

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 ^{1,2}

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.); liulimin15@sina.com (L.L.); zhaoyue806@sina.com (Y.Z.); lz0002@163.com (Z.L.); zhangchong2118@163.com (C.Z.); qihui6@gt.cn (H.Q.); haimingyuan@aliyun.com (H.Y.)

² Key Laboratory of Geriatrics (Hepatobiliary Diseases) of China General Technology Group, Beijing 100073, China; 18811777978@163.com (J.F.); yjp5209@163.com (P.Y.)

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

Abstract: (1) Background: This study seeks to employ a machine learning (ML) algorithm to forecast the risk of distant metastasis (DM) in patients with T1 and T2 gallbladder cancer (GBC); (2) Methods: Data of patients diagnosed with T1 and T2 GBC was obtained from SEER, encompassing the period from 2004 to 2015, were utilized to apply seven ML algorithms. These algorithms were appraised by the area under the receiver operating characteristic curve (AUC) and other metrics; (3) Results: This study involved 4371 patients in total. Out of these patients, 764 (17.4%) cases progressed to develop DM. Utilizing a logistic regression (LR) model to identify independent risk factors for DM of gallbladder cancer (GBC). A nomogram has been developed to forecast DM in early T-stage gallbladder cancer patients. Through the evaluation of different models using relevant indicators, it was discovered that Random Forest (RF) exhibited the most outstanding predictive performance; (4) Conclusions: RF has demonstrated high accuracy in predicting DM in gallbladder cancer patients, assisting clinical physicians in enhancing the accuracy of diagnosis. This can be particularly valuable for improving patient outcomes and optimizing treatment strategies. We employ the RF algorithm to construct the corresponding web calculator.

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



Citation: Guo, Z.; Zhang, Z.; Liu, L.; Zhao, Y.; Liu, Z.; Zhang, C.; Qi, H.; Feng, J.; Yao, P.; Yuan, H. Machine Learning Algorithm for Predicting Distant Metastasis of T1 and T2 Gallbladder Cancer Based on SEER Database. *Bioengineering* **2024**, *11*, 927. <https://doi.org/10.3390/bioengineering11090927>

Academic Editors: Yunfeng Wu and Maria Francesca Spadea

Received: 16 July 2024

Revised: 5 September 2024

Accepted: 12 September 2024

Published: 15 September 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Gallbladder cancer (GBC) is characterized by its insidious onset, early metastasis, and poor prognosis [1,2]. The 5-year survival rates for gallbladder cancer in the T3 and T4 stages are 32.4% and 3.5%, respectively [3,4]. Currently, early diagnostic methods with high specificity and sensitivity for GBC remain lacking, and most GBC cases are detected in the middle to late stages [5]. Research indicates that the incidence of lymph node and distant metastasis (DM) in GBC patients ranges from 17.9% to 64.5%, with the liver, lungs, and peritoneum being the most commonly affected organs by metastasis. [6–8]. Among gallbladder cancer patients, those with DM have a poorer prognosis compared to those without [7,9]. Distant metastasis is a strongly correlated factor that significantly impacts the prognosis of GBC [10]. Early assessment of distant metastasis risk is vital for timely intervention and enhancing the prognosis of GBC in the T1 and T2 stages. Although, nomogram is currently the most widely utilized clinical tool. At the same time, there is also a continuous development of new technologies. Machine learning (ML) algorithms are increasingly employed to build clinical models due to their practicality, innovation, and accuracy [11]. Machine learning algorithms hold great promise in harnessing complex and massive clinical data for disease diagnosis and outcome prediction. Previous studies have

demonstrated that machine learning offers more advantages compared to traditional big data clinical prediction research methods [12].

Therefore, our objective is to utilize an ML algorithm to predict the occurrence of DN in GBC. Our study can assist clinicians in making more precise predictions, thus enhancing their ability to tailor treatment plans to individual patients, improve patient prognosis through early intervention, and enhance the patient prognosis even further.

2. Materials and Methods

2.1. Data Sources

The patient data originates from the SEER, utilizing SEER*stat 8.4.2 software. This research encompassed individuals who were diagnosed with GBC during the period from 2004 to 2015. The process of patient screening is illustrated in Figure 1. The data criteria are outlined as follows: (1) The 6th edition of the AJCC TNM staging system served as the foundation; (2) Clear histological diagnosis; (3) For a single tumor.

