Machine Learning–Based Identification of Depressive Symptoms in Patients with Cancer: Multidimensional Feature Analysis and Model Development

Shuai Wei, Xingcai Gao*
The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
**Corresponding Author*

Abstract: This study sought to determine the factors linked to depressive symptoms among cancer patients and to establish a machine learning model for identifying those at high risk of such symptoms, with interpretability supported by SHAP. The dataset derived from the 2018 wave of the China Health and Retirement Longitudinal Study (CHARLS) was used, and a total of 466 middle-aged and elderly cancer patients were enrolled. Feature selection was conducted via LASSO regression, and the data were randomly partitioned into training (60%) and test (40%) subsets. Eight machine learning models were built and compared. Their performance was evaluated using the ROC curves, PR curves, Brier score, and decision curve analysis, while SHAP was applied to interpret the model outputs. Seven key features were identified. Among all models, XGBoost achieved the optimal overall performance in the test cohort, presenting an AUC of 0.771, PR-AUC of 0.790, Brier score of 0.1596, and favorable net clinical benefit. SHAP analysis indicated that self-rated health, life satisfaction, and IADL limitation were the most important contributors to model predictions. Subgroup analyses showed stable performance across age and sex strata, with all AUCs above 0.70. Overall, the XGBoost model demonstrated good discrimination and interpretability, suggesting its potential as an auxiliary tool for early warning, further screening, and risk stratification of depressive symptoms in cancer patients.

Keywords: Patients with Cancer; Symptoms of Depression; Machine Learning; XGBoost; SHAP

1. Introduction

With advances in cancer diagnosis and treatment, survival among cancer patients has improved, and increasing attention has been paid to their quality of life and mental health [1]. Depression is one of the most common psychiatric comorbidities in patients with cancer. It not only significantly affects patients' subjective well-being and quality of life, but also can lower adherence to anticancer treatment, prolong hospitalization, and increase the risk of adverse outcomes [2,3]. Therefore, early identification of patients with cancer who are at high risk of depression, followed by timely and targeted psychological assessment and intervention, is of great clinical significance.

At present, depressive symptoms in cancer patients are primarily identified using self-report instruments, such as the Center for Epidemiologic Studies Depression Scale (CES-D). These tools play an important role in psychological assessment and are widely used for the clinical screening of depressive symptoms. However, in real-world clinical settings—especially when outpatient volume is high, follow-up workloads are heavy, or patients have substantial cognitive and physical burdens—routine screening using only rating scales may still be limited by factors such as time costs, patient cooperation, and implementation conditions. Therefore, constructing an auxiliary high-risk identification model based on routinely available information, such as sociodemographic characteristics, health status, and functional status, may provide a useful supplement for the early warning and further screening of depressive symptoms in cancer patients.

Previous studies investigating factors related to depressive symptoms among cancer patients have mainly employed traditional statistical approaches, such as generalized linear regression [4]. These methods are advantageous in identifying independent correlates; however, when faced with multidimensional and heterogeneous real-world data, they still have

certain limitations in handling multicollinearity, complex nonlinear relationships, and high-order interactions, which may affect their ability to characterize individual differences. In recent years, machine learning (ML) methods have shown substantial potential in the analysis of high-dimensional and nonlinear medical data, and have gradually been applied in prediction studies of various chronic diseases and clinical outcomes[5, 6]. Compared with traditional statistical models, ML methods can more flexibly integrate multisource feature information, thereby providing a new analytical approach for risk identification in complex clinical problems.

However, although complex ML models may achieve better predictive performance, they are commonly criticized for their limited transparency and interpretability, which to some extent restricts their clinical application. SHapley Additive exPlanations (SHAP) provides an effective tool for improving model transparency, enabling the estimation of each feature's impact on model predictions and the explanation of prediction outcomes at both group and individual levels. Therefore, combining machine learning modeling with interpretability analysis may help improve identification performance while enhancing the clinical interpretability of the results.

Based on this, the present study draws on data from the 2018 China Health and Retirement Longitudinal Study (CHARLS) to compare the performance of several mainstream ML algorithms in identifying high risk of depression among patients with cancer, and to construct and validate the best-performing model. Meanwhile, through the integration of least absolute shrinkage and selection operator (LASSO) feature selection and the SHAP interpretability framework, this study seeks to explore the core factors related to depressive symptoms in cancer patients and analyze their possible nonlinear associations. The findings may provide a basis for early warning, further screening, and risk stratification of psychological problems in patients with cancer.

2. Materials and Methods

2.1 Data Collection and Processing

The data analyzed in this study were obtained from the 2018 fourth-wave nationally representative survey of the CHARLS database. The CHARLS protocol was approved by the Ethics Review Committee of Peking University, with written informed consent obtained from all participants [7].

Eligibility criteria were defined as: respondents aged ≥45 years who explicitly reported a physician-confirmed diagnosis of cancer or malignant tumor in the CHARLS questionnaire. The exclusion criteria were: (1) missing outcome data, specifically missing CESD-10 scores; and (2) severe missingness in key covariates that precluded subsequent analysis. Eligible patients with cancer were finally included in the analysis.

