

Comparison the Accuracy of Hepatocellular Carcinoma Recurrence Prediction Models: LASSO Regression, Random Forest and Support Vector Machine Models

Kun Xiong*

School of Clinical Medicine, Ningxia Medical University, Yinchuan, Ningxia, 750000, China

*Corresponding author: 202302115304@nxmu.edu.cn

Abstract. The diagnostic effects of Lasso-Cox Logistic regression, random forest (RF) and support vector machine (SVM) combined with hepatocellular carcinoma were compared and analyzed. First, the data quality is improved through preliminary data preprocessing, including the establishment of binary features, categorical variable conversion, missing value processing and lifetime zero value processing. Then COX regression is used to check whether there is a high-dimensional problem in the data, and the stability of the model is improved by multicollinearity test. Then the Lasso regression model is trained to screen significant feature variables and improve the model performance. In the random forest model, the parameters are adjusted by Out-of-Bag (OBB) error rate analysis, and the optimal model is obtained. In SVM model, the optimal penalty parameter and kernel function parameter are determined by cross-validation of optimization parameters. When evaluating the prediction rate of the model, the accuracy rate is taken as the main reference index, and the Lasso-Cox regression, RF and SVM are analyzed in detail. The results show that the RF model in this study has the highest accuracy. The results suggest that RF model may be more suitable for clinical indicators to predict Hepatocellular carcinoma (HCC) recurrence. In the future, larger data sets can be utilized for validation, and the model content can be improved to further optimize and improve the prediction accuracy. This study provides a reference for the selection of HCC recurrence risk prediction model, and is expected to provide a certain degree of reliable support for clinical judgment.

Keywords: Hepatocellular carcinoma; prediction model; Lasso-Cox; random forest; support vector machine.

1. Introduction

According to the 2022 global cancer statistics, liver cancer has become the fourth leading cause of cancer death. About 70.1% of liver cancer cases occur in Asia, and the number of liver cancer cases in China exceeds 196,494, which is estimated to account for 42.4% of the total liver cancer cases in the world [1]. Hepatocellular carcinoma (HCC) is a highly heterogeneous and highly prevalent subtype of liver cancer. Many factors induce the occurrence of HCC, including hepatitis C virus (HCV) and hepatitis B virus (HBV) infection, as well as chronic hepatitis caused by factors such as alcohol and drug-induced liver injury, whose mechanism is to induce the occurrence of cancer through the continuous stimulation of inflammation [2].

Although computed tomography (CT) + nuclear magnetic resonance imaging (MR) Dual radiomics has been used in the clinical examination of HCC, and the diagnostic ability has been confirmed to be improved, advanced HCC has the characteristics of high recurrence rate, which greatly affects the five-year survival rate of patients after surgery [3]. More than a decade ago, researchers developed immune checkpoint inhibitor (ICI) therapy for patients with advanced HCC, which has been used in clinical treatment and has achieved certain results [4]. Studies have shown that the efficacy of ICI is closely related to tumor microenvironment (TME) [5]. The expression of some key immune cells, such as myeloid-derived suppressor cells (MDCS), CD8⁺ cytotoxic T lymphocytes (CTLs) and tumor macrophages (TMS), participate in immune escape and promote tumor growth through specific mechanisms [6-8]. By exploring the resistance mechanism of ICI and the means of regulating TME, the treatment of HCC will move towards more individualized and precise treatment, and compete for the survival time of patients with advanced HCC.

Wang et al used the Lasso-Cox regression model to predict the recurrence of early HCC patients, which was better than the C-index value of random forest (RF), but the number of samples was small, and HCC patients were affected by nonlinear factors, which led to prediction bias of the model [9]. Compared with support vector machine (SVM) model, SVM can resist overfitting and reduce the interference of irrelevant factors. According to the research of Guo and his team, the recursive feature elimination (RFE) method based on the RF model was used to screen out biomarkers. Compared with SVM, RF can realize probability estimation, which is convenient for survival risk quantification and combined with clinical practice [10]. This paper focuses on People's Daily living habits, analyzes the pathogenic factors of HCC patients, establishes a prediction model, and compares the analysis and prediction effects of Lasso-Cox, RF and SVM, so as to provide valuable reference for clinical practice.