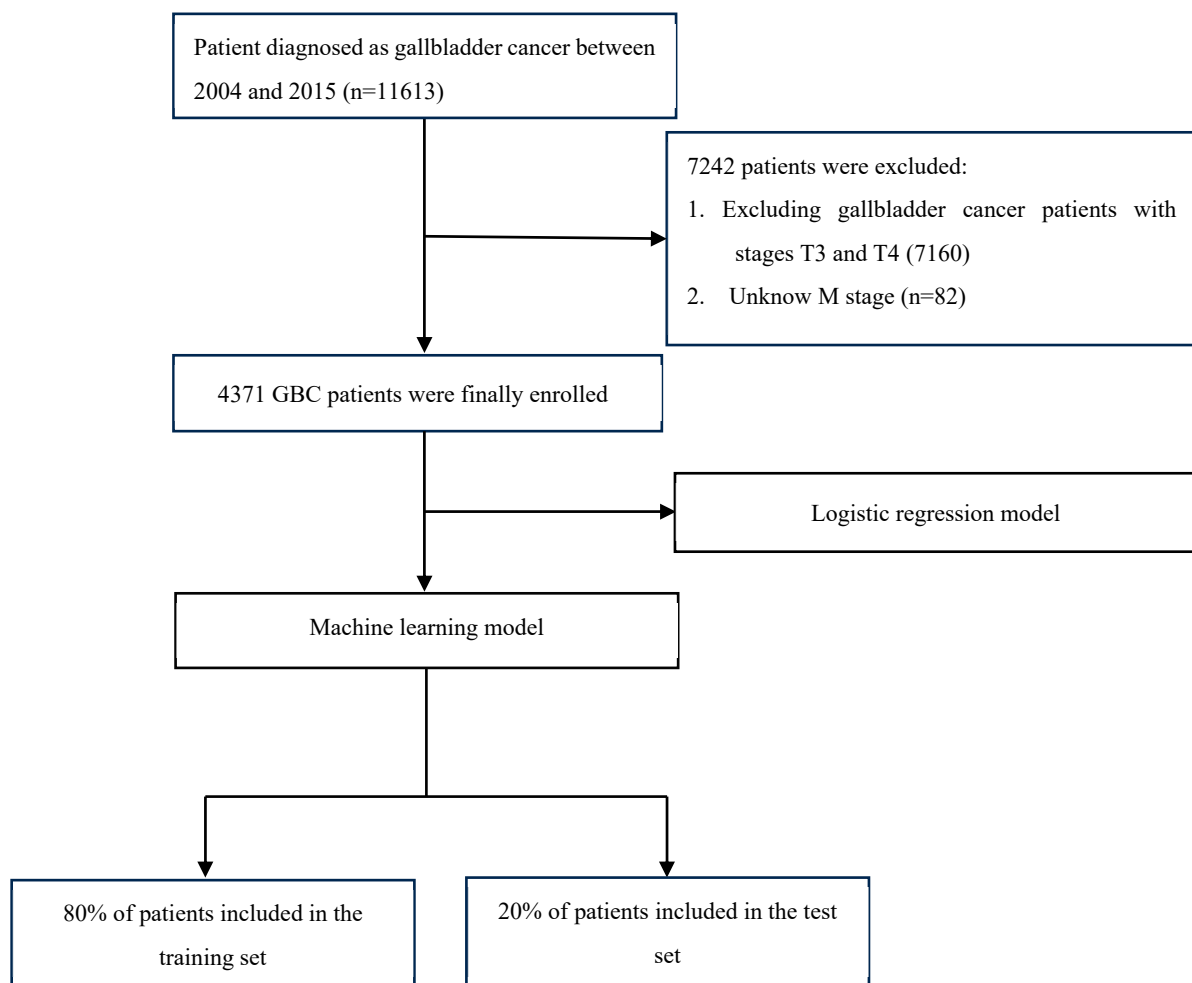


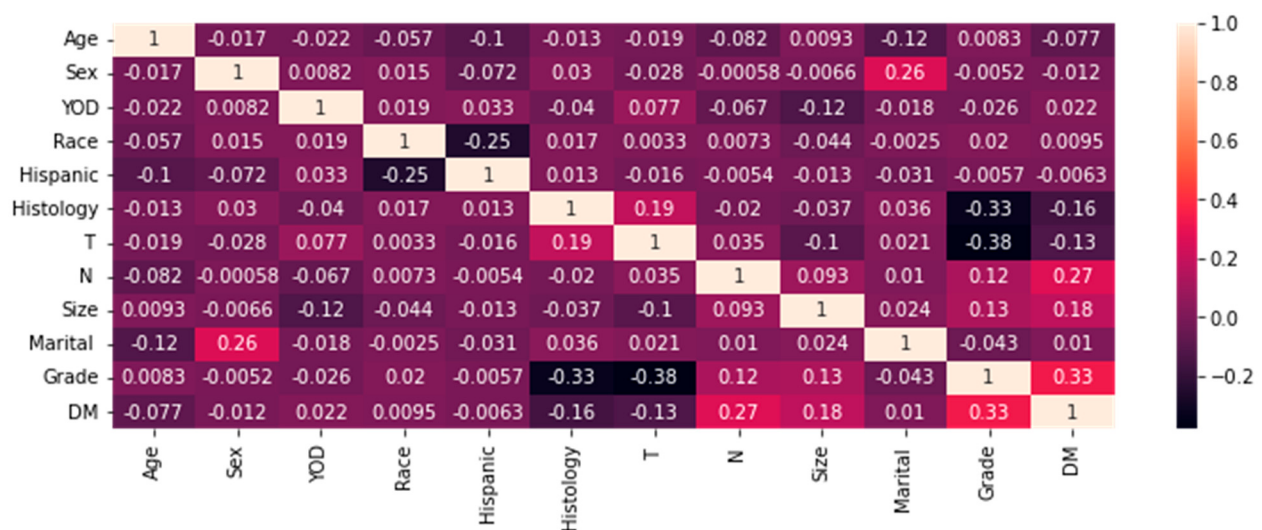
Figure 1. The flow diagram of the selection process for the study.

Exclusion criteria: (1) missing or incomplete data, including T staging, M staging, etc. Variables included age, sex, race, tumor size, year of diagnosis, N stage, Hispanic, histology, marital status, T stage, grade, and DM. Distant metastasis refers to the spread of a tumor to distant organs, such as the liver, bone, brain, lung, peritoneum, and other organs. SEER does not require informed consent from the patient concerned, nor does it require ethical approval. This study has received approval from the National Cancer Institute, with reference number 19238-Nov2021.

2.2. Model Construction

Statistical analysis of the data in this study was performed utilizing the SPSS software (version 26.0). Construct a nomogram prediction model for DM using R 4.3.2 and draw a calibration curve. All patients were randomly allocated into two subsets (including the training set and the test set) in an 8:2 ratio. Categorical variables are represented in the form of numerical values and corresponding percentages, providing a comprehensive and detailed presentation. Comparisons within the groups are carried out utilizing the chi-square test, Fisher's exact test, and Mann-Whitney U test. The utilization of these advanced testing methods enhances the depth and sophistication of our research findings, ultimately contributing to a more captivating and illuminating presentation of our results. Additionally, multiple categorical variables are also expressed as numbers and percentages for between-group comparisons, employing the chi-square test and Kruskal-Wallis test. The significance is adjusted using the Bonferroni correction. We establish a logistic regression (LR) based on univariate and multivariate logistic regression analysis and display them as a nomogram. A nomogram is a graphical representation that converts mathematical formulas into geometric expressions and explains the interactions between predicted variables. It is mainly used in logistic regression models and COX proportional risk models [13]. All computed *p*-values were two-sided, and statistical significance was considered at a level of less than 0.05.

