A Transformer-MLP Fusion Model with SHAP Analysis for Predicting Postoperative Survival in Intrahepatic Cholangiocarcinoma

Supplementary Material

Table of Contents

Method S1. Surgical, Adjuvant Chemotherapy and follow-up Protocols

All patients underwent curative-intent radical hepatectomy. The specific surgical strategy, including the choice between anatomical and non-anatomical resection and the extent of resection, was individualized based on a comprehensive preoperative evaluation. This evaluation encompassed tumor location and its relationship to major hepatic vasculature, the degree of underlying liver cirrhosis, quantitative assessment of liver reserve function, and the estimated future liver remnant volume.

For patients with an indication for adjuvant therapy, postoperative chemotherapy regimens were administered per contemporary standards for biliary tract cancers. Primary regimens included gemcitabine-based therapy (as monotherapy or combined with platinum agents), gemcitabine plus nab-paclitaxel, or capecitabine monotherapy. The specific agents, doses, and treatment cycles were determined and adjusted by the treating medical oncologist based on individual patient tolerance, organ function, and hematologic parameters.

Patients were evaluated one month after surgery and subsequently at regular 3-month intervals. Each follow-up visit included a clinical examination, laboratory tests (complete blood count, liver function tests, and serum tumor markers such as CA19-9 and CEA), and abdominal imaging. Imaging assessment was performed using contrast-enhanced computed tomography (CT) or magnetic resonance imaging (MRI) of the liver. The primary endpoints of the study was overall survival (OS). OS was defined as the time from surgery to death from any cause. All surviving patients were censored at the last follow-up date, which was October 30, 2025

Method S2. Collection of clinicopathological data

Clinical and pathological data were systematically collected from medical records. Body mass index (BMI) was calculated from recorded height and weight. All patients had a preoperative Child-Pugh grade A and an Eastern Cooperative Oncology Group (ECOG) performance status of 0 or 1. A history of viral hepatitis (hepatitis B or C) was documented. The presence of liver cirrhosis was evaluated based on preoperative imaging findings.

Comorbidities were documented and categorized. Hepatobiliary diseases included benign conditions such as intrahepatic or extrahepatic bile duct stones, cholecystitis, hepatic cysts, and hemangiomas (excluding viral hepatitis). Cardiopulmonary diseases encompassed hypertension, coronary heart disease, chronic obstructive pulmonary disease (COPD), and other chronic cardiac or respiratory disorders.

Surgical parameters were extracted from operative notes. The surgical approach was classified as either laparoscopic or open; procedures initiated laparoscopically but converted to open were analyzed within the laparoscopic group. Data on the performance of lymph node dissection and estimated intraoperative blood loss were recorded. The extent of hepatectomy was categorized as major (resection of three or more Couinaud segments) or minor.

Preoperative laboratory values collected included a complete blood count (from which the platelet-to-lymphocyte ratio, PLR, and neutrophil-to-lymphocyte ratio, NLR, were derived), liver function tests, coagulation profiles, and serum tumor markers (alpha-fetoprotein, carcinoembryonic antigen, and carbohydrate antigen 19-9).

Preoperative imaging characteristics assessed were maximum tumor diameter, tumor number, and tumor location. Histopathological findings from the resected specimens included tumor differentiation grade (well, moderate, or poor), and the presence or absence of microvascular invasion, perineural invasion, and liver capsule invasion. The receipt of postoperative adjuvant chemotherapy was also recorded.

Method S3. Specific data preprocessing process

The primary outcome was overall survival (OS), measured in months from the date of surgery. To mitigate the influence of extreme outliers, OS duration was truncated at 180 months. For the purpose of training certain models, a binary outcome variable was also created based on 24-month survival, classifying patients as either survivors or non-survivors at this postoperative time point.

Missing values in numerical variables were imputed using the median value calculated from the training set only, preventing data leakage into the validation and test sets. All continuous variables were then standardized using StandardScaler from scikit-learn, which centers the data to a mean of zero and scales it to unit variance. The scaling parameters were fitted on the training set and subsequently applied to the validation and test sets to maintain consistency.

Method S4. Specific steps of five-fold cross-validation

Patients were randomly divided into 5 subsets, with stratification based on key prognostic factors (e.g., tumor differentiation grade, vascular invasion status) to ensure balanced distribution. In each iteration, 4 folds (n=262) served as the training set for model construction and hyperparameter tuning, while the remaining fold (n=66) was used as the validation set for preliminary performance assessment. This process was repeated 5 times so that each subset was used once for validation. This approach makes efficient use of limited data, provides a robust performance estimate across multiple cycles, and reduces bias from a single random split.