2.2 Disease Definition, Outcome, and Predictor Variables

According to the standard CHARLS questionnaire, respondents who self-reported that they had ever been diagnosed by a physician with "cancer or a malignant tumor (excluding minor skin cancer)" were defined as patients with cancer. Considering that disease reporting might be missed in a single survey wave, cancer status was identified by combining information across previous survey waves; respondents who reported cancer in any wave were assigned to the cancer group.

The primary endpoint of this study was depressive symptoms, measured using the CESD-10. Total scores on the CESD-10 range from 0 to 30. Individuals with a score ≥10 were defined as having depressive symptoms (coded 1), whereas those with a score <10 were categorized as non-cases (coded 0).

Candidate variables were extracted from multiple domains, including sociodemographic characteristics, socioeconomic status, lifestyle, physical functioning, subjective health perception, chronic disease burden, and cancer treatment history. These variables included age, sex, marital status, educational level, place of residence, health insurance, household size, per capita household consumption expenditure, out-of-pocket inpatient medical expenses, body mass index (BMI), smoking history, drinking history, daily sleep duration, physical exercise, participation in social activities, childhood health status, guardian history of addiction, limitations in instrumental activities of daily living (IADL), limitations in activities of daily living (ADL), self-rated health, life satisfaction, life expectancy perception, number of comorbidities, and history of cancer treatment (surgery, chemotherapy, radiotherapy, and

anticancer medication).

Among these, the number of comorbidities was calculated based on the coexistence of 13 chronic conditions: dyslipidemia, diabetes, heart disease, stroke, asthma, chronic lung disease, kidney disease, liver disease, hypertension, stomach or digestive disease, arthritis or rheumatism, memory-related disease, and psychological or psychiatric problems. The variable was categorized into 0, 1, 2, and ≥3 conditions. To ensure consistent variable direction in the model, ordinal variables such as life satisfaction, self-rated health, life expectancy perception, and number of comorbidities were uniformly recoded so that increasing values represented worsening status or higher risk.

2.3 Feature Selection

Extreme values of continuous variables were handled using the Winsorization method. The study population was then randomly split into training and test sets at a 6:4 ratio. The training set was used for data preprocessing, feature selection, and model development, whereas the test set served as an independent validation cohort.

For missing covariate data, multiple imputation was implemented in the training set via the mice package, with predictive mean matching (PMM) as the imputation method, generating five imputed datasets for subsequent analyses. The test set was reserved solely for independent validation and was not involved in feature selection or hyperparameter tuning.

LASSO regression was performed in the training subset to screen key features. A binomial logistic regression model was constructed with the glmnet package, and the optimal penalty parameter λ was identified via 10-fold cross-validation. Variables corresponding to lambda.min were chosen as the input features for subsequent model development.

2.4 ML Model Development and Evaluation

Based on the features identified via LASSO regression, eight ML models were developed in the training set, including a generalized linear model (GLM), extreme gradient boosting (XGBoost), support vector machine (SVM), gradient boosting machine (GBM), random forest (RF), adaptive boosting (AdaBoost), neural network (NNET), and K-nearest neighbor (KNN). All models were developed within the caret framework, and hyperparameter tuning was performed using repeated 10-fold cross-validation with five repetitions, with the area under the receiver operating characteristic curve (AUC) used as the optimization criterion. The optimal hyperparameter settings for all models, including XGBoost, were determined based on predefined parameter grids and are presented in Table 1.

Table 1. Parameter Settings and Tuning Ranges of Machine Learning Models

Model	Key Parameter Settings / Tuning Range
GLM	No tuning grid set; default parameters adopted
XGBoost	nrounds = 10; eta = 0.001; max_depth = 3; gamma = 0.5; subsample = 0.6; colsample_bytree = 0.5; min_child_weight = 1;
SVM	C = 0.09; sigma = 0.001
GBM	n.trees = 100; shrinkage = 0.1; interaction.depth = 3; n.minobsinnode = 5
NNET	decay = 0.6; size = 6
RF	numRandomCuts = 3; mtry = 11
AdaBoost	mfinal = 2; maxdepth = 2; coeflearn = Zhu
KNN	distance = 1; kmax = 12; kernel = optimal

Model performance was assessed in both the training and independent test sets, and the final model was selected primarily based on performance in the independent test set. Discriminative performance was assessed via the receiver operating characteristic (ROC) curve and AUC. Furthermore, the precision-recall (PR) curve and area under the PR curve (PR-AUC) were adopted to assess the model's capacity for detecting positive cases. Prediction error was measured with the Brier score, where lower values signified superior agreement between predicted probabilities and observed outcomes. Calibration performance was presented through calibration curves. Clinical usefulness was further assessed by decision curve analysis (DCA), which assessed the net benefit of each model over a spectrum of threshold probabilities. The model with the best overall performance was selected on the basis of these

comprehensive evaluation metrics.

2.5 Model Interpretability and Subgroup Analysis

To enhance interpretability, the optimal model was further explained using SHAP. SHAP values were computed with the fastshap package, and the shapviz package was used to produce feature importance, beeswarm, dependence, and waterfall plots for visualizing the global and local effects of features on model output.