Use Python software (version 3.9.12). Include all variables in the machine learning algorithm, and a prediction model is built. In the SEER database, there are fewer cases of DM in T1 and T2 gallbladder cancer, and the original dataset is imbalanced. We use under-sampling and over-sampling techniques to process the raw data. After conducting the sampling, the correlation between variables is illustrated in Figure 2. The darker the color, the higher the negative correlation coefficient; bright colors do the opposite; a and b represent the correlation between data after different sampling methods. We utilized seven commonly used ML algorithms to construct the predictive model, 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 accuracy, precision, recall, F1 score, and AUC value, and the model with the highest ROC value and F1 score is the optimal model.



(a)

Figure 2. Cont.

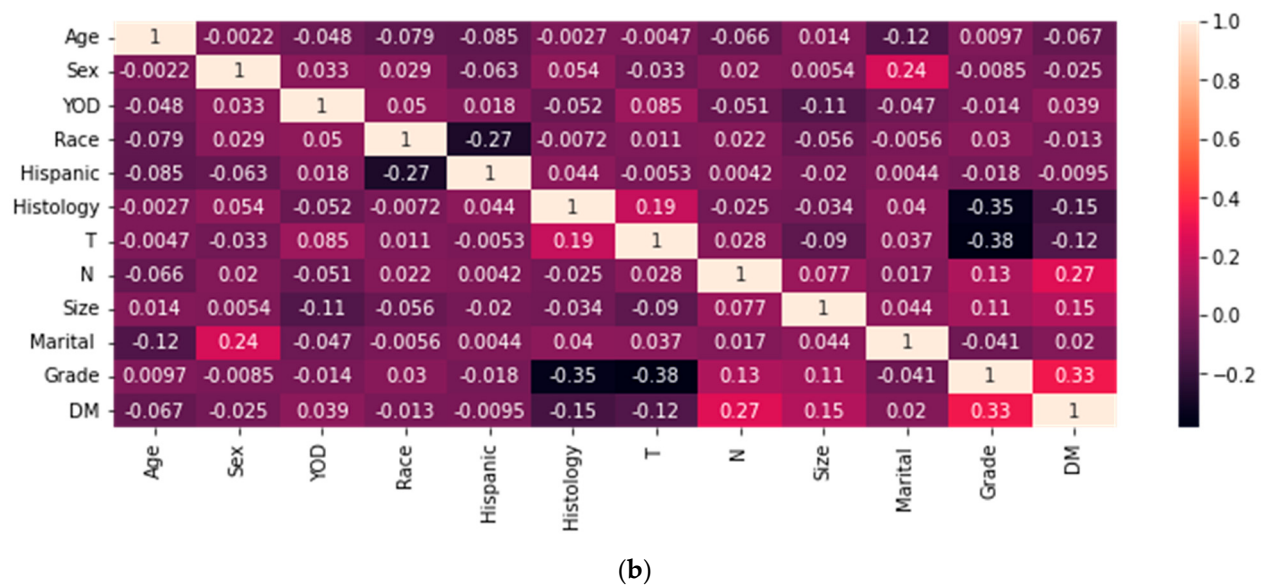


Figure 2. Correlation heatmaps of characteristics are featured in various datasets. (a): Data processed using over-sampling. (b): Data processed using under-sampling.

3. Results

3.1. Analysis of Patient Information

Our research encompassed 4371 patients diagnosed with T1 and T2 GBC. Among them, 764 patients presented with distant metastasis, while the remaining 3607 patients did not have distant metastasis. The vast majority of patients were of advanced age (≥ 70 years old, 56.9%), female (70.3%), and white (76.5%). Significant differences were observed in age, histology, tumor size, grade, N stage, and T stage between patients with or without DM ($p < 0.05$). The information of all patients is detailed in Table 2. During our exploration into the relationship between the training set, test set, and total set, we discovered that there were no significant differences in various indicators among the three datasets, as depicted in Table 2.

Table 1. Characteristics of the gallbladder cancer patients in T1 and T2.

Characteristic	Without DM (N = 3607)	With DM (N = 764)	<i>p</i> -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%)	
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
Adenocarcinoma	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%)	

Table 1. *Cont.*

Characteristic	Without DM (N = 3607)	With DM (N = 764)	p-Value
Tumor Size(cm)			<0.001
<2	838 (23.2%)	81 (10.6%)	
≥2	1543 (42.8%)	321 (42%)	
Unknown	1226 (34%)	362(47.4%)	<0.001
T Stage			
T1	1259 (34.9%)	361 (47.3%)	
T2	2348 (65.1%)	403 (52.7%)	<0.001
N stage			
N0	2871 (79.6%)	422 (55.2%)	
N1	644 (17.8%)	257 (33.7%)	0.531
NX	92 (2.6%)	85 (11.1%)	
Marital Status			
Single	1839 (51.0%)	380 (49.7%)	<0.001.
Married	1768 (49.0%)	384 (50.3%)	
Grade			
Grade I	737 (20.4%)	39 (5.1%)	<0.001.
Grade II	1536 (42.6%)	219 (28.6%)	
Grade III	894 (24.8%)	255 (33.4%)	
Grade IV	55 (1.5%)	18 (2.4%)	<0.001.
Unknown	385 (10.7)	233 (30.5%)	

Table 2. The relationship between the training set, testing set, and total set.