2. Methods

2.1. Data Source

The data used in this study came from Kaggle, and the clinical dataset was collated and presented by ridgie_mo of The Cancer Genome Atlas Program (TCGA). The dataset availability score was 5.29, and a total of 368 patients and related clinical data were recorded. After processing and screening, there are 13 variables in the original data, which can be used to analyze the risk factors related to HCC pathogenesis and establish a prediction model.

2.2. Variables and Data Preprocessing

The original data set has a total of 11 variables, and the name and explanation of each variable are shown in Table 1.

Table 1. Attribute Information

Variables	Explanation
Age	Age of each individual
Gender	The gender of each individual, categorized as "male" and "female"
BMI	$\text{Weight(kg)} / (\text{height(cm)} / 100)^2$
Race_list	Individual races, categorized as "ASIAN", "WHITE", "BLACK OR AFRICAN AMERICAN" or "AMERICAN INDIAN OR ALASKA NATIVE"
Person_neoplasm_cancer_status	Whether an individual has a tumor or not, classified as "WITHTUMOR", "TUMOR FREE" or "UNKOWN".
Neoplasm_histologic_grade	Histological grading of the tumor, categorized as "G1", "G2", "G3", "G4", "UNKOWN"
Stage_event	Staging of the tumor, categorized as "5th", "6th", "7th", "UNKOWN"
Vital_status	The living state of the individual, 0 for deceased, 1 for alive
Survival_time	Actual survival time of the patient (unit: days)
Hepatitis_b	whether the patient has Hepatitis B, 1 for Hepatitis B, 0 for no Hepatitis B
Hepatitis_c	whether the patient has Hepatitis C, 1 for Hepatitis C, 0 for no Hepatitis C,
Alcohol	whether the patient drink alcohol, 1 for alcohol, 0 for no alcohol
Smoking	whether the patient smoke, 1 for smoking, 0 for no smoking

For data preprocessing, this paper uses the "mice" package in R to check for missing values. The results indicated that there were no missing values and no need to delete variables. In the case of a survival of zero, the survival with a value of zero was replaced by the median of the survival without zero. Categorical variables were transformed into factors and categories were combined with very few samples. Binary characteristics of risk factors were created and classifications were refined. BMI was used instead of height/weight to provide more intuitive clinical indicators.

2.3. Method Introduction

In this paper, three methods are selected to build the prediction model, namely Lasso-Cox regression model, random forest model (RF) and vector machine model (SVM). For all models, the

80:20 ratio is used to divide the training set and the validation set. In order to test the accuracy of the model's prediction of HCC recurrence rate, the receiver operating characteristic (ROC) or confusion matrix was used to calculate the accuracy of each prediction model as a reference for evaluation. The definition and calculation formula of the confusion matrix are shown in Table 2.

Table 2. Confusion Matrix

		Prediction Condition	
Actual Condition	Positive (P)	Positive (P) True Positive (TP)	Negative (N) False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TN+TP}{TP+FN+FP+TN} \quad (1)$$

A simple regression analysis was performed on the data to detect whether the variables had high dimensions, and the variance inflation factor (VIF) was used to test multicollinearity. Training with Lasso-Cox with a training set, screening highly correlated variables, and then making predictions and accuracy assessments on the test set to produce results.

In RF model, the optimal parameters are selected through cross-validation to improve the prediction ability of RF model. The RF model is used to predict the test set, and the accuracy values are evaluated.

The SVM model also uses dimensionality reduction data, adopts radial basis function (RBF) and cross-validation methods to select the optimal parameters, runs the SVM model on the test set for prediction, and evaluates according to the accuracy. Finally, the accuracy of Lasso-Cox model, RF model and SVM model is compared.

3. Results and Discussion

3.1. Data Distribution

As can be seen from the frequency histogram of the binary variables in Figure 1, most patients were still alive before data collection, while a small number of patients were recorded as dead, accounting for about 1/3 of the total data. This variable indicates that HCC is a cancer with high mortality and large number of survivors, which has certain research value in predicting the occurrence and recurrence of this disease.

