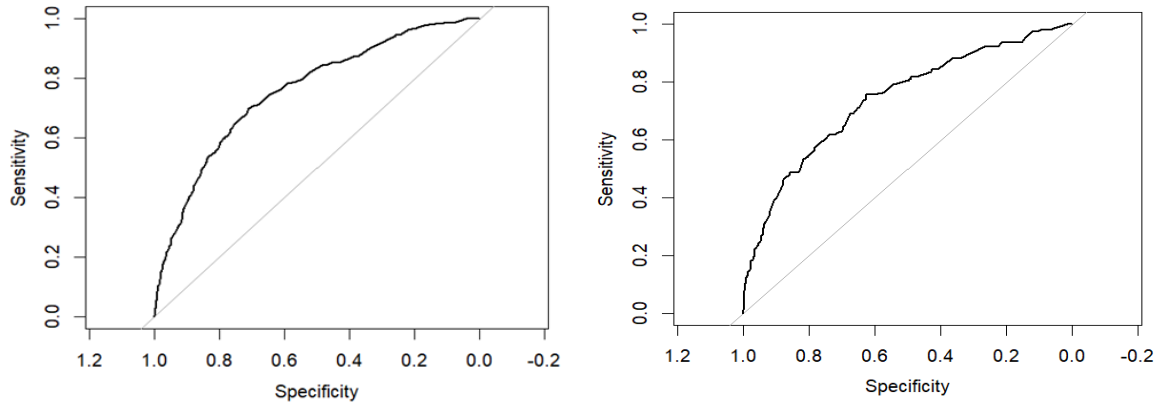
black	400 (11.1%)	97 (12.7%)	
other	437 (12.1%)	89 (11.6%)	
Hispanic			0.572
YES	808 (22.4%)	164 (21.5%)	
NO	2799 (77.6)	600 (78.5%)	
Histology			<0.001
Adenocarcinom	3308 (91.7%)	611 (80.0%)	
Others	299 (8.3%)	153 (20.0%)	
Year of diagnosis			0.262
2004-2009	1624 (45.0%)	327 (42.8%)	
2010-2015	1983 (55.0%)	437 (57.2%)	
Tumor size(cm)			<0.001
<2	2270 (76.8%)	578 (75.7%)	
≥2	400 (11.1%)	97 (12.7%)	
Unknown	437 (12.1%)	89 (11.6%)	
T stage			<0.001
T1	1259 (34.9%)	361 (47.3%)	
T2	2348 (65.1%)	403 (52.7%)	
N stage			<0.001
N0	2871 (79.6%)	422 (55.2%)	
N1	644 (17.8%)	257 (33.7%)	
NX	92 (2.6%)	85 (11.1%)	
Marital status			0.531
Single	1839 (51.0%)	380 (49.7%)	
Married	1768 (49.0%)	384 (50.3%)	
Grade			<0.001.
Grade I	737 (20.4%)	39 (5.1%)	
Grade II	1536 (42.6%)	219 (28.6%)	
Grade III	894 (24.8%)	255 (33.4%)	
Grade IV	55 (1.5%)	18 (2.4%)	
Unknown	385 (10.7)	233 (30.5%)	

Table 2. Univariate and multivariate analysis in the training cohort.

	Univariable analysis			**Multivariable analysis**		
	OR	95%CI	P value	OR	95%CI	P value
Age(year)						
<70	Ref			Ref		
≥70	0.723	0.607-0.861	<0.001	0.705	0.583-0.852	<0.001
Gender						
Female	Ref					
male	0.881	0.726-1.069	0.200			
Race						
white	Ref					
black	1.116	0.850-1.464	0.431			
other	0.980	0.746-1.287	0.885			
Hispanic						
YES	0.997	0.810-1.228	0.977			
NO	Ref					
Histology						
Adenocarcinom	0.345	0.274-0.436	<0.001	0.595	0.456-0.777	<0.001
Others	Ref			Ref		