Characteristic	Training Set (N = 3498)	Test Set (N = 873)	Total Set (N = 4371)	p-Value
Age (Year)				1.000
<70	1506 (43.1%)	376 (43.1%)	1882 (43.1%)	
≥70	1992 (56.9%)	497 (56.9%)	2489 (56.9%)	
Gender				0.386
Female	2445 (69.9%)	631 (72.3%)	3076 (70.4%)	
Male	1053 (30.1%)	242 (27.7%)	1295 (29.6%)	
Race				0.883
White	2686 (76.8%)	662 (75.8%)	3348 (76.6%)	
Black	390 (11.1%)	107 (12.3%)	497 (11.34%)	
Other	422 (12.1%)	104 (11.9%)	526 (12.0%)	
Hispanic				0.256
YES	796 (22.8%)	176 (20.2%)	972 (22.2%)	
NO	2702 (77.2)	697 (79.8%)	3399 (77.8%)	
Histology				0.207
Adenocarcinoma	3122 (89.3%)	797 (91.3%)	3919 (89.7%)	
Others	376 (10.7%)	76 (8.7%)	452 (10.3%)	
Year of Diagnosis				0.844
2004–2009	1569 (44.9%)	382 (43.8%)	1951 (44.6%)	
2010–2015	1929 (55.1%)	491 (56.2%)	2420 (55.4%)	
Tumor size(cm)				0.969
<2	735 (21%)	184 (21.1%)	919 (21.1%)	
≥2	1488(42.5%)	376 (43%)	1864 (42.6%)	
Unknown	1275(36.4%)	313 (35.9%)	1588 (36.3%)	0.756
T Stage				
T1	1306 (37.3%)	314 (36%)	1620 (37.1%)	
T2	2192 (62.7%)	559 (64%)	2751 (62.9%)	0.104
N stage				
N0	2611 (74.6%)	682 (78.1%)	3293 (75.3%)	
N1	741 (21.2%)	160 (18.3%)	901 (20.6%)	
NX	146 (4.2%)	31 (3.6%)	177 (4.1%)	

Table 2. Cont.

Characteristic	Training Set (N = 3498)	Test Set (N = 873)	Total Set (N = 4371)	p-Value
Marital Status				0.998
Single	1775 (50.7%)	444 (50.9%)	2219 (50.8%)	
Married	1723 (49.3%)	429 (49.1%)	2152 (49.2%)	
Grade				0.620
Grade I	619 (17.7%)	157 (18.0%)	776 (17.7%)	
Grade II	1384 (39.6%)	371 (42.5%)	1755 (40.2%)	
Grade III	949 (27.1%)	200 (22.9%)	1149 (26.3%)	
Grade IV	51 (1.5%)	22 (2.5%)	73 (1.7%)	
Unknown	495 (14.1%)	123 (14.1%)	618 (14.1%)	
Distant Metastasis				0.998
YES	612 (17.5%)	152 (17.4%)	764 (17.5%)	
NO	2886 (82.5%)	721 (82.6%)	3607 (82.5%)	

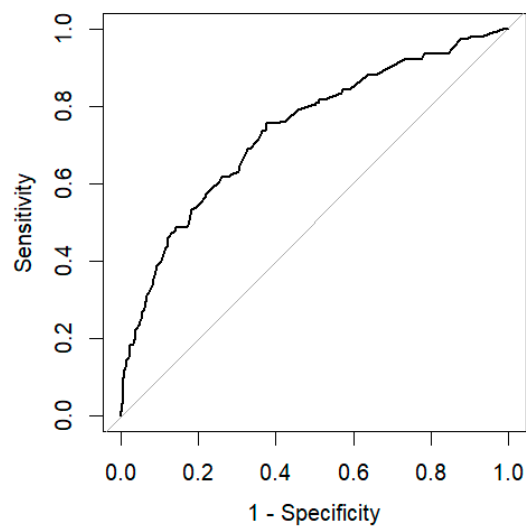
We employed LR to identify clinical factors that affect DM. Grade, histology, tumor size, T stage, N stage, and age were all identified as risk factors for DM in gallbladder cancer in univariate and multivariate logistic regression (Table 3). An LR model was developed with an AUC of 0.755 (95% CI: 0.734–0.776) in the test set and an AUC of 0.738 (95% CI: 0.693–0.783) in the training set (Figure 3). Figure 4 displays the calibration curves of LR in different data sets. The calibration curves indicate that the predicted probability curve is closely aligned with the actual values, indicating that LR has good calibration and is consistent with real-world observations. Significant independent factors were incorporated to establish the nomogram (Figure 5a), which clearly illustrates the impact of each risk factor on the outcome. Each factor corresponds to a specific score, which, when combined, provides the predicted probability of DM of early GBC. According to our nomogram, the risk of DM in early GBC ranges from 0.1 to 0.8. The DCA of the DM nomogram (Figure 5b) shows that within the threshold probability range of 1–40%, the net benefit of the model’s decision curve is higher than the net benefit of the two invalid lines. The decision curve illustrates the utility of the model’s predictions for decision-making at different probability thresholds. If the decision curve surpasses the two extreme case curves (i.e., the gray line assumes all patients have the disease, and the black line assumes none), the model is considered valuable.

Table 3. LR analysis in the training cohort.

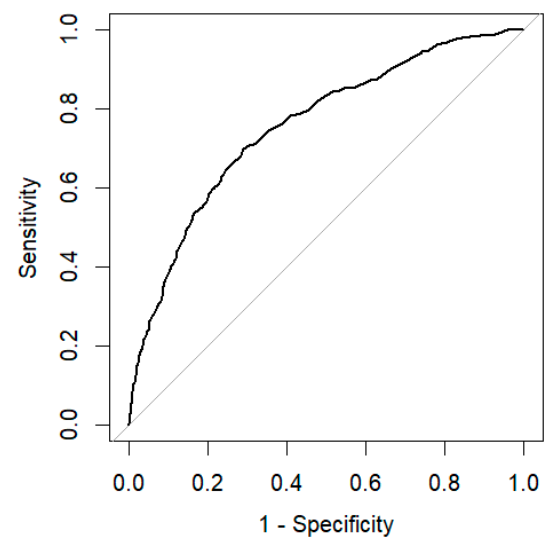
	Univariable Analysis			Multivariable Analysis		
	OR	95%CI	p Value	OR	95%CI	p Value
Age (Year)						
<70	Ref			Ref		
≥70	0.723	0.607–0.861	<0.001	0.705	0.583–0.852	<0.001
Gender						
Female	Ref					
Male	0.881	0.726–1.069	0.200			
Race						
White	Ref					
Black	1.116	0.850–1.464	0.431			
Other	0.980	0.746–1.287	0.885			
Hispanic						
YES	0.997	0.810–1.228	0.977			
NO	Ref					
Histology						
Adenocarcinoma	0.345	0.274–0.436	<0.001	0.595	0.456–0.777	<0.001

Table 3. Cont.