To assess model stability across different populations, subgroup analyses were further conducted according to age tertiles and sex. For each subgroup, the AUC and its 95% confidence interval were calculated to assess the discriminative performance of the model across demographic strata.

2.6 Statistical Analysis

Continuous variables are presented as mean ± SD and compared via independent-samples t-tests. Categorical data are expressed as frequencies (%) and analyzed using Chi-square or Fisher's exact tests. Statistical significance is defined as a two-sided $P < 0.05$. All analyses and modeling were performed in R (v4.4.3).

3. Results

3.1 Study Population Characteristics and Development of the Modeling Dataset

This study utilized the standardized CHARLS 2018 dataset, with an initial sample size of 25,586 participants. After screening, 466 cancer patients were enrolled in the final analysis, comprising 230 with depressive symptoms and 236 without depressive symptoms. The detailed sample flowchart is illustrated in Figure 1.

Table 2 presents the baseline characteristics. Relative to the non-depressive symptom group, patients with depressive symptoms showed higher out-of-pocket inpatient medical expenses and shorter sleep duration per day (both $P \leq 0.001$). In terms of sociodemographic and lifestyle characteristics, the depressive symptom group had higher proportions of illiteracy, rural residence, and smoking, as well as a lower proportion of participation in physical exercise (all $P < 0.05$). Regarding physical functioning and health status, the depressive symptom group had higher proportions of ADL/IADL limitations, a heavier comorbidity burden, and poorer life satisfaction, self-rated health, and expected longevity (all $P < 0.001$). No significant between-group differences were observed for the remaining characteristics, including age, sex, household size, and cancer treatment history (surgery, chemotherapy, radiotherapy, and long-term medication) (all $P > 0.05$).

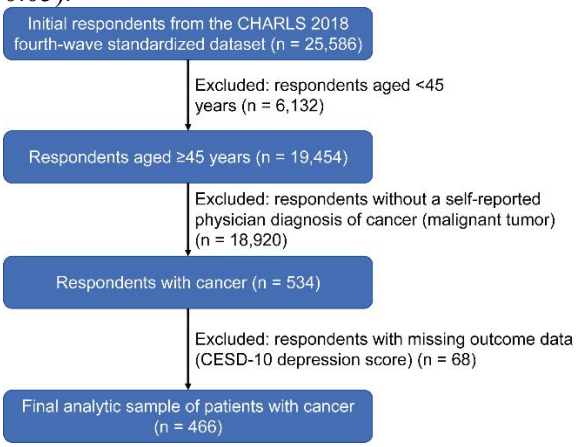

Figure 1. Flowchart of Study Population Inclusion

Table 2. Baseline Characteristics

Characteristic	Total (n=466)	Without depressive symptoms (n=236)	With depressive symptoms (n=230)	P-value
Continuous variables: Mean (SD)				
Age, years	62.5 (9.2)	62.9 (9.5)	62.2 (8.8)	0.382
Annual household consumption expenditure	23902.4 (27386.2)	22494.0 (24195.6)	25296.5 (30212.8)	0.313
Out-of-pocket inpatient medical expenses	8106.0 (23772.4)	4571.2 (13279.7)	11766.5 (30722.2)	0.001
Household size	2.8 (1.5)	2.9 (1.5)	2.6 (1.5)	0.087
Daily sleep duration, hours	5.7 (2.0)	6.1 (1.7)	5.2 (2.1)	<0.001
Body mass index (BMI)	24.6 (8.0)	24.4 (4.1)	24.8 (10.5)	0.649
Categorical variables: n (%)				
Female	309 (66.3)	148 (62.7)	161 (70.0)	0.117
Educational level				0.002
Illiterate	197 (42.3)	83 (35.2)	114 (49.6)	
Primary school	124 (26.6)	65 (27.5)	59 (25.7)	
Junior high school	96 (20.6)	53 (22.5)	43 (18.7)	
Senior high school or above	49 (10.5)	35 (14.8)	14 (6.1)	
Married	394 (84.5)	206 (87.3)	188 (81.7)	0.126