 doi:10.20944/preprints202407.1736.v1

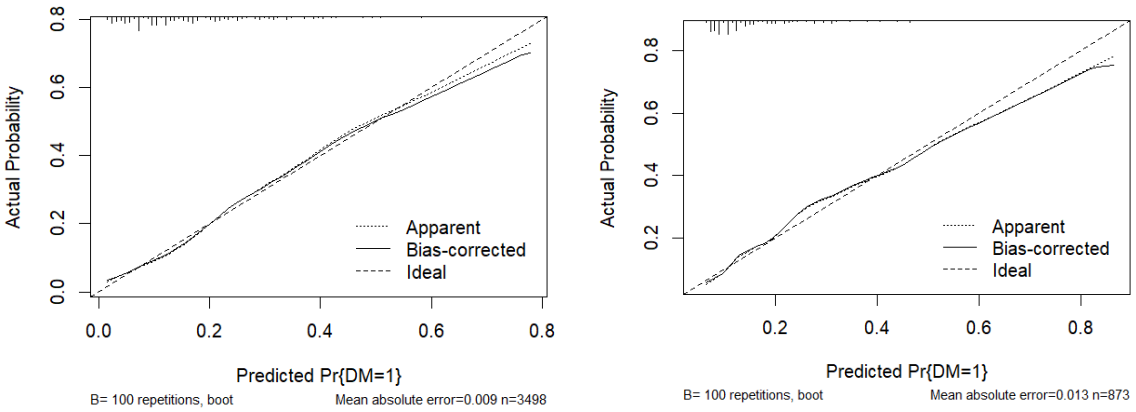
Year of diagnosis						
2004-2009	Ref					
2010-2015	1.151	0.965-1.374	0.117			
Tumor size(cm)						
<2	Ref			Ref		
≥2	1.916	1.449-2.534	<0.001	1.507	1.121-2.027	0.007
Unknown	2.729	2.067-3.602	<0.001	2.023	1.509-2.714	<0.001
T stage						
T1	Ref					
T2	0.594	0.498-0.708	<0.001	0.679	0.547-0.843	<0.001
N stage						
N0	Ref			Ref		
N1	2.656	2.155-3.197	<0.000	2.377	1.920-2.944	<0.001
NX	6.067	4.299-8.563	<0.000	4.913	3.398-7.105	<0.001
Marital status						
Single	Ref					
Married	1.096	0.920-1.305	0.304			
Grade						
Grade I	Ref			Ref		
Grade II	3.507	2.281-5.391	<0.001	3.236	2.090-5.010	<0.001
Grade III	6.835	4.453-10.489	<0.001	5.776	3.721-8.966	<0.001
Grade IV	8.990	4.316-18.725	<0.001	6.316	2.932-13.605	<0.001
Unknown	13.936	8.977-21.635	<0.001	8.684	5.475-13.774	<0.001



(**a**) (**b**)

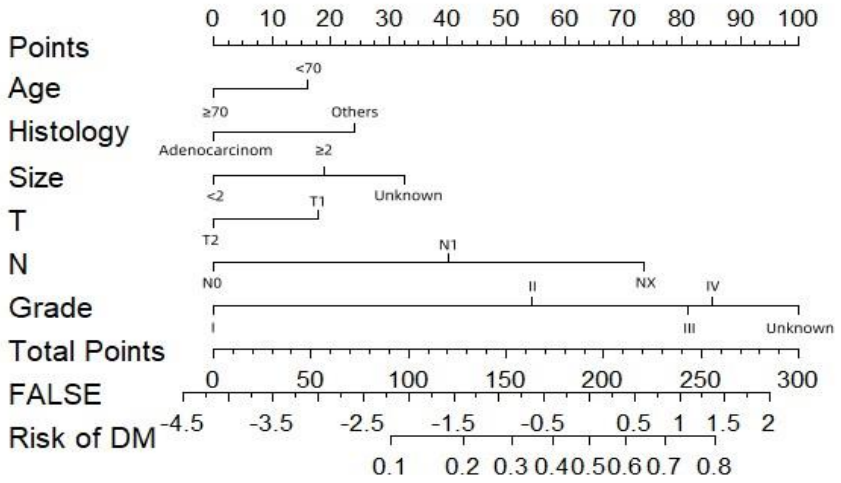
Figure 3. Prediction of ROC curves for distant metastasis in GBC patients using LR models in test set and training set. (a):ROC curve of LR model in test set. (b): ROC curve of LR model in training set.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2024 doi:10.20944/preprints202407.1736.v1