	Univariable Analysis			Multivariable Analysis		
	OR	95%CI	p Value	OR	95%CI	p Value
Others	Ref			Ref		
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 DM in GBC using LR models in the test set and training set. (a): ROC curve generated by the LR model in the test set. (b): ROC curve generated by the LR model in the training set.

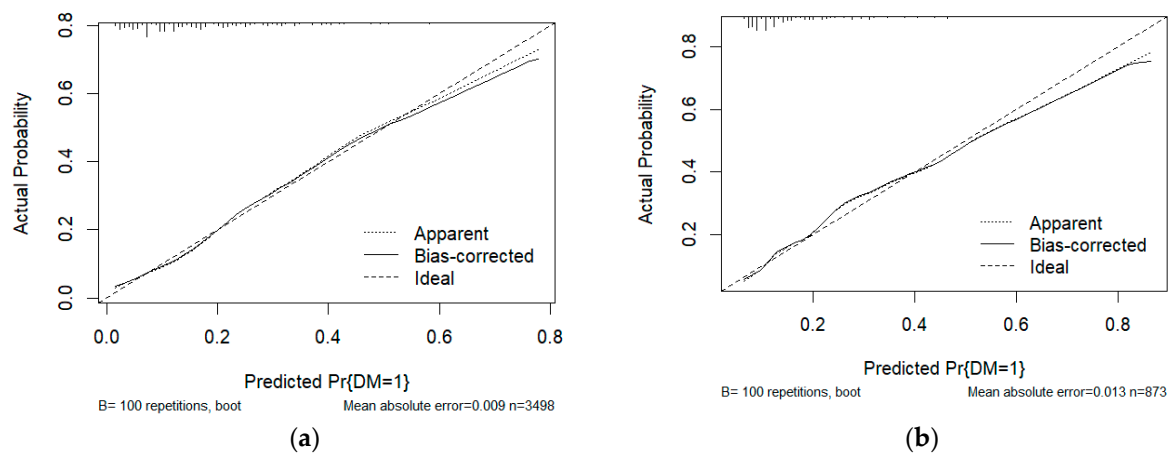


Figure 4. The calibration plot of the LR. (a): Calibration curve of LR in the test set. (b): Calibration curve of LR in the training set.

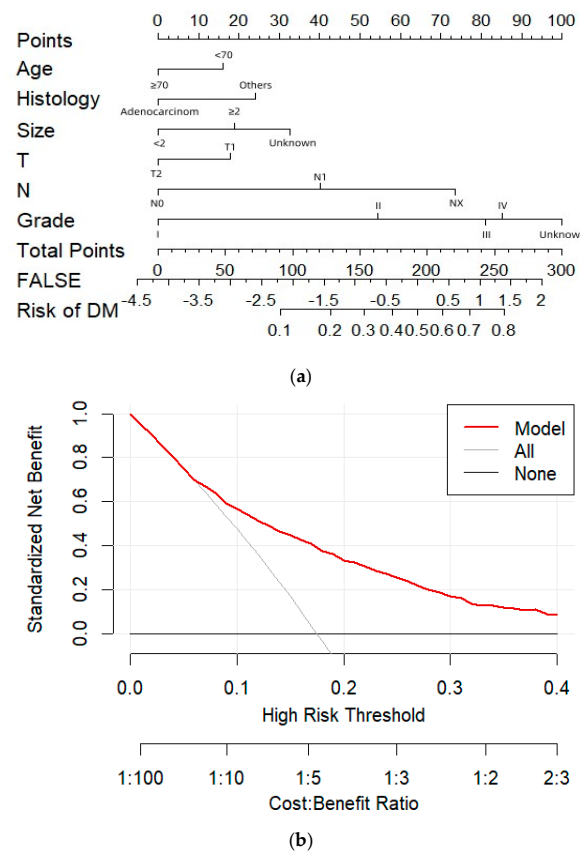


Figure 5. Nomogram and decision curve for predicting DM of early GBC. (a): The nomogram of the LR. (b): Decision curve analysis of GBC distant metastasis.

3.2. Analysis of Results Obtained from Machine Learning Algorithms

ML algorithms were evaluated by subsequent metrics, encompassing accuracy, precision, recall, F1 scores, and AUC values. The evaluation of these models was conducted with a comprehensive analysis of their performance across various key indicators. ML algorithms trained using over-sampling data outperformed those trained with under-sampling data. Detailed comparisons of the seven models constructed can be found in Tables 4 and 5. By employing over-sampling and under-sampling techniques, seven ML algorithms were constructed, and the performance of different data sets is depicted in Figure 6. Among them, the Random Forest (RF) model outperformed the others, achieving an accuracy rate

of 0.828, precision of 0.811, recall rate of 0.862, F1 score of 0.836, and an AUC of 0.913. The model also demonstrated excellent performance in the 5-fold cross-validation, with accuracies of 0.808, 0.819, 0.814, 0.817, and 0.812, respectively, with an average accuracy of 0.814, indicating robust performance. The calibration curves of the RF in both the test and training sets are depicted in Figure 7. RF model exhibited strong calibration in different data sets. Feature selection using RF, as shown in Figure 7c, identified grade as a key predictor of DM.

Table 4. Different model performance in over-sampling.