Rural residence	254 (54.5)	117 (49.6)	137 (59.6)	0.038
Basic medical insurance (yes)	449 (96.4)	230 (97.5)	219 (95.2)	0.297
Smoking history (yes)	73 (15.7)	28 (11.9)	45 (19.6)	0.031
Drinking history (yes)	110 (23.6)	58 (24.6)	52 (22.6)	0.696
Physical exercise (yes)	413 (88.6)	217 (91.9)	196 (85.2)	0.032
ADL limitation (yes)	133 (28.5)	39 (16.5)	94 (40.9)	<0.001
IADL limitation (yes)	180 (38.6)	67 (28.4)	113 (49.1)	<0.001
Social activity participation (yes)	209 (44.8)	110 (46.6)	99 (43.0)	0.496
Life satisfaction				<0.001
Fair	230 (49.4)	120 (50.8)	110 (47.8)	
Good	152 (32.6)	105 (44.5)	47 (20.4)	
Poor	84 (18.0)	11 (4.7)	73 (31.7)	
Self-rated health status				<0.001
Very good	24 (5.2)	22 (9.3)	2 (0.9)	
Good	29 (6.2)	22 (9.3)	7 (3.0)	
Fair	170 (36.5)	104 (44.1)	66 (28.7)	
Poor	170 (36.5)	72 (30.5)	98 (42.6)	
Very poor	73 (15.7)	16 (6.8)	57 (24.8)	
Life expectancy perception				<0.001
Almost impossible	113 (27.8)	34 (17.0)	79 (38.3)	
Unlikely	93 (22.9)	39 (19.5)	54 (26.2)	
Uncertain	128 (31.5)	75 (37.5)	53 (25.7)	
Likely	21 (5.2)	16 (8.0)	5 (2.4)	
Almost certain	51 (12.6)	36 (18.0)	15 (7.3)	
Childhood health status				0.259
Excellent	59 (12.8)	36 (15.5)	23 (10.1)	
Very good	204 (44.3)	103 (44.2)	101 (44.3)	
Good	84 (18.2)	44 (18.9)	40 (17.5)	
Fair	85 (18.4)	39 (16.7)	46 (20.2)	
Poor	29 (6.3)	11 (4.7)	18 (7.9)	
Guardian addictive behaviors (yes)	24 (5.9)	11 (5.4)	13 (6.4)	0.804
Number of comorbidities				<0.001
None	50 (10.7)	30 (12.7)	20 (8.7)	
1	93 (20.0)	59 (25.0)	34 (14.8)	
2	92 (19.7)	53 (22.5)	39 (17.0)	
≥3	231 (49.6)	94 (39.8)	137 (59.6)	
Cancer surgery history (yes)	213 (45.7)	108 (45.8)	105 (45.7)	1
Cancer chemotherapy history (yes)	109 (23.4)	47 (19.9)	62 (27.0)	0.092
Cancer radiotherapy history (yes)	61 (13.1)	25 (10.6)	36 (15.7)	0.138
Long-term cancer medication history (yes)	259 (55.6)	128 (54.2)	131 (57.0)	0.619

3.2 Results of Feature Selection

The sample was randomly split into a training set and a test set in a 6:4 ratio, including 280 cases in the training set and 186 cases in the test set. The training set was used for feature selection and model development, whereas the test set was used for model performance validation. LASSO regression was performed in the training set (n = 280) for feature selection, and ultimately 7 features linked to the detection of depressive symptoms in cancer patients were selected (Figure 2). These included 3 continuous variables—age, out-of-pocket inpatient medical expenses, and daily sleep duration—and 4 categorical variables: self-rated health, life satisfaction, IADL limitation, and childhood health status. Figure 2A shows the coefficient paths of each variable across varying values of the regularization parameter λ, and Figure 2B shows the cross-validation error curve.

3.3 Construction and Performance Assessment of the Machine Learning Models

Using the seven selected features, this study established and compared eight machine learning models: GLM, SVM, XGBoost, GBM, RF, NNET, KNN, and AdaBoost. Their main performance indicators in the training and test sets are presented in Table 3.

Among all models in the training set, Random Forest attained the maximum AUC of 0.988 (Figure 3A); however, its AUC decreased to 0.763 in the test set (Figure 3B). The KNN and GBM models also performed relatively well in the training set, but their performance declined in the test set. Conversely, the XGBoost model demonstrated relatively stable performance

across both the training and test sets, achieving an AUC of 0.783 in the training set and 0.771 in the test set, and showing the best overall performance among all models. In the test set, XGBoost exhibited a sensitivity of 0.754 and a specificity of 0.753, respectively. By comparison, the Neural Network and SVM models had relatively high specificity but lower sensitivity, with sensitivities of 0.665 and 0.646, respectively. According to the precision–recall curve analysis, the XGBoost model achieved PR-AUC values of 0.787 in the training set and 0.790 in the test set (Figures 3C and 3D).

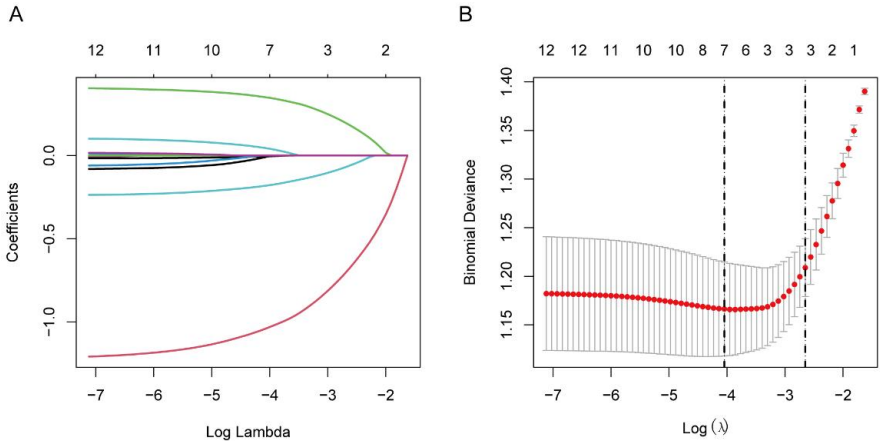

Figure 2. LASSO Regression–Based Feature Selection Results
(A) Coefficient Trajectories of Candidate Variables in the LASSO Regression Model; (B) Cross-Validation Error Curve for Determining the Optimal Penalty Parameter Λ By 10-Fold Cross-Validation.