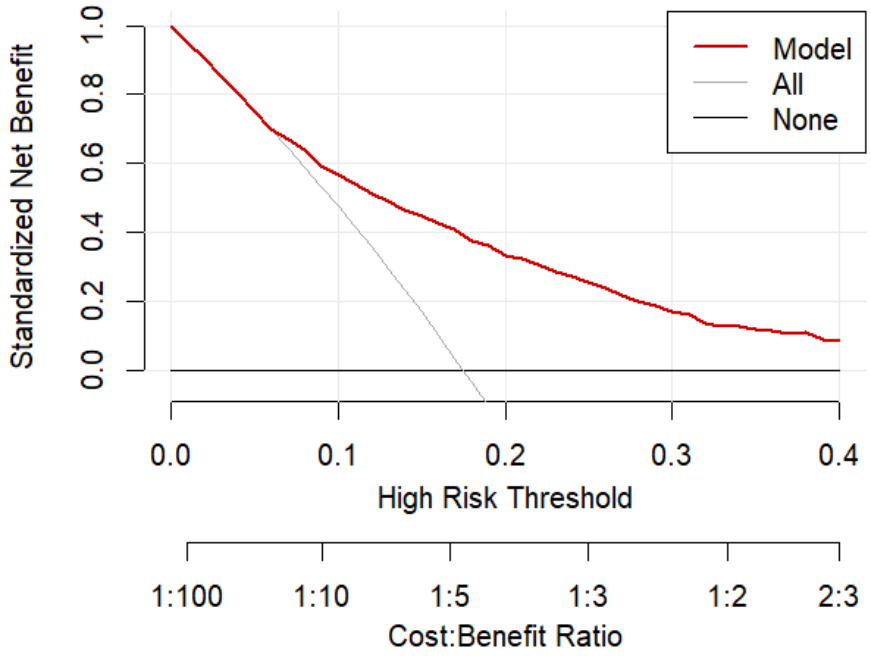
(**a**) (**b**)

Figure 4. The calibration plot of the LR model.(a): Calibration curve of LR model in test set. (b): Calibration curve of LR model in training set.



(**a**)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 July 2024 doi:10.20944/preprints202407.1736.v1



(**b**)

Figure 5. (a)The nomogram of the LR model(b): Decision curve analysis of GBC distant metastasis.

3.2. Analysis of Machine Learning Algorithm Results

Based on accuracy, precision, recall, F1 score, and AUC value, 7 machine learning models are developed and compared. The machine learning model trained by over-sampling data is better than that trained by under-sampling data, see Table 3 and Table 4 for the details of 7 machine learning models constructed by over-sampling and under-sampling data.Using over-sampling and under-sampling to build seven machine learning models, the performance of the training set and test set is shown in Figure 6. Among them, the performance of the RF model is better than other models, with accuracy rate of 0.828, precision 0.811, recall rate 0.862, F1 score 0.836, and AUC 0.913. The calibration curves of the RF model in the test and training sets are shown in Figure 7, The RF model has good calibration in both the training and testing sets. Using the RF model for feature selection, as shown in Figure 7C, it can be seen that grade is a key predictor of distant metastasis in T1 and T2 GBC patients.

Table 3. Comparison prediction performances of different models for Over-sampling.

Model	Accuracy	AUC	Precision	Recall rate	F1-score
NB	0.681	0.739	0.734	0.587	0.652
SVC	0.707	0.781	0.722	0.690	0.706
KNN	0.738	0.822	0.721	0.791	0.761
DT	0.681	0.891	0.686	0.688	0.687
RF	0.828	0.913	0.811	0.862	0.836
XGBoost	0.784	0.877	0.781	0.799	0.790
GBM	0.704	0.789	0.711	0.704	0.707

Table 4. Comparison prediction performances of different models for under-sampling.

Model	Accuracy	AUC	Precision	Recall rate	F1-score
NB	0.689	0.735	0.715	0.549	0.621

 doi:10.20944/preprints202407.1736.v1

SVC	0.702	0.763	0.691	0.647	0.669
KNN	0.604	0.715	0.562	0.661	0.687
DT	0.699	0.649	0.676	0.676	0.676
RF	0.686	0.739	0.643	0.725	0.682
XGBoost	0.656	0.712	0.624	0.654	0.639
GBM	0.702	0.765	0.683	0.669	0.676