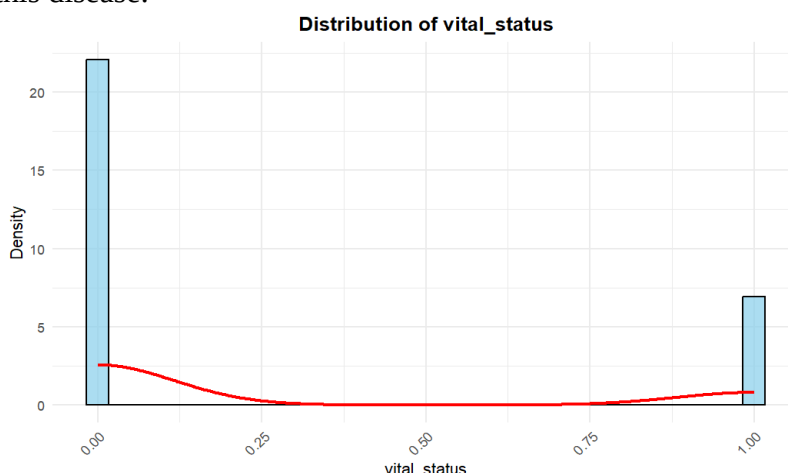


Fig. 1 Histogram of vital status

As shown in Figures 2 and 3, in the multigrade frequency histogram, the majority of patients had moderately differentiated tumor grade (G2) and poorly differentiated tumor grade (G3), which had a significant impact on survival events. For the "survival time" variable, most patients live no more than three years. These results indicate that the survival rate and prognosis of HCC patients are

relatively poor. The research on the prediction of tumor occurrence and recurrence is of public health value.

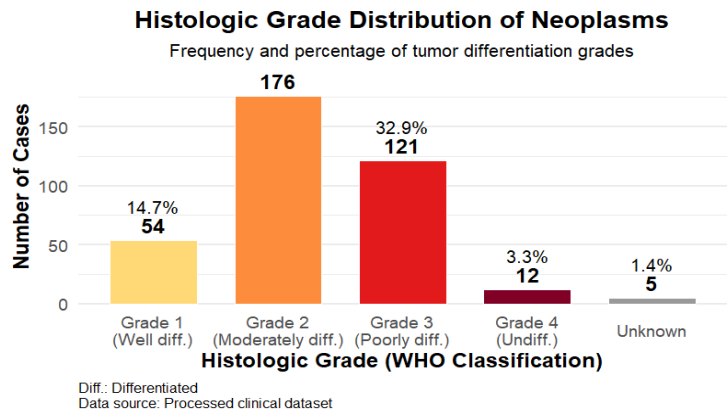


Fig. 2 Histologic Grade Distribution of Neoplasms

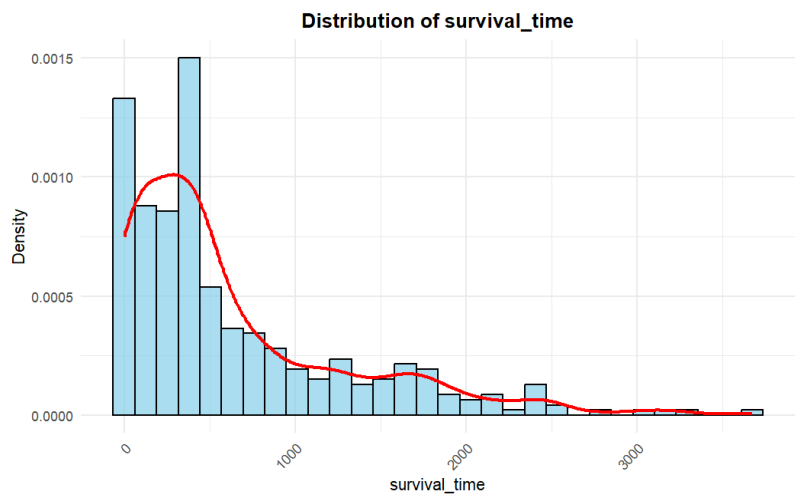


Fig. 3 Histogram of Survival time distribution

Figure 4 shows a pie chart of partial dichotomous variables, showing the distribution of patients across different risk factors, among the three variables of smoking, hepatitis B and hepatitis C, the number of people without this risk factor is higher than the number of people with this risk factor, and the occurrence of hepatitis B is the highest among all the variables, indicating that this factor has a high relationship with the disease among all the relevant factors.