Method S5. Feature number optimization process

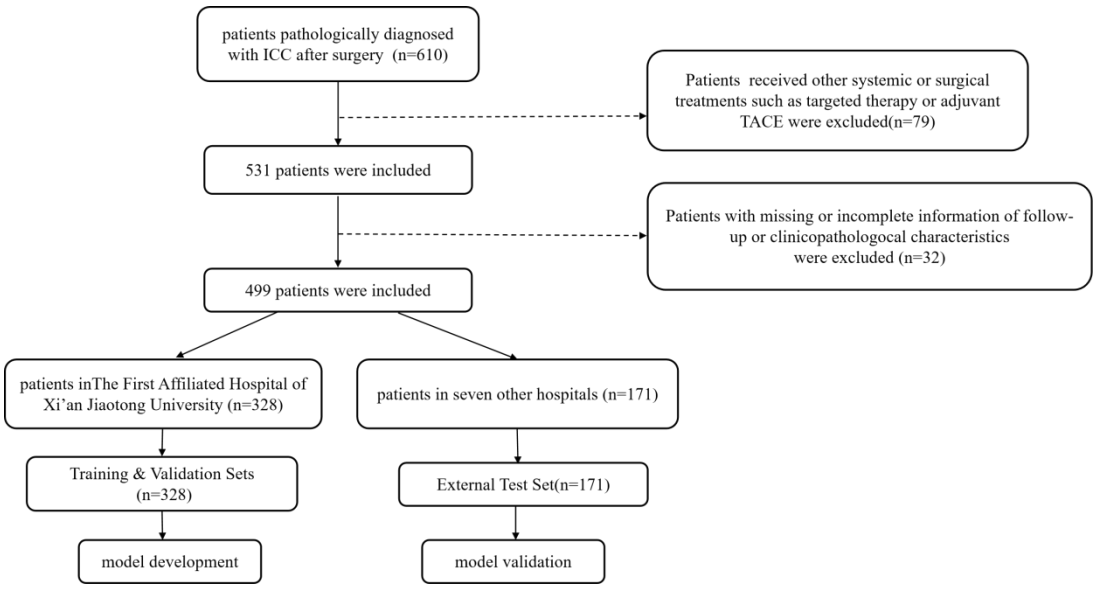
This process adhered to the principle of strict data separation: all feature quantity evaluations were conducted solely within the model development set (n = 328), with the independent test set (n = 171) remaining entirely excluded until final model evaluation. The procedure was as follows: based on the comprehensive feature importance scores derived in Section 2.7.1, the 31 candidate features were ranked in descending order of importance. For each candidate feature subset size k (ranging from 1 to 31), the top-k features were selected. The performance of each subset was evaluated using 5 repetitions of 5-fold cross-validation on the development set. In each cross-validation iteration, the data were partitioned into five folds; four folds were used for model training, and the remaining fold was used for validation, with the area under the curve (AUC) computed on the validation fold. The mean AUC across the five folds was recorded as the performance metric for that repetition. The overall performance and stability for a given k were summarized by the mean and standard deviation of the AUC values obtained from the five independent repetitions.

Method S6. The traditional and deep-learning models and their theoretical rationale

As a generalized linear model, logistic regression predicts the probability of a binary outcome by transforming a linear combination of input features via the Sigmoid function, which maps the result to a probability between 0 and 1. Its parameters are directly interpretable as log-odds, providing clear clinical insights into the relationship between predictors and the outcome, making it a benchmark for evaluating linear prognostic relationships. SVM performs classification by identifying the maximum-margin hyperplane that best separates different classes in a high-dimensional feature space. This study utilized a radial basis function (RBF) kernel, which allows the model to capture complex, non-linear decision boundaries by implicitly mapping input features into higher-dimensional spaces. As an ensemble learning method, Random Forest constructs a multitude of decision trees during training and aggregates their predictions. This approach enhances model robustness and predictive accuracy by reducing overfitting. A key advantage is its ability to handle high-dimensional data and provide intrinsic measures of feature importance, which aids in interpreting the contribution of individual predictors. XGBoost is an advanced implementation of gradient-boosted decision trees. It operates by iteratively training a sequence of weak learners (trees), where each subsequent model focuses on correcting the errors of its predecessors. Known for its high predictive accuracy and computational efficiency, XGBoost effectively manages complex feature interactions and avoids overfitting through regularization. LightGBM is another gradient-boosting framework optimized for performance and scalability. It employs a histogram-based algorithm to bucket continuous features into discrete bins, significantly accelerating the training process. Furthermore, its leaf-wise tree growth strategy expands the nodes that yield the largest loss reduction, leading to high efficiency and lower memory usage, particularly beneficial for large datasets.

As a foundational deep learning model, the Multi-Layer Perceptron (MLP) stacks multiple fully connected layers with nonlinear activation functions, enabling strong nonlinear mapping capabilities. It automatically learns effective intermediate representations from raw features through hierarchical transformations . The Recurrent Neural Network (RNN) processes sequential data via internal recurrent connections, capturing temporal dependencies. In this study, static patient feature vectors were reshaped into single-time-step sequences to assess their potential in modeling intrinsic feature correlations .To address training instability in traditional RNNs, the Long Short-Term Memory (LSTM) network incorporates input, forget, and output gates, mitigating gradient vanishing or explosion issues while capturing long-range dependencies . The Gated Recurrent Unit (GRU) simplifies this architecture by combining the forget and input gates into a single update gate, resulting in a more parameter-efficient model that often matches LSTM performance with reduced computational overhead .In contrast to recurrent models, the Transformer architecture relies entirely on self-attention mechanisms to compute global dependencies among input features. By discarding recurrent and convolutional structures, its core self-attention layer dynamically weights interactions between all feature pairs, efficiently capturing complex feature interdependencies