Table 3. Performance Comparison of Models

Model	AUC (Training)	AUC (Test)	Sensitivity	Specificity	Brier (Training)	Brier (Test)
GLM	0.799	0.749	0.627	0.867	0.1807	0.2054
SVM	0.797	0.764	0.646	0.861	0.2193	0.2381
GBM	0.879	0.734	0.702	0.904	0.1432	0.2129
XGBoost	0.783	0.771	0.754	0.753	0.1493	0.1596
RF	0.988	0.763	0.975	1	0.0122	0.2093
KNN	0.898	0.73	0.826	0.813	0.1395	0.2119
NNET	0.805	0.764	0.665	0.849	0.1797	0.1972
AdaBoost	0.682	0.705	0.429	0.892	0.224	0.2132

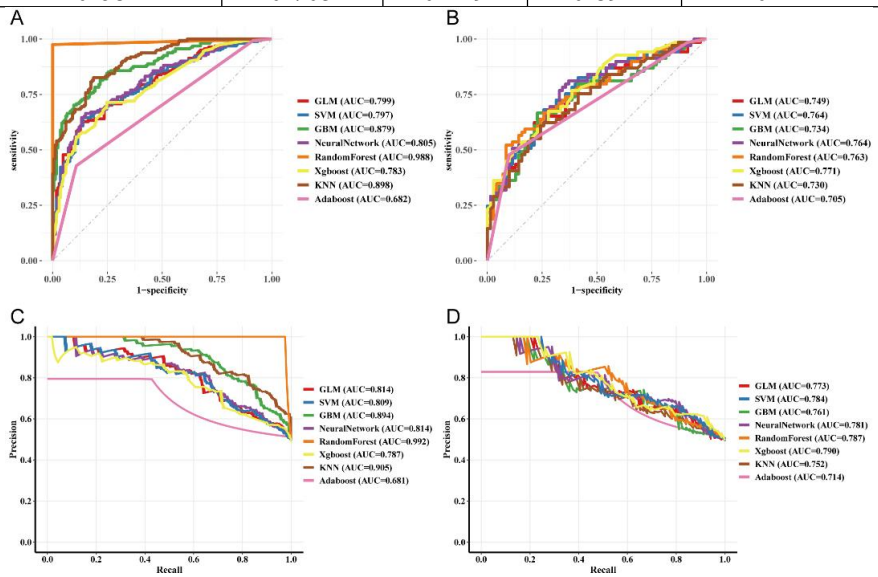

Figure 3. ROC and PR Curves of Different ML Models in the Training and Test Sets
(A, B) ROC Curves in the Training and Test Sets, Respectively; (C, D) PR Curves in the Training and Test Sets, Respectively

In terms of prediction error, XGBoost had the lowest Brier score in the test set, at 0.1596, which was lower than that of SVM (0.2381) and Random Forest (0.2093) (Table 3). The calibration curves indicated good consistency between the predicted probabilities and the observed outcomes for the XGBoost model (Figures 4A and 4B). Decision curve analysis further demonstrated that the XGBoost model achieved a greater net benefit across a wide spectrum of threshold probabilities (Figures 4C and 4D). Taken together, these evaluation metrics identified XGBoost as the best-performing model in this study.

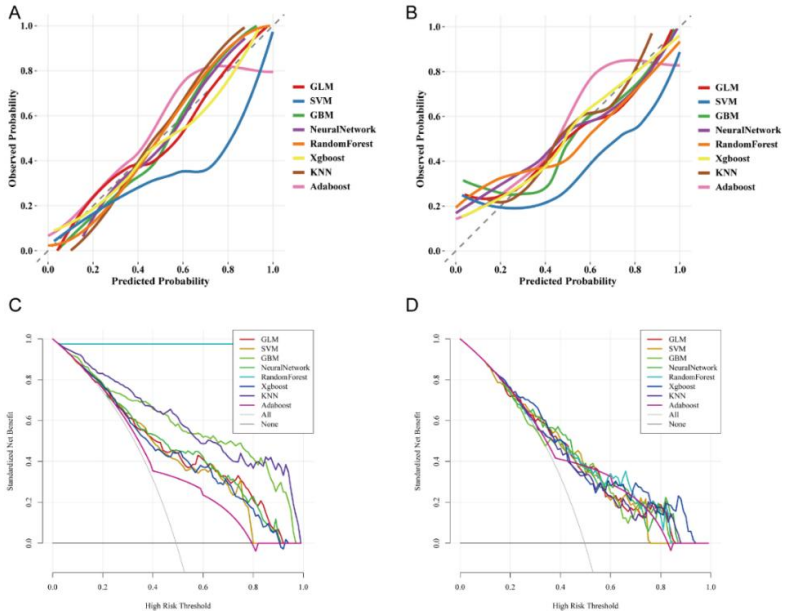

Figure 4. Calibration and decision curves of different ML models in the training and Test Sets(A, B) Calibration Curves in the Training and Test Sets, Respectively; (C, D) Decision Curves in the Training and Test Sets, Respectively