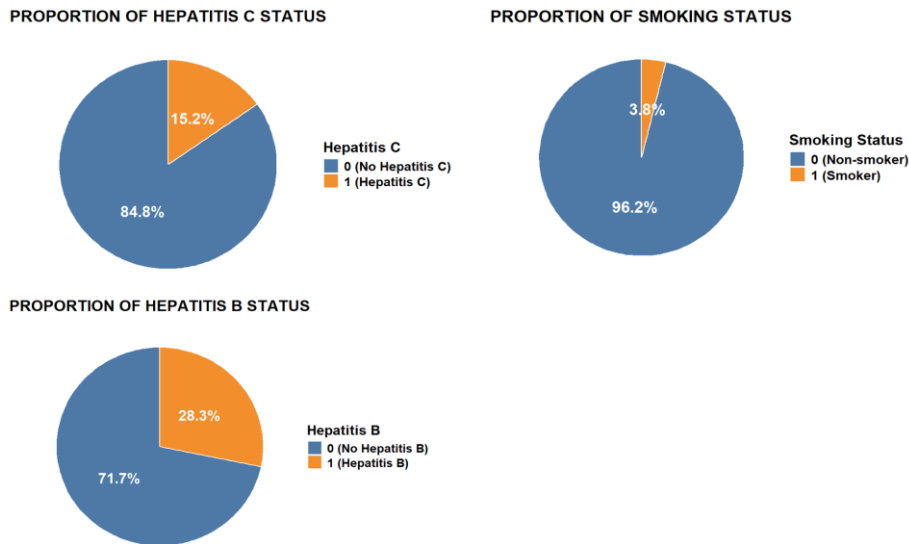


Fig. 4 Pie charts of smoking, hepatitis B and hepatitis C

As shown in Figure 5, among multiple categorical variables, the age distribution is close to normal, and it can be seen that the age range of the data set is about 16-88 years old, and the patients are mainly concentrated in the 35-80 years old age group. HCC occurs most frequently in patients over 50 years of age, suggesting that middle-aged and elderly people should maintain a healthier lifestyle, which is conducive to disease prevention and longevity.

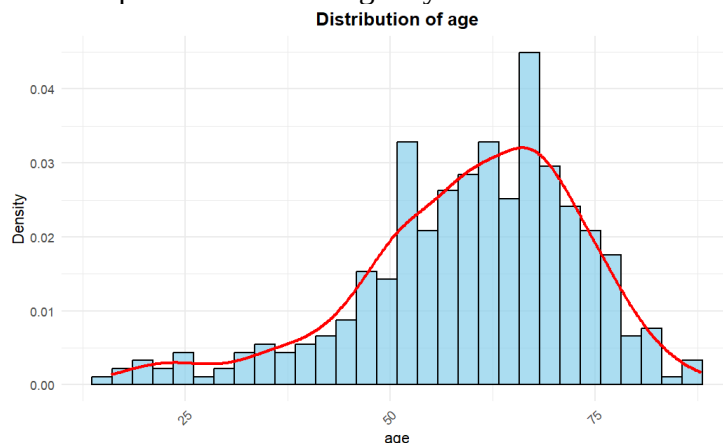


Fig. 5 Histogram of age distribution

As shown in Figures 6 and 7, the variable BMI is also close to a normal distribution. HCC is most commonly seen in overweight patients, according to BMI standards published by the World Health Organization. These data show that obesity is associated with various diseases, the proportion of healthy weight is very small, and maintaining a good weight can reduce the possibility of disease. In this analysis, the author classified Native Americans and blacks as "other," with a more balanced distribution of cases among Whites and Asians.