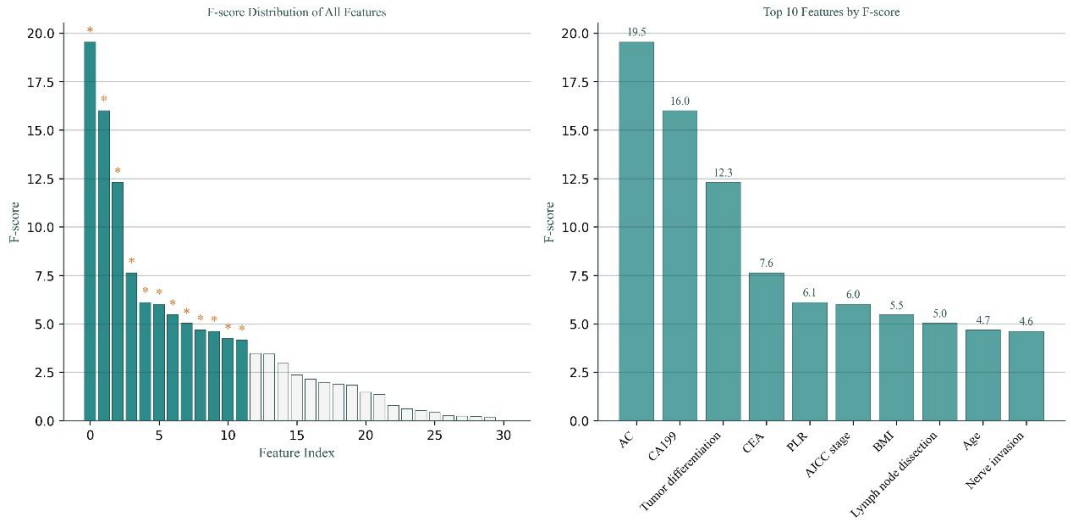
Figure S1. Flowchart of patient selection



the inclusion and exclusion criteria: The inclusion criteria were: (1) receipt of radical liver resection for ICC; (2) definitive postoperative pathological diagnosis of ICC; and (3) no prior history of liver-directed systemic therapy or surgery. The exclusion criteria were: (1) non-radical resection (microscopically positive surgical margin, R1/R2); (2) perioperative mortality (death within 90 days of surgery); (3) presence of concurrent other active malignancies; and (4) diagnosis of recurrent ICC.

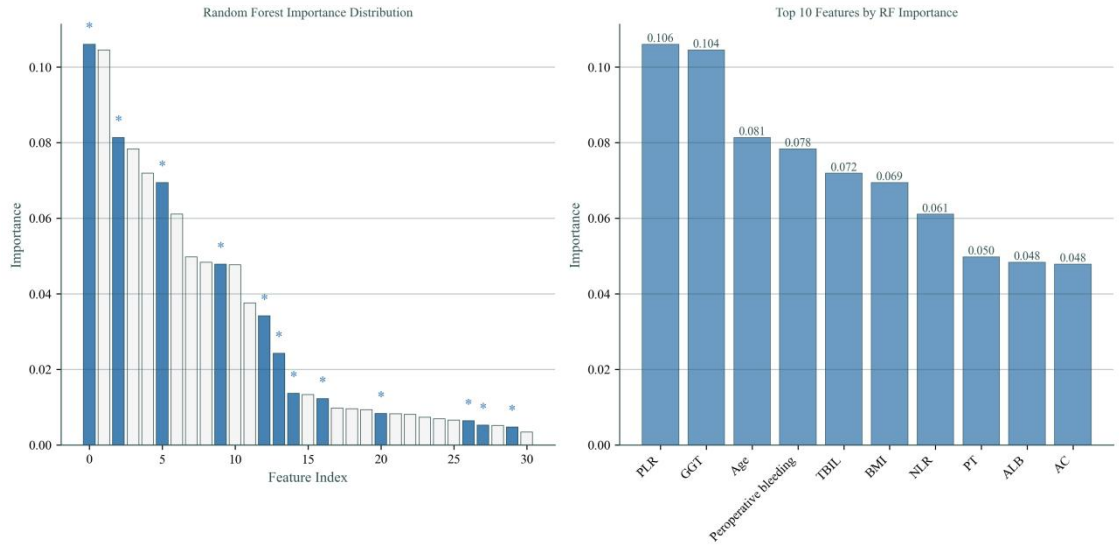
This multicenter, retrospective cohort study was conducted using data from eight tertiary hospitals in China: The First Affiliated Hospital of Xi'an Jiaotong University, West China Hospital of Sichuan University, Shengjing Hospital of China Medical University, Xijing Hospital and Tangdu Hospital of Air Force Medical University, Baoji Central Hospital, Baoji People's Hospital, and Yulin Hospital of The First Affiliated Hospital of Xi'an Jiaotong University. The cohort distribution by contributing center was as follows: The First Affiliated Hospital of Xi'an Jiaotong University (n=328), West China Hospital of Sichuan University (n=58), Tangdu Hospital of Air Force Medical University (n=53), Shengjing Hospital of China Medical University (n=26), Xijing Hospital of Air Force Medical University (n=21), Baoji Central Hospital (n=8), Baoji People's Hospital (n=3), and Yulin Hospital (n=2)