3.4 Key Features of the XGBoost Model and SHAP Interpretation

To further elucidate the output characteristics of the optimal XGBoost model, SHAP was used to illustrate the contribution of each feature. The SHAP beeswarm plot (Figure 5A) showed that different feature values had varying effects on the direction and magnitude of model output. Overall, samples with poorer self-rated health, lower life satisfaction, and the presence of IADL limitations had more data points distributed in the region with SHAP values greater than 0.

The feature importance bar plot (Figure 5B) showed that self-rated health, life satisfaction, and IADL limitation were the three features contributing most to model output. Daily sleep duration, out-of-pocket inpatient medical expenses, age, and childhood health status also made notable contributions. The SHAP dependence plots (Figures 5C–E) suggested certain interaction patterns between life satisfaction and out-of-pocket inpatient medical expenses, self-rated health, and IADL limitation.

In addition, the SHAP waterfall plot (Figure 5F) illustrated the formation process of the prediction result for an individual sample. The baseline model output for this sample was E[f(x)] = 0.518; after integrating the individual's feature values, the final model output was f(x) = 1, indicating that the sample was classified by the model as having depressive symptoms. Poorer life satisfaction, poorer self-rated health, and abnormal daily sleep duration were the main factors driving the prediction upward.

3.5 Subgroup Analysis

To further assess the performance of the XGBoost model across different demographic subgroups, stratified analyses were conducted according to age tertiles and sex (Figure 6).

Among the age subgroups, the AUCs of the model in the low-, middle-, and high-age groups in the training set were 0.772, 0.725, and 0.740, respectively (Figure 6A), and 0.761, 0.705, and 0.710, respectively, in the test set (Figure 6B). Across the sex subgroups, the AUCs were 0.752 for male patients and 0.765 for female patients in the training set (Figure 6C), and 0.732 and 0.715, respectively, in the test set (Figure 6D). Overall, the AUCs in all subgroups were above 0.70, and the 95% confidence intervals

overlapped to some extent, suggesting that the model performed relatively stably across different age and sex groups.

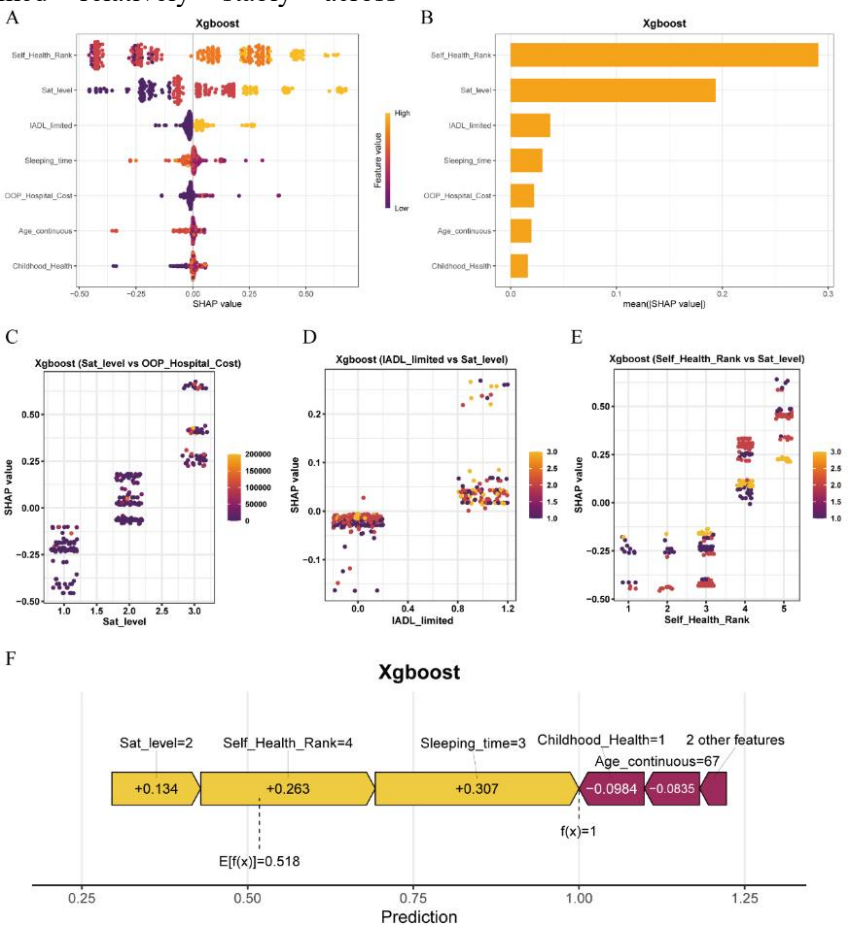

Figure 5. Key Features of the XGBoost Model and SHAP-based Interpretability Analysis (A) SHAP Beeswarm Plot of Features in the XGBoost Model; (B) Bar plot Showing The Mean Absolute SHAP Values of the Features in the XGBoost Model; (C) SHAP Dependence Plot for Life Satisfaction and out-Of-Pocket Inpatient Medical Expenses; (D) SHAP dependence Plot for IADL Limitation and Life Satisfaction; (E) SHAP Dependence Plot for Self-Rated Health and Life Satisfaction; (F) SHAP Waterfall Plot for the Prediction of an Individual Sample