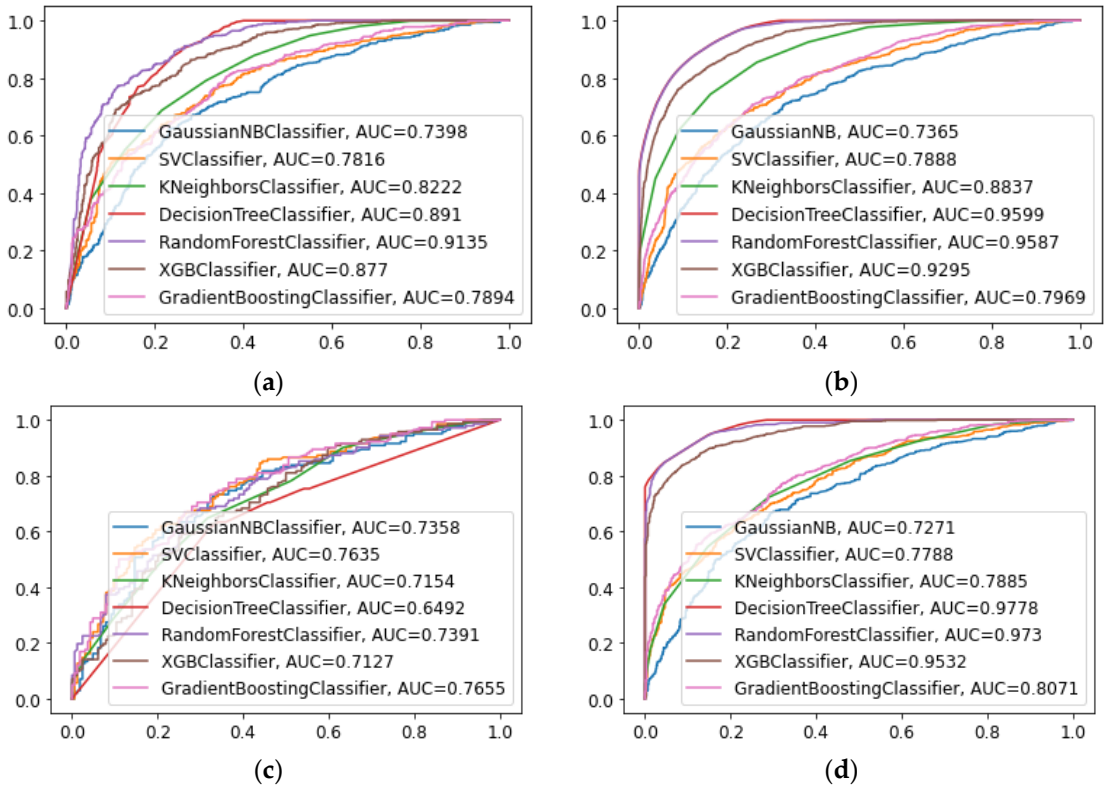
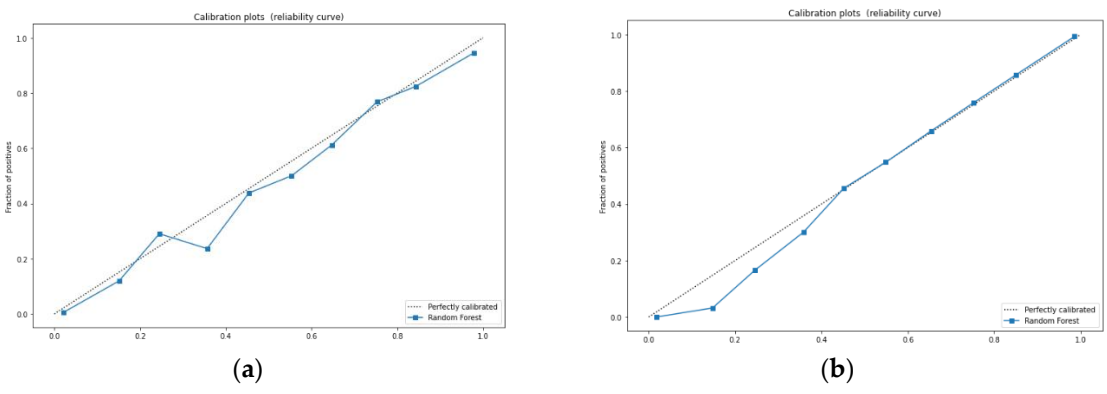
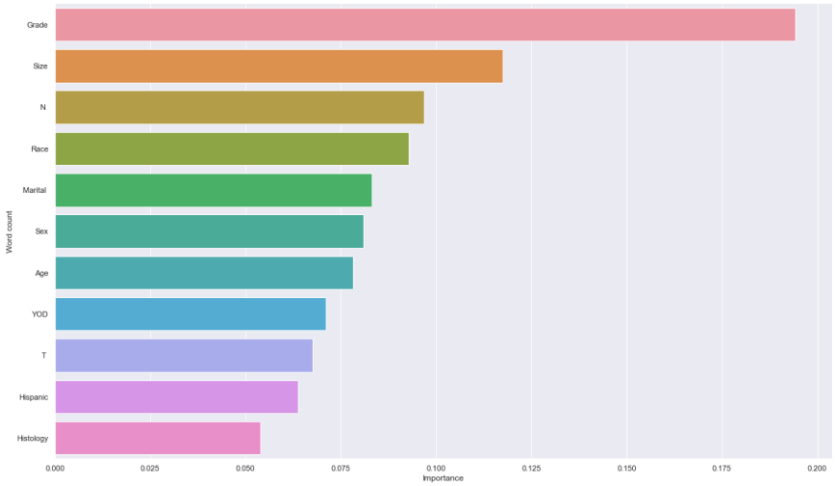


Figure 6. ROC curves of 7 ML algorithms in different datasets. (a): The ROC curves of the 7 ML algorithms model in the test set with over-sampling. (b): The ROC curves of the 7ML algorithms model in the training set with over-sampling. (c): The ROC curves of the 7 ML algorithms model in the test set with under-sampling. (d): The ROC curves of the 7 ML algorithms model in the training set with under-sampling.



 doi:10.20944/preprints202407.1736.v1



(c)

Figure 7. (a): Calibration curve of RF model in test set. (b): Calibration curve of RF model intraining set. (c): Feature importance derived from the RF model.