Model	Accuracy	AUC	Precision	Recall Rate	F1-Score
NB	0.681	0.740	0.734	0.587	0.652
SVC	0.707	0.782	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 5. Different model performance in under-sampling.

Model	Accuracy	AUC	Precision	Recall Rate	F1-Score
NB	0.689	0.736	0.715	0.549	0.621
SVC	0.702	0.764	0.691	0.647	0.669
KNN	0.604	0.715	0.562	0.661	0.687
DT	0.699	0.653	0.676	0.676	0.676
RF	0.686	0.743	0.643	0.725	0.682
XGBoost	0.656	0.713	0.624	0.654	0.639
GBM	0.702	0.766	0.683	0.669	0.676

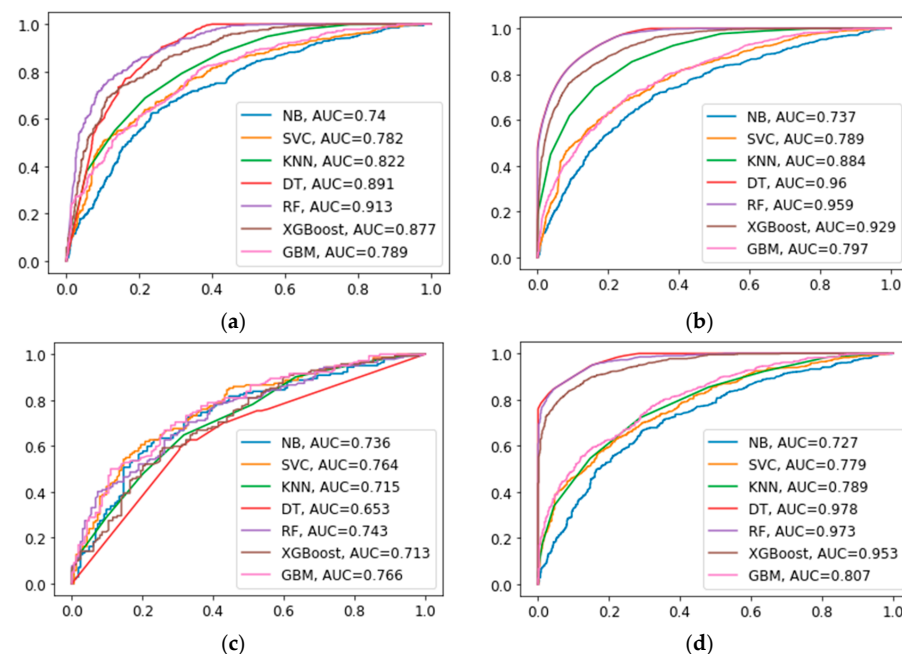


Figure 6. ROC curves for 7 machine learning algorithms across various datasets. (a): The ROC curves for the 7 machine learning algorithms in the test set were generated using over-sampling. (b): The ROC curves for the 7 machine learning algorithms in the training set were generated using over-sampling. (c): The ROC curves for the 7 machine learning algorithms in the test set were generated using under-sampling. (d): The ROC curves for the 7 machine learning algorithms in the training set were generated using over-sampling.