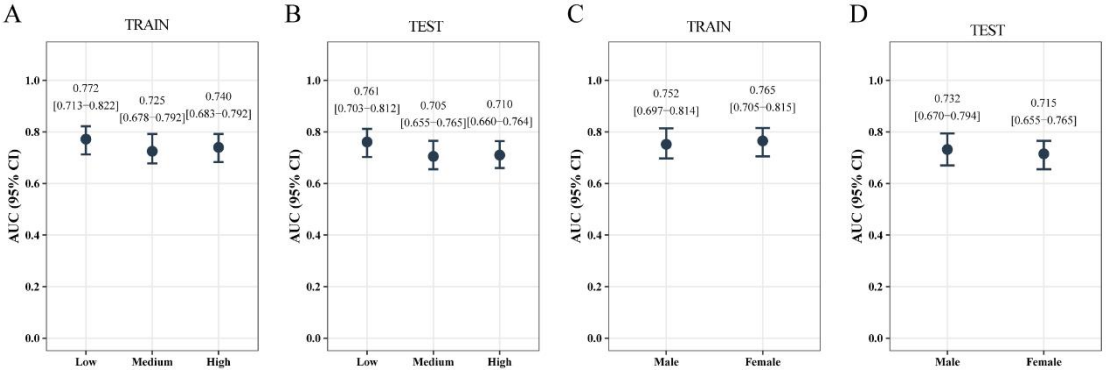

Figure 6. AUCs and 95% Confidence Intervals of the XGBoost Model Across Age and sex Subgroups(A) Model AUCs in Age Subgroups of the Training Set; (B) Model AUCs in Age Subgroups of the test Set; (C) Model AUCs in Sex Subgroups of the Training Set; (D) Model AUCs in Sex Subgroups of the Test Set.

4. Discussion

Using nationally representative CHARLS data, this study compared the performance of eight machine learning algorithms for identifying depressive symptoms in patients with cancer and developed an XGBoost model with relatively strong predictive performance. The results

showed that XGBoost maintained relatively stable discrimination in the test set as well as across age- and sex-based subgroups, and demonstrated certain advantages in calibration performance and decision curve analysis. Combined with LASSO feature selection, this study ultimately identified seven important features associated with the identification of depressive symptoms in patients with cancer, including age, out-of-pocket inpatient medical expenses, daily sleep duration, self-rated health, life satisfaction, IADL limitation, and childhood health status. Compared with previous studies that primarily relied on traditional regression methods to identify relevant factors, this study further integrated machine learning modeling with SHAP-based interpretability analysis, thereby enhancing the interpretability of the findings while focusing on identification performance and providing a new analytical perspective for the early risk warning of depressive symptoms in patients with cancer.

The SHAP analysis showed that self-rated health, life satisfaction, and IADL limitation were the three features contributing most to model output. This finding suggests that depressive symptoms in patients are related not only to physical functional status, but also closely to individuals' subjective perceptions of their own health and life conditions [8-9]. This result is generally consistent with the biopsychosocial model of medicine. During the course of disease and treatment, patients with cancer often face multiple stressors simultaneously, including fatigue, pain, functional limitations, role changes, and reduced social participation, all of which may jointly influence their psychological well-being. Poorer self-rated health and lower life satisfaction may reflect patients' overall perception of disease burden, declining quality of life, and uncertainty about the future, whereas IADL limitation indicates certain impairments in independent daily living and social functioning. These findings suggest that objective functional status and subjective health evaluation may jointly contribute to the identification of depressive symptoms in patients with cancer, and they also indicate that, in addition to focusing on anticancer treatment itself, clinicians should pay attention to the comprehensive assessment of patients' functional status and subjective health perceptions.

In addition to the above features, out-of-pocket inpatient medical expenses and daily sleep duration also made substantial contributions to the model. Higher out-of-pocket inpatient medical expenses suggest that financial burden may be an important marker of psychological vulnerability in patients with cancer. Previous research has suggested that cancer-related financial toxicity not only affects treatment adherence and quality of life, but may also aggravate negative emotional experiences through persistent economic stress [10]. In real-world clinical settings in China, the influence of economic burden on patients' psychological well-being deserves particular attention. On the other hand, abnormal sleep duration may reflect common sleep disturbances, pain burden, or overall physical and psychological imbalance among patients with cancer. Previous research has suggested that sleep problems are strongly linked to negative emotional states such as anxiety and depression [11]. Therefore, financial burden and abnormal sleep may not only be important model features, but also serve as warning signs that warrant priority attention when screening for psychological problems in patients with cancer.