Figure S2. F-score distribution of features



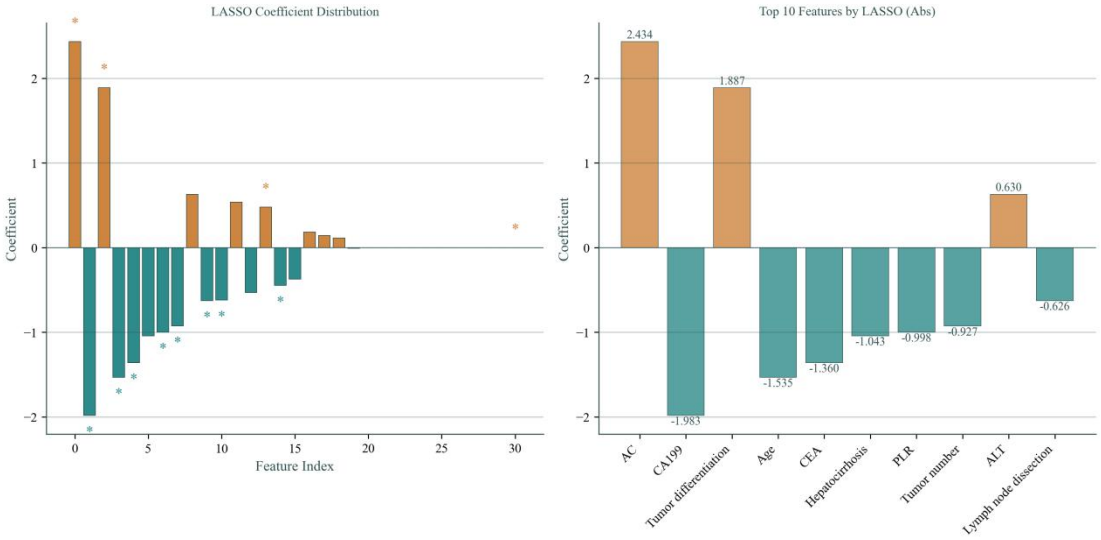
The univariate F-value analysis, a filter method, assessed the linear correlation strength between each feature and overall survival (OS). Features such as adjuvant chemotherapy (AC), CA19-9, and the platelet-to-lymphocyte ratio (PLR) demonstrated significantly higher F-values ($P < 0.05$) compared to others. The top-ranked features predominantly fell into categories of treatment-related factors (e.g., AC), tumor markers (e.g., CA19-9), and systemic inflammatory indicators (e.g., PLR), aligning with established clinical prognostic knowledge for ICC .

Figure S3. Random Forest importance distribution



The random forest algorithm, an ensemble learning model, evaluates feature importance based on the mean decrease in impurity (Gini Importance) across all decision trees, capturing non-linear relationships. This method assigned higher importance to tumor differentiation grade, microvascular invasion (MVI), and intraoperative blood loss, reflecting the impact of tumor biology and surgical details on prognosis.

Figure S4. LASSO Coefficient Distribution

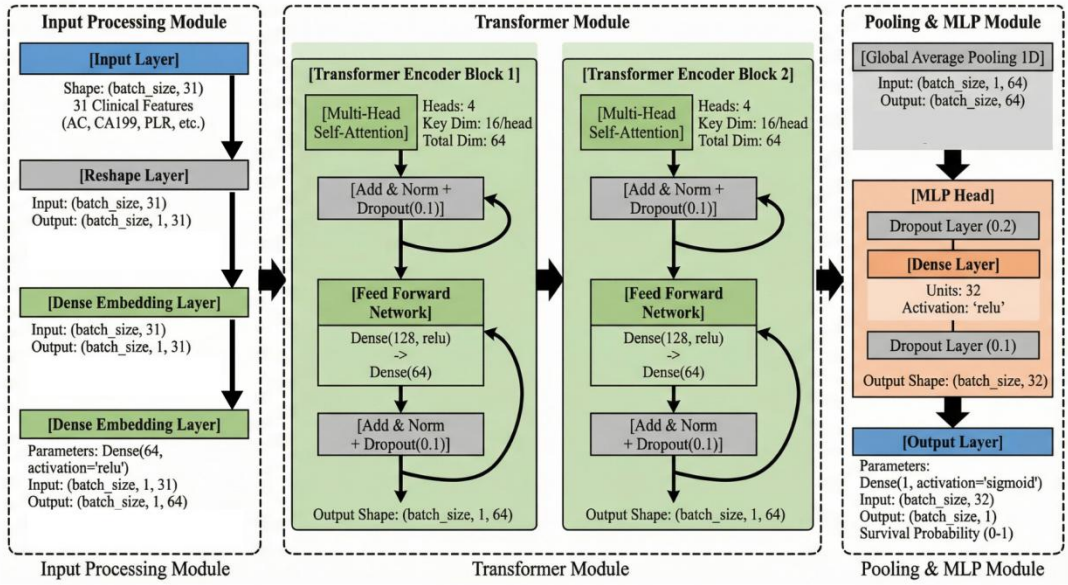


The LASSO (Least Absolute Shrinkage and Selection Operator) regression technique, which applies L1 regularization, shrunk the coefficients of 12 less relevant features (e.g., patient height and weight) to zero, retaining 19 features. The features with the largest absolute coefficient values were preoperative CA19-9, AC, and tumor size, corroborating the core role of these variables while enhancing model parsimony by eliminating redundant predictors.