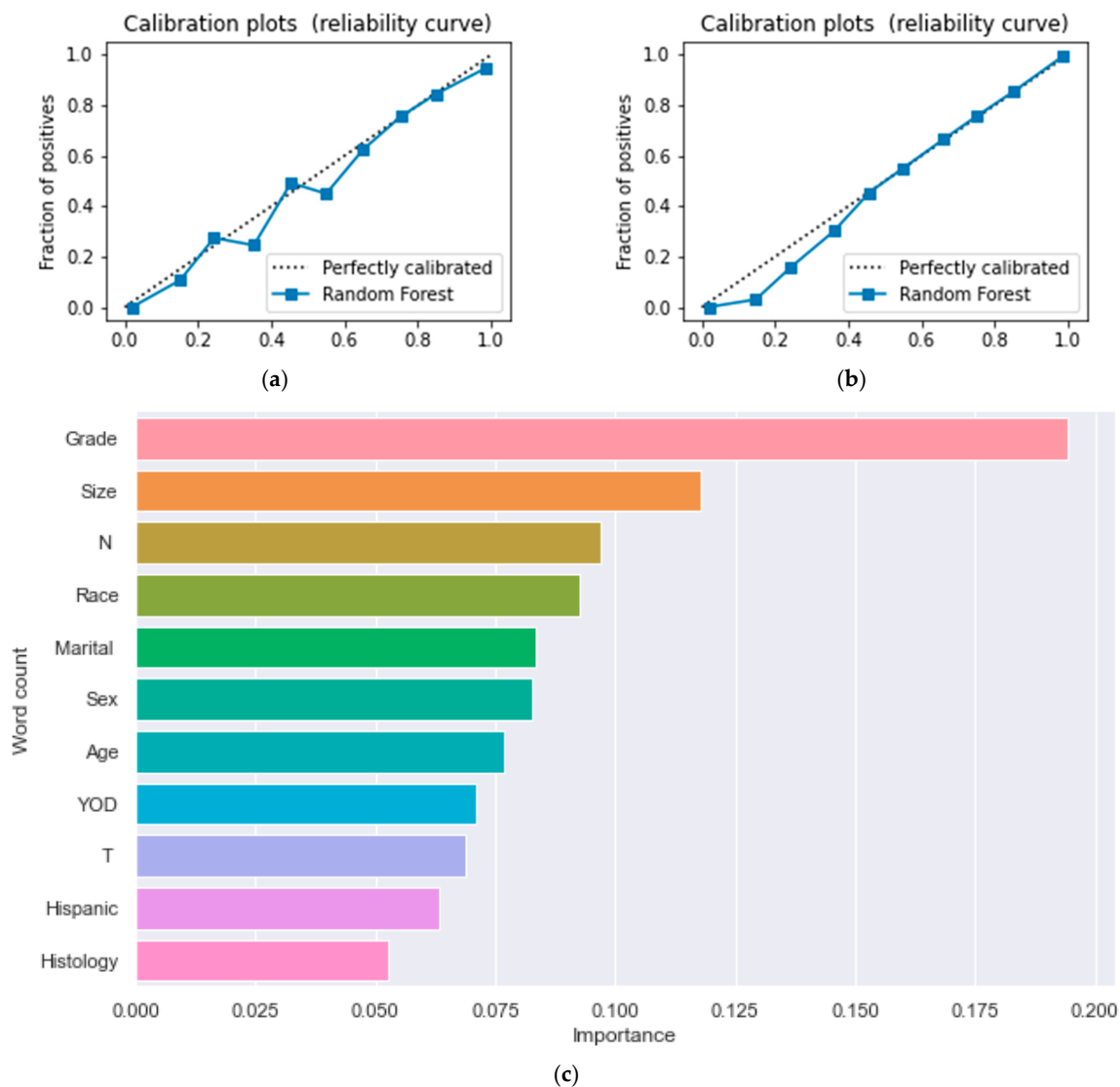


Figure 7. Calibration plots of RF in training and test sets and the importance of RF features. (a): Calibration curve of RF in the test set. (b): Calibration curve of RF in the training set. (c): Feature importance derived from the RF.

This research has established a web calculator for assessing the risk of DM in T1 and T2 GBC. The tool is intended for use in clinical settings and can be accessed on 1 September 2024 at <http://121.43.117.60:8004/>.

4. Discussion

In this study, we integrated ML algorithms with clinical pathological features to develop a model for DM in gallbladder cancer. Unlike previous studies, this research predicts and analyzes distant metastasis in GBC patients by constructing an ML algorithm. The results, based on the SEER database, demonstrated that among the seven machine learning algorithms compared, Random Forest demonstrated the most superior performance and achieved superior predictive accuracy.

Gallbladder cancer is relatively rare in the population and exhibits a slow increase in incidence. However, within the bile duct system, it is the most prevalent form of malignant tumor. [2,14]. Treatment outcomes are poor once GBC progresses to the middle and late stages. The overall survival (OS) rate for GBC patients is roughly 17.8–21.7%, with a 5-year OS of only 5% [15–17]. The 5-year survival rate for T1 stage GBC is between 95% and 100%, whereas, for T3 and T4 stage patients, it drops to only 23% and 12% respectively [18]. The

prognosis for GBC patients with distant metastasis is worse than those without, with a 1-year survival rate ranging between 20% and 50% [7,9]. Therefore, assessing the risk of DM in early GBC and developing predictive models are crucial for early identification and clinical intervention, ultimately improving patient prognosis. Currently, research on distant metastasis of gallbladder cancer primarily focuses on disease prognosis, relying mostly on nomograms developed from traditional logistic regression (LR) models or COX competitive risk models [6,19,20]. The traditional logistic regression model evaluates the association between risk factors and specific outcomes, quantifying the strength of the relationship through the generation of corresponding coefficients. However, LR models also have limitations, including sensitivity to multicollinearity and a lack of mechanisms to prevent overfitting [21]. With advancements in artificial intelligence technology, the application of ML in tumor diagnosis and prognosis assessment is becoming increasingly common [22,23]. The ML algorithm also addresses 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 DM of T1 and T2 stage gallbladder cancer, aiming to improve patient prognosis through early intervention effectively.

This research seeks to apply an ML algorithm to predict DM of T1 and T2 stage gallbladder cancer while identifying the key factors affecting metastasis through logistic regression analysis.

LR analysis identified age, histology, tumor size, T stage, N stage, and grade were all predictive factors for DM in GBC, consistent with previous research findings [6]. Similar to the findings from logistic regression, the feature importance in the RF model also highlights grade as a key predictive variable for evaluating distant metastasis of gallbladder cancer. Tumor grade serves as an indicator to assess the morphological and functional similarity between tumor cells and source organ tissues [25].

Previous studies have also found that grade is a significant predictor of distant metastasis and prognosis in gallbladder cancer patients [6,7,20]. The higher the grade, the poorer the cell differentiation; higher grades are typically associated with greater invasiveness, a broader range of infiltration, and an increased likelihood of DM [20].

Studies have shown that poorly differentiated GBCs are more likely to experience distant metastasis [26], a finding consistent with the conclusion of this study. Lymph node status is a widely recognized predictive factor for evaluating metastasis and prognosis in gastrointestinal malignant tumors [27,28], and a thorough assessment of lymph nodes is also essential for patient treatment [29,30]. This study identified the N stage as an important factor in predicting the occurrence of DM in GBC. This study found that gallbladder cancer patients with tumors 2 cm or larger are more prone to DM, aligning with previous research findings [6].

ML enables computers to emulate human learning capabilities, enhancing performance through iterative data analysis model improvements [31]. Over the past decade, ML algorithms have been extensively applied in the medical field, yielding remarkable results in the diagnosis, treatment, and prognosis of diseases [32]. Compared with traditional data analysis methods, ML offers several key advantages. It can efficiently process large datasets and, more efficiently, handle nonlinear data through diverse algorithms and statistical models, while traditional methods often fall short of achieving satisfactory outcomes with nonlinear data. Numerous studies have demonstrated that the predictive performance of machine learning is superior to traditional methods [13]. In this study, the RF model emerged as one of the most effective ML algorithms. RF employs advanced classification decisions and different weighting ratios, excelling in handling large feature sets and highly nonlinear data. This model enhances the utilization of analytical information, resulting in superior predictive performance [12].

We applied seven predictive models using the SEER to evaluate the distant metastasis in T1 and T2 gallbladder cancer patients. These models were evaluated by accuracy, precision, recall, F1 score, and AUC value. Amongst them, the RF model demonstrated the

highest predictive ability (AUC = 0.913, F1 score = 0.836), making it the optimal choice for predicting distant metastasis in gallbladder cancer.