From a methodological perspective, one strength of this study is that it considered both identification performance and model interpretability. Compared with traditional linear models, machine learning methods have certain advantages in handling multidimensional data, nonlinear relationships, and complex interactions, while SHAP helps improve the transparency of model output. Through both global and local SHAP explanations, this study not only identified key features contributing substantially to model output, but also demonstrated the direction and relative magnitude of each variable's impact on model predictions at the individual level. This strategy of "interpretable machine learning" may help improve the understandability of model results. It should be noted that the model established in this study is better suited as an auxiliary tool for early risk warning and further screening based on routinely available information, rather than as a replacement for standardized depression scales or professional psychiatric and psychological assessment.

Several limitations of this study warrant consideration. First, cancer diagnosis in the CHARLS dataset was primarily self-reported, and detailed clinical information such as

pathological type, TNM stage, and specific treatment regimens was unavailable; therefore, the influence of tumor heterogeneity on the identification of depressive symptoms could not be further evaluated. Second, as this study used cross-sectional data, the identified features mainly reflected associations with depressive symptoms rather than causal relationships. Third, the sample size was relatively limited, and only internal validation was conducted; therefore, the stability and generalizability of the model still need to be further examined in independent populations. Fourth, some variables included in the model, such as self-rated health and life satisfaction, are subjective indicators and conceptually overlap to some extent with depressive symptoms. Although these variables may improve identification performance, they may also partly enhance the apparent performance of the model. Finally, due to limitations in the original questionnaire design of the database, some potentially important factors affecting psychological status—such as social support, family functioning, history of psychological intervention, and more detailed cancer treatment information—could not be included in the analysis. Future studies should further evaluate the robustness, transportability, and practical value of the model by incorporating prospective cohorts, more detailed clinical variables, and external validation.

In summary, the XGBoost model developed based on CHARLS data showed relatively good overall performance in identifying depressive symptoms among patients with cancer, and life satisfaction, self-rated health, and IADL limitation were the key contributing features. This model may serve as an auxiliary tool for early warning and further screening of psychological problems in patients with cancer, and may provide a reference for risk stratification and subsequent assessment. However, its clinical utility still requires further validation in larger samples, prospective studies, and external populations.

5. Conclusion

In the present study, we constructed an XGBoost model based on CHARLS data to screen cancer patients at elevated risk of depressive symptoms. The model demonstrated relatively good discriminatory performance and interpretability, with self-rated health, life satisfaction, and IADL limitation emerging as the key contributing features. This model may provide a useful reference for the early warning, further screening, and risk stratification of psychological problems in patients with cancer, although further external validation is still needed.

Acknowledgments

The authors are grateful to the CHARLS team for making the data publicly available and to all participants for their input. We also thank those involved in data collection and management for their essential help.

References

[1] GETIE A, AYALNEH M, BIMEREW M. Global prevalence and determinant factors of pain, depression, and anxiety among cancer patients: an umbrella review of systematic reviews and meta-analyses. BMC Psychiatry, 2025, 25(1): 156.

[2] WALKER J, MULICK A, MAGILL N, et al. Major Depression and Survival in People With Cancer. Psychosom Med, 2021, 83(5): 410-6.

[3] SPIEGEL D. Cancer and depression. Br J Psychiatry Suppl, 1996, (30): 109-16.

[4] XU H, TANG W, LIANG Y, et al. Risk prediction models for depression in cancer survivors: A systematic review and meta-analysis. Medicine (Baltimore), 2025, 104(34): e43978.

[5] ALHUMAIDI N H, DERMAWAN D, KAMARUZAMAN H F, et al. The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review. JMIR Med Inform, 2025, 13: e68898.

[6] THOTTAKKARA P, OZRAZGAT-BASLANTI T, HUPF B B, et al. Application of Machine Learning Techniques to High-Dimensional Clinical Data to Forecast Postoperative Complications . PLoS One, 2016, 11(5): e0155705.

[7] ZHAO Y, HU Y, SMITH J P, et al. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol, 2014, 43(1): 61-8.

[8] AMANI O, MAZAHERI M A, MALEKZADEH MOGHANI M, et al. Mediating effects of rumination on insomnia in cancer survivors: Influences of cancer-related fatigue, fear of recurrence,

and psychological distress. Cancer Med, 2024, 13(18): e70189.

[9] STORENG S H, SUND E R, KROKSTAD S. Factors associated with basic and instrumental activities of daily living in elderly participants of a population-based survey: the Nord-Trondelag Health Study, Norway . BMJ Open, 2018, 8(3): e018942.

[10]CHEN X, TAN S, LI Y, et al. Financial toxicity, social support, and negative emotions among caregivers of children with cancer: a cross-sectional study in Western China . Front Public Health, 2025, 13: 1677962.

[11]PALAGINI L, MINIATI M, RIEMANN D, et al. Insomnia, Fatigue, and Depression: Theoretical and Clinical Implications of a Self-reinforcing Feedback Loop in Cancer. Clin Pract Epidemiol Ment Health, 2021, 17(1): 257-63.