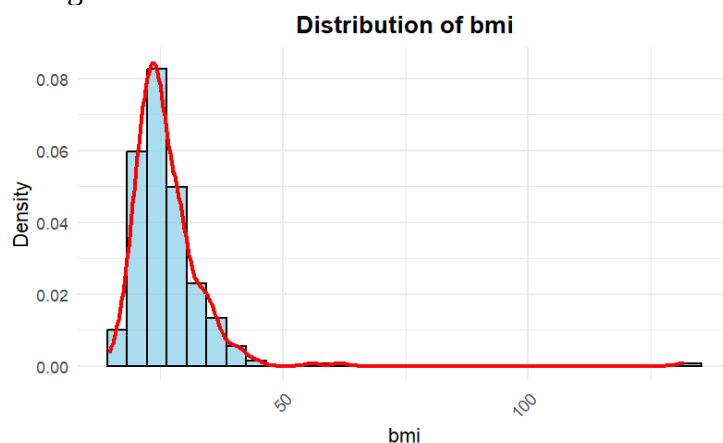


Fig. 6 Histogram of BMI

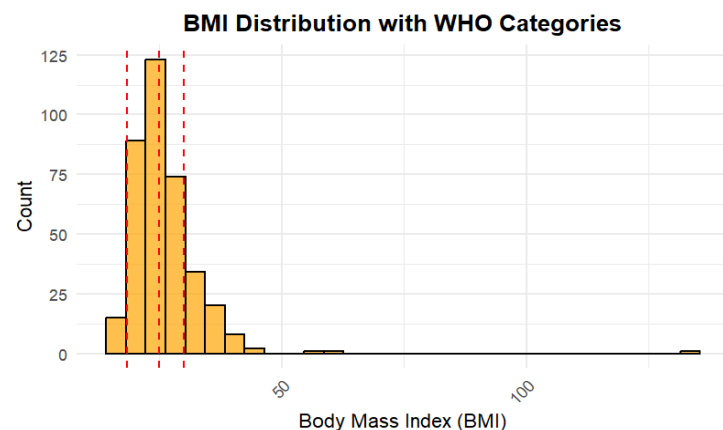


Fig. 7 Histogram of BMI distribution with WHO Categories

3.2. Lasso-Cox Regression

Before model training for Lasso-Cox regression, a simple Cox regression analysis was performed on the data set. It was found that "gender" and "bmi" showed $P > 0.05$, and there was mild collinearity, the generalized variance inflation factor is 1.2 ($G\text{VIF} > 1.2$). These results all reflect the high dimensional problem of the variables, and the multicollinearity between the independent variables is checked by the square inflation factor (VIF). As shown in Table 3, although the VIF values among independent variables were all less than 5, the value of the number of death events/number of variables was about 9.7, which was slightly lower than the empirical threshold, so LASSO regression was performed to screen the set of variables with the most predictive power.

Table 3. The VIF value of the variable

Variable	VIF
Gender	1.276
BMI	1.337
Race_list	1.980
Person_neoplasm_cancer_status	1.216
Neoplasm_histologic_grade	1.210
Stage_event	1.224
Hepatitis_b	1.527
Hepatitis_c	1.155
Alcohol	1.256
Smoking	1.119

In order to make the results repeatable and improve the accuracy of the model, the random seed was set as "123" during LASSO regression. The model is trained using the "glmnet" package in R for the training set. Table 4 shows the results after training, and the "Coefficient" column indicates the coefficient of this feature.

Table 4. Regression Coefficient

Variable	Coefficient
race_list WHITE	1.359
race_listBLACK OR AFRICAN AMERICAN	0.514
stage_event6th	-0.678
stage_event7th	-1.782
neoplasm_histologic_gradeG2	0.668
Neoplasm_histologic_grade G3	1.397
Neoplasm_histologic_gradeG4	1.114
hepatitis_b(1)	0.530

When the 10-fold cross-validation method is applied to iterative analysis, a model with better predictive performance but more variables is obtained when λ is 0.0105, as shown in Figure 8. The selected variables included white race, tumor histological grade, tumor stage and hepatitis B.