4. Discussion

In this study, we used machine learning algorithms combined with clinical pathological features to construct a predictive model for predicting distant metastasis of gallbladder cancer. Compared with previous studies, this study predicts and analyzes the distant metastasis of GBC patients by constructing a machine learning algorithm model. The results showed that based on the SEER database, by comparing the predictive performance of seven machine learning algorithms, we found that the model based on the RF algorithm performed the best and had higher predictive performance.

Although gallbladder cancer is relatively rare and its incidence rate increases slowly, it is still the most common malignant tumor in the bile duct system [2,14]. The treatment effect is poor when GBC progresses to the middle and late stages. The overall survival rate (OS) of GBC patients is about 17.8% -21.7%, and the OS in 5 years is only 5% [15–17]. The 5-year survival rate of T1 stage GBC patients is as high as 95% -100%, while the 5-year survival rates of T3 and T4 stage patients are only 23% and 12% [18]. The prognosis of GBC patients with distant metastasis is worse than that of GBC patients without metastasis, and the 1-year survival rate is between 20% -50% [7,9]. Therefore, exploring the risk of distant metastasis of early gallbladder cancer and establishing corresponding predictive models are crucial for early identification and clinical intervention of distant metastasis of gallbladder cancer, thereby improving prognosis. At present, research on distant metastasis of gallbladder cancer mainly focuses on exploring disease prognosis, and mostly relies on nomograms established based on traditional LR models or COX competitive risk models [6,19,20]. The traditional logistic regression model evaluates the association between risk factors and specific outcomes, and reflects the strength of the relationship between risk factors and outcomes by generating corresponding coefficients. At the same time, logistic regression models also have some shortcomings, such as being sensitive to multicollinearity and lacking mechanisms to prevent overfitting [21]. With the continuous progress of artificial intelligence technology, the application of ML models in tumor diagnosis and prognosis assessment is becoming increasingly common [22,23]. The ML algorithm also compensates for the shortcomings of traditional logistic regression models, such as overfitting and imbalanced data distribution [24]. In this study, we applied the ML algorithm for the first time to predict distant metastasis of T1 and T2 stage gallbladder cancer, with the aim of effectively improving patient prognosis through early intervention.

The aim of this study is to construct a machine learning model to predict the distant metastasis of T1 and T2 stage gallbladder cancer patients, and to predict the relevant factors affecting the distant metastasis of GBC patients through logistic regression analysis.

Univariate and multivariate logistic regression analysis showed that age, history, tumor size, T stage, N stage, and grade were all predictive factors for distant metastasis of gallbladder cancer,This

 doi:10.20944/preprints202407.1736.v1

is consistent with previous research findings [6]. Similar to the results presented by logistic regression,The feature importance of the RF model also indicate that grade is a key predictive variable for evaluating distant metastasis of gallbladder cancer.Tumor grade is an indicator used to evaluate the similarity of morphological and functional features between tumor cells and source organ tissues [25].

Previous studies have also found that grade plays an important predictive role in the distant metastasis and prognosis of gallbladder cancer patients [6,7,20]. The higher the grade, the poorer the cell differentiation, while higher grades typically have higher invasiveness, a wider range of infiltration, and are more prone to distant metastasis [20].

Studies have shown [26] that poorly differentiated GBCs are more likely to undergo distant metastasis, which is similar to the conclusion of this study.Lymph node status is a commonly used predictive factor for evaluating the metastasis and prognosis of gastrointestinal malignant tumors [27,28], and a thorough evaluation of lymph node status is also a necessary condition for patient treatment [29,30]. This study found that N stage is an important factor in predicting the occurrence of distant metastasis in gallbladder cancer. LR regression shows that when lymph node metastasis is detected, the probability of GBC developing distant metastasis is higher. This study found that gallbladder cancer patients with tumors larger than or equal to 2cm are more likely to develop distant metastasis, which is consistent with previous research results [6].