Figure S5. Model performence vs. Number of features on 11 models

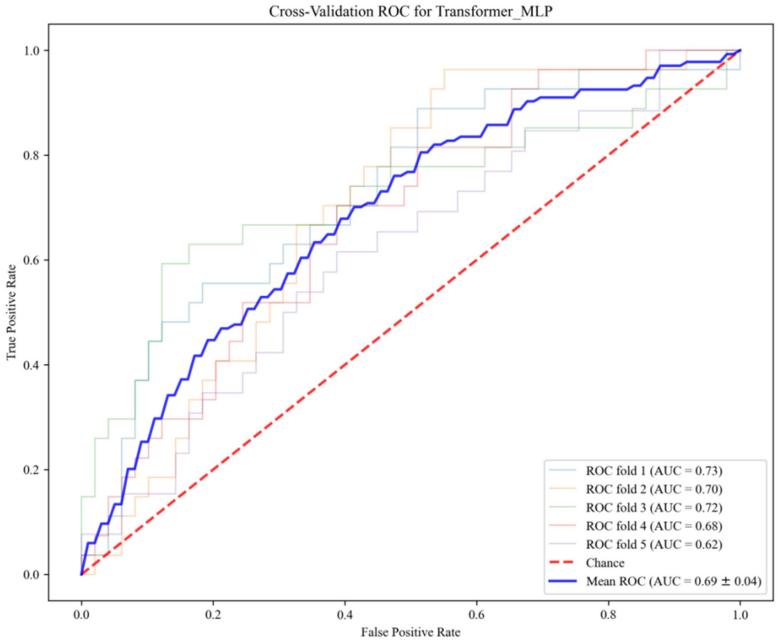
(a)Aggregated line chart depicting the maximum and mean AUC values across five traditional machine learning models as the number of input features increases; (b) Line charts illustrating the AUC variations of each of the five individual traditional machine learning models with changing feature quantities; (c) Aggregated line chart presenting the maximum and mean AUC values across six deep learning models with an increasing number of input features; (d) Line charts showing the AUC trends of each of the six individual deep learning models as a function of the number of features entering the models. AUC: Area Under the Receiver Operating Characteristic Curve, XGBoost: eXtreme Gradient Boosting, LightGBM: Light Gradient Boosting Machine.

Figure S6. Architecture Design of Transformer-MLP Model



The proposed Transformer-MLP architecture processes 31 clinical features by projecting them into a 64-dimensional embedding space (d_{model}=64). The core encoder consists of 2 Transformer blocks, each utilizing 4-head self-attention and a feed-forward network with a hidden dimension of 128, stabilized by layer normalization and residual connections. The resulting sequence is aggregated via global average pooling into a 64-dimensional vector. Subsequently, this vector feeds into an MLP classification head containing a 32-unit dense layer with ReLU activation and Dropout for feature fusion, culminating in a Sigmoid output layer for survival probability prediction. MLP: Multilayer Perceptron.