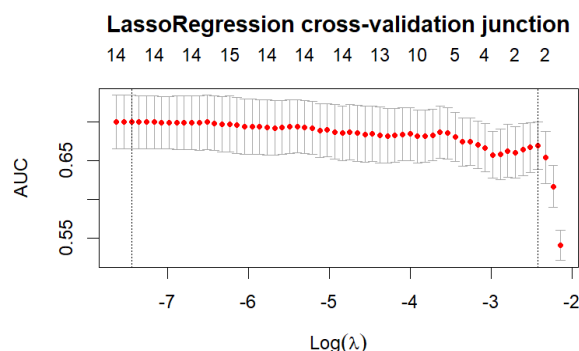


Fig. 8 The process of selecting the best value of parameter λ in Lasso regression model by cross-validation method

The established prediction model was verified, and the negative prediction accuracy (NPV) was 0.740. The confusion matrix obtained was shown in Figure 9, and the calculated accuracy was 0.691.

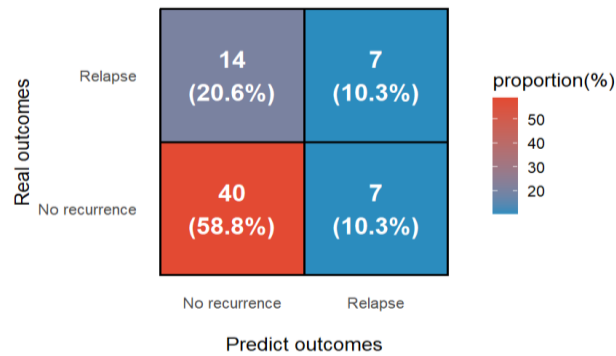


Fig. 9 Four-quadrant analysis of hepatocellular carcinoma recurrence prediction

As can be seen from the confusion matrix, among 68 patients, 7 patients were predicted to relapse, but actually did not relapse; The number of patients correctly predicted to be recurrence-free was 40, the number of patients correctly predicted to be recurrence-free was 7, and 14 patients were predicted to be recurrence-free but actually relapsed.

3.3. Random Forest Model

The random forest model was implemented to predict tumor recurrence, and the relevant clinical data was used to predict tumor recurrence. The random seed is set to 123 to ensure repeatability. The training set and the test set are divided into 80:20. Through cross-validation, the best parameter of "ntree" is determined to be 100. The error rate stabilized at approximately 30% after 200 trees, showing good convergence, as shown in Figure 10.

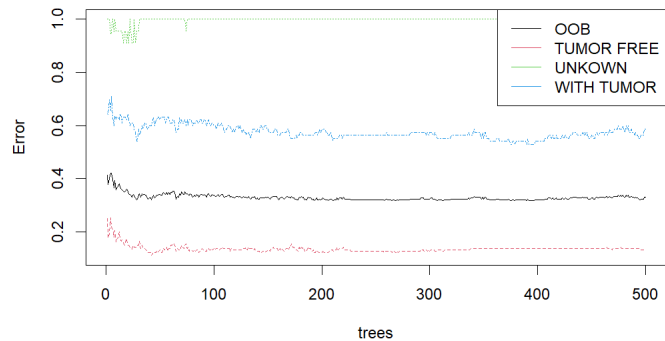


Fig. 10 Out-of-Bag (OBB) Error Rate

The established model was verified by prediction on the test set. The confusion matrix obtained was shown in Figure 11, and the overall accuracy of the calculation was about 0.736.