This study also has some limitations: (1) As it relies on North American demographic data, it needs to be validated with external populations in future studies. (2) The performance of the RF algorithm requires further enhancement by incorporating additional risk factors, as well as tumor markers, which may also be important predictive factors for distant cancer metastasis. In response to the above issues, we will gather more comprehensive data and conduct in-depth supplementary studies in future research.

5. Conclusions

The efficacy of this model is expected to be further enhanced through the incorporation of additional risk factors, thereby elevating its performance to a higher level and quantitative indicators to forecast DM of T1 and T2 GBC. Among the seven predictive models, the RF algorithm demonstrated superior predictive power. Offer individualized predictive outcomes to enhance patient results. We have applied an online network calculator to assess the risk of DM, which applies to clinical use (<http://121.43.117.60:8004/>). This tool serves as a valuable resource for evaluating the likelihood of DM in these patient populations.

Author Contributions: Conceptualization, Z.G. and Z.Z.; methodology, Z.G. and Z.Z.; software, Z.G.; validation, L.L. and Y.Z.; investigation, Z.L. and C.Z.; resources, J.F., H.Y. and P.Y.; data curation, Z.G. and H.Q.; 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 the 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: Not applicable.

Informed Consent Statement: We did not require informed consent from relevant patients for the use of the SEER database for research purposes, as it contains public data. Ethical approval was also not necessary. 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. This data can be accessed on 1 September 2024. (<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. *J. Cancer Res. Clin. Oncol.* **2023**, *149*, 12297–12313. [CrossRef]
2. Huang, J.; Patel, H.K.; Boakye, D.; Chandrasekar, V.T.; Koulaouzidis, A.; Lucero-Prisno, D.E., III; 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 Lett.* **2021**, *521*, 238–251. [CrossRef]
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? *Int. J. Surg.* **2020**, *82*, 82–86. [CrossRef]
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. *J. Clin. Gastroenterol.* **2013**, *47*, 443–448. [CrossRef]
5. Sharma, A.; Sharma, K.L.; Gupta, A.; Yadav, A.; Kumar, A. Gallbladder cancer epidemiology, pathogenesis and molecular genetics: Recent update. *World J. Gastroenterol.* **2017**, *23*, 3978–3998. [CrossRef]
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. *Int. J. Med. Sci.* **2020**, *17*, 1704–1712. [CrossRef]
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 Surg.* **2021**, *21*, 59. [CrossRef]
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 Med.* **2018**, *7*, 5006–5014. [CrossRef]

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* **2020**, *22*, 1490–1495. [\[CrossRef\]](#)
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. [\[CrossRef\]](#)
11. Zhong, X.; Lin, Y.; Zhang, W.; Bi, Q. Predicting diagnosis and survival of bone metastasis in breast cancer using machine learning. *Sci. Rep.* **2023**, *13*, 18301. [\[CrossRef\]](#)
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 Med.* **2021**, *10*, 2802–2811. [\[CrossRef\]](#)
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. *Clin. Endocrinol.* **2023**, *98*, 98–109. [\[CrossRef\]](#)
14. Wernberg, J.A.; Lucarelli, D.D. Gallbladder cancer. *Surg. Clin. N. Am.* **2014**, *94*, 343–360. [\[CrossRef\]](#)
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 Int.* **2019**, *19*, 169. [\[CrossRef\]](#)
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. *J. Exp. Clin. Cancer Res. CR* **2019**, *38*, 247. [\[CrossRef\]](#)
17. Hundal, R.; Shaffer, E.A. Gallbladder cancer: Epidemiology and outcome. *Clin. Epidemiol.* **2014**, *6*, 99–109. [\[CrossRef\]](#)
18. Shindoh, J.; de Aretxabala, X.; Aloia, T.A.; Roa, J.C.; Roa, I.; Zimmiti, G.; Jayle, 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. *Ann. Surg.* **2015**, *261*, 733–739. [\[CrossRef\]](#)
19. Zhang, W.; Chen, Z.; Sa, B. Construction and validation of the predictive model for gallbladder cancer liver metastasis patients: A SEER-based study. *Eur. J. Gastroenterol. Hepatol.* **2024**, *36*, 129–134. [\[CrossRef\]](#)
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. *Front. Surg.* **2022**, *9*, 899896. [\[CrossRef\]](#)
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. *J. Surg. Res.* **2022**, *275*, 181–193. [\[CrossRef\]](#)
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. *Int. J. Endocrinol.* **2023**, *2023*, 9965578. [\[CrossRef\]](#)
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 Int.* **2022**, *22*, 28. [\[CrossRef\]](#)
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. *Front. Oncol.* **2021**, *11*, 614398. [\[CrossRef\]](#)
25. Osório, F.M.; Vidigal, P.V.; Ferrari, T.C.; Lima, A.S.; Lauer, G.M.; Couto, C.A. Histologic Grade and Mitotic Index as Predictors of Microvascular Invasion in Hepatocellular Carcinoma. *Exp. Clin. Transplant.* **2015**, *13*, 421–425. [\[CrossRef\]](#)
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* **2011**, *13*, 463–472. [\[CrossRef\]](#)
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 J. Surg. Oncol.* **2012**, *10*, 87. [\[CrossRef\]](#)
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. *Front. Oncol.* **2022**, *12*, 1065468. [\[CrossRef\]](#)
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. *Ann. Surg. Oncol.* **2010**, *17*, 1831–1840. [\[CrossRef\]](#)
30. Negi, S.S.; Singh, A.; Chaudhary, A. Lymph nodal involvement as prognostic factor in gallbladder cancer: Location, count or ratio? *J. Gastrointest. Surg.* **2011**, *15*, 1017–1025. [\[CrossRef\]](#)
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. *J. Clin. Transl. Hepatol.* **2023**, *11*, 1184–1191. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Bhinder, B.; Gilvary, C.; Madhukar, N.S.; Elemento, O. Artificial Intelligence in Cancer Research and Precision Medicine. *Cancer Discov.* **2021**, *11*, 900–915. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.