ML can use computers to mimic human learning abilities and improve its performance by rebuilding data analysis models [31], In the past decade, machine learning algorithms have been widely applied in the medical field and have achieved remarkable results in the diagnosis, treatment, and prognosis of diseases [32]. Compared with traditional data analysis methods, machine learning has significant advantages. On the one hand, it can process large datasets more efficiently; On the other hand, machine learning can handle nonlinear data more reasonably through different algorithms and statistical models, while traditional methods may not achieve satisfactory expected results when dealing with nonlinear data. In many studies [13], the predictive performance of machine learning is superior to traditional methods. In this study, RF is one of the effective machine learning models. The RF model adopts advanced classification decisions and different weighting ratios, which not only outperforms other technologies in processing large amounts of features and highly nonlinear data, but also improves the utilization of analytical information, thereby constructing a prediction model with better predictive performance [12].

We constructed 7 predictive models based on the SEER database to evaluate the distant metastasis of T1 and T2 gallbladder cancer patients. The 7 algorithm models were evaluated by accuracy, precision, recall, F1 score, and AUC value Amongst them, RF has good predictive ability (AUC=0.913, F1 score=0.836). The RF algorithm is the best model for predicting distant metastasis of gallbladder cancer using the SEER database.

This study also has some limitations: 1) As it is based on North American demographic data, it needs to be validated with external populations in future studies. 2) The efficiency of this model is expected to be further improved, and more risk factors can be incorporated in the future. 3) The SEER database lacks important information such as tumor family history and bilirubin，as well as tumor markers, which may also be important predictive factors for distant cancer metastasis. In response to the above issues, we will collect more information and conduct in-depth supplementary research in future research.

5. Conclusions

This study developed and validated a prediction model based on machine learning algorithms, which utilizes clinical features and quantitative indicators to predict distant metastasis of T1 and T2 gallbladder cancer. Among these seven predictive models, the RF algorithm is more predictive, providing personalized treatment and more efficient allocation of medical resources for patients.

Author Contributions: Author Contributions: Conceptualization, Z.G. and Z.Z.; methodology, Z.G. and Z.Z; software, Z.G.; vali-dation, L.L. and Z.Y.; investigation, Z.L. and C.Z.; resources,J.F. and P.Y.; data curation,

Z.G.;writing—original draft preparation, Z.G.; writing—review and editing, Z.Z.; supervision, Z.Z. funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Beijing Municipal Science & Technology Commission (No. Z171100000417056) and the Key Support Project of Guo Zhong Health Care of China General Technology Group (GZKJ-KJXX-QTHT-20230626).

Institutional Review Board Statement: Adopting AJCC 7th edition TNM stage, as the SEER database contains public data, informed consent from relevant patients for the use of the SEER database for research purposes was not required, nor was ethical approval. Our request for access to the SEER data was approved by the National Cancer Institute, USA (reference number 19238-Nov2021).

Data Availability Statement: Details of the data access process are available online. (https://github.com/990305/GBC).

Conflicts of Interest: The authors declare no competing interests.