Figure S7. 5-Fold Cross-Validation ROC for Transformer-MLP Model



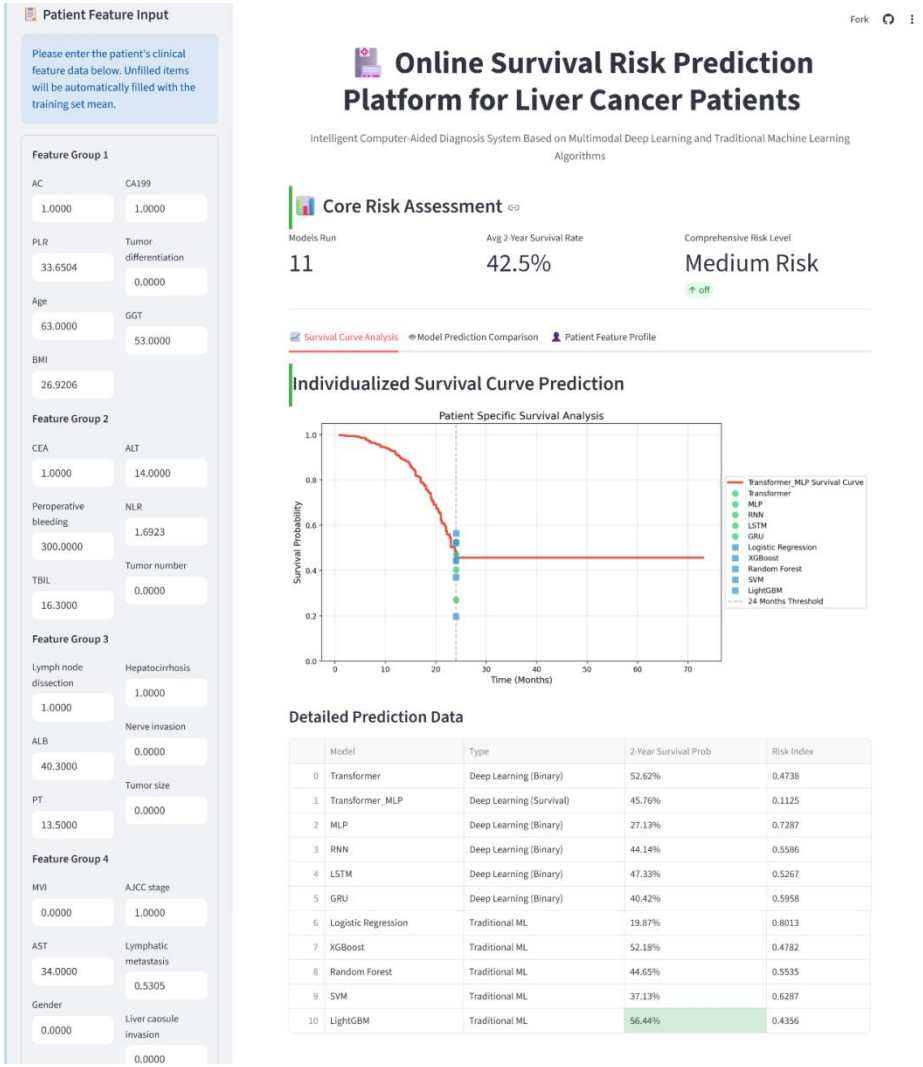
This figure presents the ROC curves of the Transformer-MLP model across five independent folds of cross-validation. MLP: Multilayer Perceptron, ROC: Receiver Operating Characteristic.

Figure S8. Online Predictor

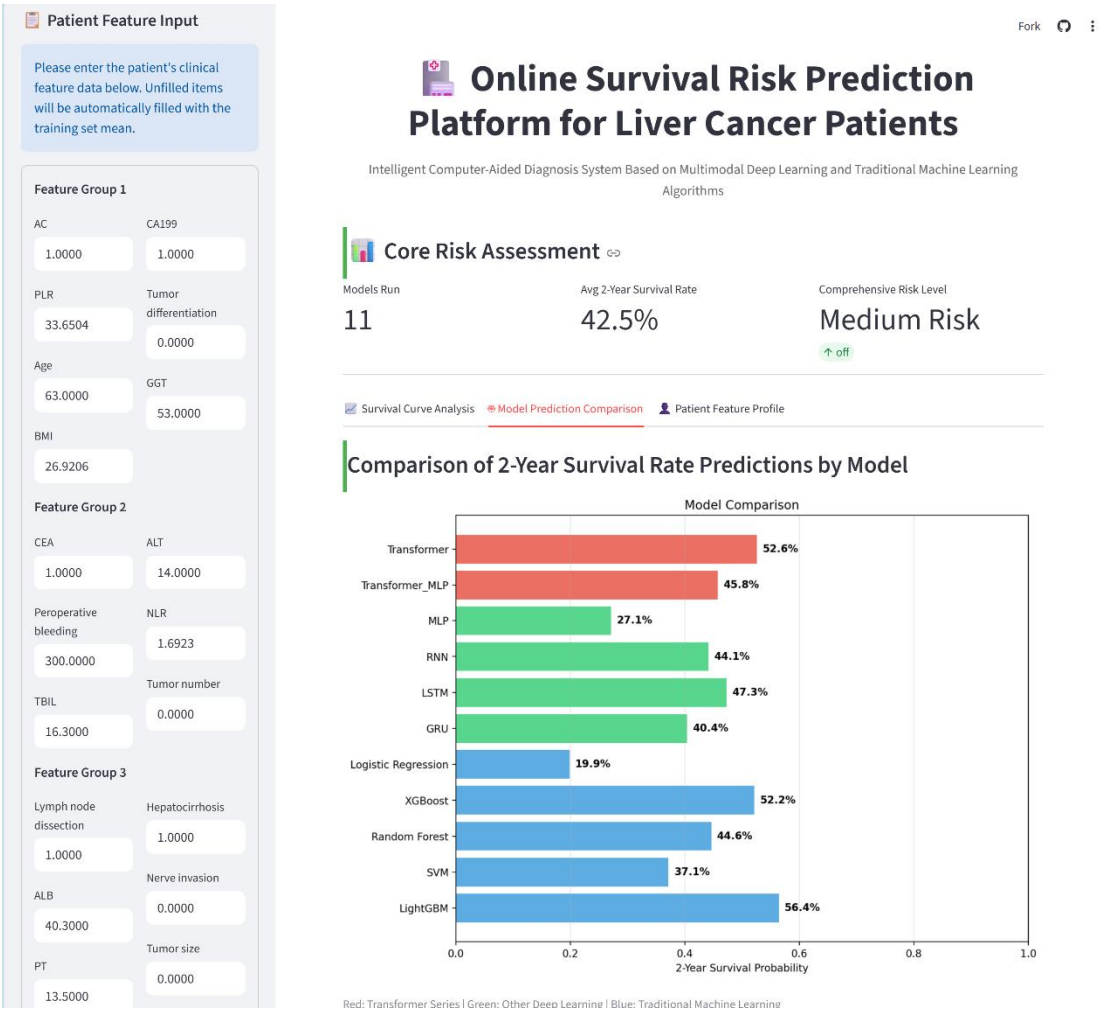
To transform the predictive models developed in this study into clinically applicable tools and enhance the translational value and practicality of research findings, we have developed an online survival risk predictor for intrahepatic cholangiocarcinoma (ICC) following surgery (https://icc-prediction.streamlit.app/). This platform integrates all 11 predictive models developed in this study (5 traditional machine learning models and 6 deep learning models). Through an intuitive visual interface and systematic analysis workflow, it provides clinicians with personalized postoperative mortality risk assessment services for individual patients.