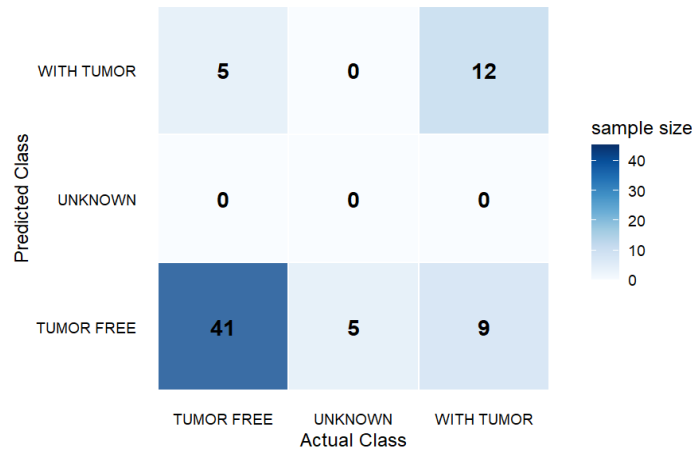


Fig. 11 Confusion matrix for the random forest model

It can be seen from the confusion matrix that 9 of the 78 patients did not actually have tumor, but were predicted to have tumor. The number of patients correctly predicted to be tumor-free was 41, and the number of patients correctly predicted to relapse was 12.

3.4. Support Vector Machines

In the process of implementing the support vector machine (SVM) model, in order to find the model with the best prediction performance, the Radial Basis Function (RBF) is used, and the cross-validation method is used to optimize the prediction accuracy of the model. The parameters were adjusted for the training set, and the optimal penalty parameter (C=100) and kernel function parameter ($\sigma=0.001$) were determined.

The SVM model was run to predict the data set, and the confusion matrix was shown in Figure 12. The accuracy of the model was 0.726, and the Recall rate was 0.8462.

		Actual Class	
		0	1
Predicted Class	0	44 (60.3%)	12 (16.4%)
	1	8 (11.0%)	9 (12.3%)

Fig. 12 Confusion matrix of the SVM model

From the confusion matrix, it can be seen that of the 68 patients, 12 participants did not actually relapse but were predicted to relapse. There were 44 cases without recurrence and 9 cases with recurrence.

4. Conclusion

To sum up, in this paper, a clinical dataset of HCC was adopted. With the accuracy of the model as a reference, the performance of Lasso-Cox, RF and SVM models in predicting HCC recurrence was compared. This paper found that the RF model has the best prediction accuracy on the test set, and this dataset is more suitable for predictive analysis using RF. However, it should be noted that the predictive ability of the model is influenced by many factors, including sample sources, data preprocessing, feature selection, and the choice of model parameters, etc. Future research can further refine the model and conduct tests on more comprehensive datasets to enhance the predictive ability of diseases and provide patients with more precise and personalized treatments.

References

[1] Bray F, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 2024, 74(3): 229-263.

[2] Singal A G, El-Serag H B. Hepatocellular Carcinoma From Epidemiology to Prevention: Translating Knowledge into Practice. Clinical gastroenterology and hepatology: the official clinical practice journal of the American Gastroenterological Association, 2015, 13(12): 2140-2151.

[3] Ma Y, et al. Research on multi-model imaging machine learning for distinguishing early hepatocellular carcinoma. BMC Cancer, 2024, 24(1): 363.

[4] Sharma P, et al. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. Cell, 2017, 168(4): 707-723.

- [5] Oura K, et al. Tumor Immune Microenvironment and Immunosuppressive Therapy in Hepatocellular Carcinoma: A Review. *International Journal of Molecular Sciences*, 2021, 22(11): 5801.
- [6] Van Vlerken-Ysla L, et al. Functional states of myeloid cells in cancer. *Cancer Cell*, 2023, 41(3): 490-504.
- [7] Chiu D K C, et al. Hypoxia induces myeloid-derived suppressor cell recruitment to hepatocellular carcinoma through chemokine (C-C motif) ligand 26. *Hepatology*, 2016, 64(3): 797-813.
- [8] Kersten K, et al. Spatiotemporal co-dependency between macrophages and exhausted CD8⁺ T cells in cancer. *Cancer Cell*, 2022, 40(6): 624-638.e9.
- [9] Wang Q, et al. Nomogram established on account of Lasso-Cox regression for predicting recurrence in patients with early-stage hepatocellular carcinoma. *Frontiers in Immunology*, 2022, 13: 1019638.
- [10] Guo L, et al. Random-forest algorithm based biomarkers in predicting prognosis in the patients with hepatocellular carcinoma. *Cancer Cell International*, 2020, 20: 251.