References

1. Ji, Z.; Ren, L.; Liu, F.; Liu, L.; Song, J.; Zhu, J.; Ji, G.; Huang, G. Effect of different surgical options on the long-term survival of stage I gallbladder cancer: a retrospective study based on SEER database and Chinese Multi-institutional Registry. Journal of cancer research and clinical oncology 2023, 149, 12297-12313, doi:10.1007/s00432-023-05116-z.
2. Huang, J.; Patel, H.K.; Boakye, D.; Chandrasekar, V.T.; Koulaouzidis, A.; Lucero-Prisno Iii, D.E.; Ngai, C.H.; Pun, C.N.; Bai, Y.; Lok, V.; et al. Worldwide distribution, associated factors, and trends of gallbladder cancer: A global country-level analysis. Cancer letters 2021, 521, 238-251, doi:10.1016/j.canlet.2021.09.004.
3. Torres, O.J.M.; Alikhanov, R.; Li, J.; Serrablo, A.; Chan, A.C.; de Souza, M.F.E. Extended liver surgery for gallbladder cancer revisited: Is there a role for hepatopancreatoduodenectomy? International journal of surgery (London, England) 2020, 82s, 82-86, doi:10.1016/j.ijsu.2020.05.085.
4. Lim, H.; Seo, D.W.; Park, D.H.; Lee, S.S.; Lee, S.K.; Kim, M.H.; Hwang, S. Prognostic factors in patients with gallbladder cancer after surgical resection: analysis of 279 operated patients. Journal of clinical gastroenterology 2013, 47, 443-448, doi:10.1097/MCG.0b013e3182703409.
5. Sharma, A.; Sharma, K.L.; Gupta, A.; Yadav, A.; Kumar, A. Gallbladder cancer epidemiology, pathogenesis and molecular genetics: Recent update. World journal of gastroenterology 2017, 23, 3978-3998, doi:10.3748/wjg.v23.i22.3978.
6. Cai, Y.L.; Lin, Y.X.; Jiang, L.S.; Ye, H.; Li, F.Y.; Cheng, N.S. A Novel Nomogram Predicting Distant Metastasis in T1 and T2 Gallbladder Cancer: A SEER-based Study. International journal of medical sciences 2020, 17, 1704-1712, doi:10.7150/ijms.47073.
7. Yang, Y.; Tu, Z.; Ye, C.; Cai, H.; Yang, S.; Chen, X.; Tu, J. Site-specific metastases of gallbladder adenocarcinoma and their prognostic value for survival: a SEER-based study. BMC surgery 2021, 21, 59, doi:10.1186/s12893-021-01068-8.
8. Wang, X.; Yu, G.Y.; Chen, M.; Wei, R.; Chen, J.; Wang, Z. Pattern of distant metastases in primary extrahepatic bile-duct cancer: A SEER-based study. Cancer medicine 2018, 7, 5006-5014, doi:10.1002/cam4.1772.
9. Mady, M.; Prasai, K.; Tella, S.H.; Yadav, S.; Hallemeier, C.L.; Rakshit, S.; Roberts, L.; Borad, M.; Mahipal, A. Neutrophil to lymphocyte ratio as a prognostic marker in metastatic gallbladder cancer. HPB : the official journal of the International Hepato Pancreato Biliary Association 2020, 22, 1490-1495, doi:10.1016/j.hpb.2020.02.002.
10. Zhu, X.; Zhang, X.; Hu, X.; Ren, H.; Wu, S.; Wu, J.; Wu, G.; Si, X.; Wang, B. Survival analysis of patients with primary gallbladder cancer from 2010 to 2015: A retrospective study based on SEER data. Medicine 2020, 99, e22292, doi:10.1097/md.0000000000022292.
11. Zhong, X.; Lin, Y.; Zhang, W.; Bi, Q. Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning. Scientific reports 2023, 13, 18301, doi:10.1038/s41598-023-45438-z.
12. Liu, W.C.; Li, Z.Q.; Luo, Z.W.; Liao, W.J.; Liu, Z.L.; Liu, J.M. Machine learning for the prediction of bone metastasis in patients with newly diagnosed thyroid cancer. Cancer medicine 2021, 10, 2802-2811, doi:10.1002/cam4.3776.
13. Mao, Y.; Lan, H.; Lin, W.; Liang, J.; Huang, H.; Li, L.; Wen, J.; Chen, G. Machine learning algorithms are comparable to conventional regression models in predicting distant metastasis of follicular thyroid carcinoma. Clinical endocrinology 2023, 98, 98-109, doi:10.1111/cen.14693.