Upon accessing the online predictor interface, the left panel serves as the patient clinical and pathological feature input area—a foundational step for achieving precise predictions. To enhance the convenience and logical flow of clinical data entry, this study categorizes 31 core predictive features (covering patient baseline characteristics, comorbidities, surgery-related indicators, laboratory tests, and tumor pathology features) into 5 feature groups. Each group contains 6-7 relevant features, preventing confusion from excessive data entry. Default values in the feature input fields are set based on clinical data characteristics: For categorical variables such as gender, hepatitis history, cirrhosis, and surgical approach, the default value is set to 0. For continuous variables such as BMI, PLR, NLR, liver function indicators (ALT, AST, GGT, etc.), and tumor markers (AFP, CEA, CA199), the default values are the mean values from the model training set. Items not manually entered will automatically adopt the default values. This ensures continuity in the prediction process while minimizing the impact of missing data on prediction results. Clinicians must supplement or correct feature values item by item based on the patient's actual clinical data. After completing data entry, click the “Start Prediction” button at the bottom of the input area. The system will automatically invoke 11 models for parallel computation, generating comprehensive prediction results.

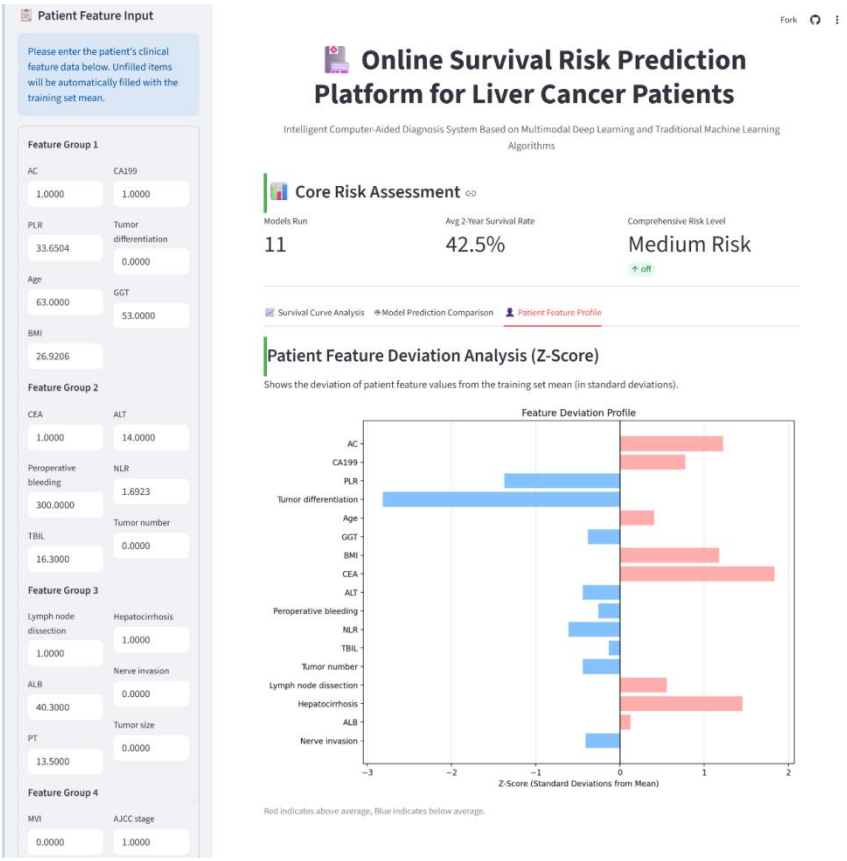
Upon completion of the prediction, the platform homepage first displays the core risk assessment panel, presenting the key predictive conclusions through concise quantitative metrics. The left side of the panel shows the number of successfully trained models (up to 11), clearly reflecting the reference dimensions of the prediction results. The central area presents the average 2-year survival rate across all model predictions. This metric synthesizes the predictive strengths of different algorithms, serving as the core quantitative reference for patients' short-term postoperative survival risk. The right side categorizes overall risk levels based on the average 2-year survival rate, with specific criteria as follows: Average survival rate >70% is classified as “Low Risk,” indicating a high probability of patient survival within 2 years post-surgery, allowing routine clinical follow-up intervals; 40%-70% is classified as “Moderate Risk,” requiring enhanced follow-up monitoring with close observation of disease progression; <40% is classified as “High Risk,” indicating elevated short-term postoperative mortality risk. Shorter follow-up intervals are recommended, and postoperative intervention strategies should be optimized based on the patient's specific condition. This grading system provides direct clinical guidance for rapid risk assessment and preliminary management planning.