 doi:10.20944/preprints202407.1736.v1

14. Wernberg, J.A.; Lucarelli, D.D. Gallbladder cancer. The Surgical clinics of North America 2014, 94, 343-360, doi:10.1016/j.suc.2014.01.009.
15. Zhong, Y.; Wu, X.; Li, Q.; Ge, X.; Wang, F.; Wu, P.; Deng, X.; Miao, L. Long noncoding RNAs as potential biomarkers and therapeutic targets in gallbladder cancer: a systematic review and meta-analysis. Cancer cell international 2019, 19, 169, doi:10.1186/s12935-019-0891-1.
16. Shen, H.; He, M.; Lin, R.; Zhan, M.; Xu, S.; Huang, X.; Xu, C.; Chen, W.; Yao, Y.; Mohan, M.; et al. PLEK2 promotes gallbladder cancer invasion and metastasis through EGFR/CCL2 pathway. Journal of experimental & clinical cancer research : CR 2019, 38, 247, doi:10.1186/s13046-019-1250-8.
17. Hundal, R.; Shaffer, E.A. Gallbladder cancer: epidemiology and outcome. Clinical epidemiology 2014, 6, 99-109, doi:10.2147/clep.S37357.
18. Shindoh, J.; de Aretxabala, X.; Aloia, T.A.; Roa, J.C.; Roa, I.; Zimmitti, G.; Javle, M.; Conrad, C.; Maru, D.M.; Aoki, T.; et al. Tumor location is a strong predictor of tumor progression and survival in T2 gallbladder cancer: an international multicenter study. Annals of surgery 2015, 261, 733-739, doi:10.1097/sla.0000000000000728.
19. Zhang, W.; Chen, Z.; Sa, B. Construction and validation of the predictive model for gallbladder cancer liver metastasis patients: a SEER-based study. European journal of gastroenterology & hepatology 2024, 36, 129-134, doi:10.1097/meg.0000000000002678.
20. Fang, C.; Li, W.; Wang, Q.; Wang, R.; Dong, H.; Chen, J.; Chen, Y. Risk factors and prognosis of liver metastasis in gallbladder cancer patients: A SEER-based study. Frontiers in surgery 2022, 9, 899896, doi:10.3389/fsurg.2022.899896.
21. Leonard, G.; South, C.; Balentine, C.; Porembka, M.; Mansour, J.; Wang, S.; Yopp, A.; Polanco, P.; Zeh, H.; Augustine, M. Machine Learning Improves Prediction Over Logistic Regression on Resected Colon Cancer Patients. The Journal of surgical research 2022, 275, 181-193, doi:10.1016/j.jss.2022.01.012.
22. Guo, Z.T.; Tian, K.; Xie, X.Y.; Zhang, Y.H.; Fang, D.B. Machine Learning for Predicting Distant Metastasis of Medullary Thyroid Carcinoma Using the SEER Database. International journal of endocrinology 2023, 2023, 9965578, doi:10.1155/2023/9965578.
23. Han, T.; Zhu, J.; Chen, X.; Chen, R.; Jiang, Y.; Wang, S.; Xu, D.; Shen, G.; Zheng, J.; Xu, C. Application of artificial intelligence in a real-world research for predicting the risk of liver metastasis in T1 colorectal cancer. Cancer cell international 2022, 22, 28, doi:10.1186/s12935-021-02424-7.
24. Ahn, J.H.; Kwak, M.S.; Lee, H.H.; Cha, J.M.; Shin, H.P.; Jeon, J.W.; Yoon, J.Y. Development of a Novel Prognostic Model for Predicting Lymph Node Metastasis in Early Colorectal Cancer: Analysis Based on the Surveillance, Epidemiology, and End Results Database. Frontiers in oncology 2021, 11, 614398, doi:10.3389/fonc.2021.614398.
25. Osório, F.M.; Vidigal, P.V.; Ferrari, T.C.; Lima, A.S.; Lauar, G.M.; Couto, C.A. Histologic Grade and Mitotic Index as Predictors of Microvascular Invasion in Hepatocellular Carcinoma. Experimental and clinical transplantation : official journal of the Middle East Society for Organ Transplantation 2015, 13, 421-425, doi:10.6002/ect.2015.0045.
26. Butte, J.M.; Gönen, M.; Allen, P.J.; D'Angelica, M.I.; Kingham, T.P.; Fong, Y.; Dematteo, R.P.; Blumgart, L.; Jarnagin, W.R. The role of laparoscopic staging in patients with incidental gallbladder cancer. HPB : the official journal of the International Hepato Pancreato Biliary Association 2011, 13, 463-472, doi:10.1111/j.1477-2574.2011.00325.x.
27. Shirai, Y.; Sakata, J.; Wakai, T.; Ohashi, T.; Ajioka, Y.; Hatakeyama, K. Assessment of lymph node status in gallbladder cancer: location, number, or ratio of positive nodes. World journal of surgical oncology 2012, 10, 87, doi:10.1186/1477-7819-10-87.
28. Qiu, B.; Su, X.H.; Qin, X.; Wang, Q. Application of machine learning techniques in real-world research to predict the risk of liver metastasis in rectal cancer. Frontiers in oncology 2022, 12, 1065468, doi:10.3389/fonc.2022.1065468.
29. Sakata, J.; Shirai, Y.; Wakai, T.; Ajioka, Y.; Hatakeyama, K. Number of positive lymph nodes independently determines the prognosis after resection in patients with gallbladder carcinoma. Annals of surgical oncology 2010, 17, 1831-1840, doi:10.1245/s10434-009-0899-1.
30. Negi, S.S.; Singh, A.; Chaudhary, A. Lymph nodal involvement as prognostic factor in gallbladder cancer: location, count or ratio? Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract 2011, 15, 1017-1025, doi:10.1007/s11605-011-1528-4.

 doi:10.20944/preprints202407.1736.v1

31. Feng, S.; Wang, J.; Wang, L.; Qiu, Q.; Chen, D.; Su, H.; Li, X.; Xiao, Y.; Lin, C. Current Status and Analysis of Machine Learning in Hepatocellular Carcinoma. Journal of clinical and translational hepatology 2023, 11, 1184-1191, doi:10.14218/jcth.2022.00077s.
32. Bhinder, B.; Gilvary, C.; Madhukar, N.S.; Elemento, O. Artificial Intelligence in Cancer Research and Precision Medicine. Cancer discovery 2021, 11, 900-915, doi:10.1158/2159-8290.Cd-21-0090.