Below the core assessment panel, three functional tabs expand detailed predictive analyses to meet diverse clinical interpretation needs. The first tab, “Survival Curve Analysis,” focuses on dynamic visualization of individual survival trends and comprehensive presentation of multi-model prediction results. See Figure 14. The central panel displays individualized survival curves calculated by the core Transformer-MLP fusion model. The horizontal axis represents postoperative time (in months), while the vertical axis shows survival probability. The curve trajectory visually reflects the patient's evolving survival risk over time. A gray dashed line marks the critical 24-month (2-year) milestone, enabling clinicians to quickly reference the patient's 2-year survival probability. Simultaneously, the 2-year survival predictions from the other 10 models are overlaid as scatter plots. Traditional machine learning models (logistic regression, SVM, LightGBM, random forest, XGBoost) are marked with blue squares, while deep learning models (MLP, Transformer, RNN, LSTM, GRU) are marked with green dots. This creates an intuitive comparison of multi-model predictions, facilitating observation of the consistency among different algorithms. Below the chart, a detailed data table systematically lists each model's type (traditional machine learning / deep learning), 2-year survival probability, and risk index. The prediction with the highest survival probability is automatically highlighted, providing data support for clinicians to analyze model prediction differences in depth and assess prediction reliability based on clinical experience.



The second tab is titled “Model Prediction Comparison,” which systematically displays the prediction results of 11 models through a horizontal bar chart. See Figure 15. The horizontal axis represents 2-year survival probability (range 0-1), while the vertical axis lists all model names sequentially. Different colored bars distinguish model types: red denotes Transformer-based models (Transformer, Transformer-MLP hybrid), green represents other deep learning models (MLP, RNN, LSTM, GRU), blue denotes traditional machine learning models (logistic regression, SVM, LightGBM, random forest, XGBoost). Each color scheme aligns with the survival curve analysis page to ensure visual consistency. Each bar chart displays the corresponding 2-year survival probability percentage on its right side, facilitating precise comparison of prediction biases across models. When most models converge toward consistent predictions, the reliability of the conclusion significantly increases; If substantial prediction discrepancies arise among models, clinicians can further validate results by considering specific patient characteristics (e.g., presence of high-risk pathological factors, comorbidities), providing flexibility for clinical interpretation of predictions.



The third tab, “Patient Feature Profile,” analyzes potential influencing factors of prediction outcomes from a feature deviation perspective, helping clinicians understand the core basis of model predictions. See Figure 16. This page displays a Z-score deviation chart illustrating the degree of deviation of 17 key features (the optimal feature subset selected by traditional machine learning models) from the training set mean. Z-scores, measured in standard deviations, visually reflect the difference between individual patient characteristics and those of the overall training population. Red bars indicate patient feature values exceeding the training set mean, while blue bars indicate values below the mean. Bar length corresponds to the magnitude of deviation. This chart enables clinicians to rapidly identify abnormal patient characteristics (e.g., significantly elevated CA199, abnormal liver function indicators, high NLR), which often serve as key factors influencing model predictions. This aids in understanding the logic behind model predictions while providing reference for further clinical investigation of core risk factors affecting patient prognosis and developing targeted intervention strategies.

The core advantage of this online predictor lies in transforming complex model prediction processes into an intuitive, user-friendly clinical tool. Clinicians can obtain multidimensional predictive results through simple feature input without requiring specialized machine learning knowledge. Its multi-model prediction strategy leverages the strengths of different algorithms while enhancing prediction reliability through result comparison. Diverse visualization charts (survival curves, horizontal bar charts, Z-score deviation plots) enable comprehensive analysis spanning survival trends, model comparisons, and feature interpretation. This provides robust support for personalized clinical survival risk assessment, effectively advancing the translation of research findings from basic science to clinical application. The tool offers practical utility for precision management following ICC surgery.

Table S1. Baseline characteristics of patients in Training & Validation sets and external Test set

Variables	Overall (n=499)	Training set (n=328)	test set (n=171)	P
Gender				0.411
Female	211	143	68	
male	288	185	103	
Age (yrs)	59			0.070
BMI (kg/m²)	22.6			0.444
Hepatitis				0.323
absence	384	248	136	
presence	115	80	35	
Hepatocirrhosis				0.723
absence	351	229	122	
presence	148	99	49	
Hepatobiliary disease				0.638
absence	376	245	131	
presence	123	83	40	
Cardiopulmonary disease				0.761
absence	366	242	124	
presence	133	86	47	
AC				0.688
absence	295	196	99	
presence	204	132	72	
Surgical type				0.012
Open surgery	174	127	47	
Laparoscopic surgery	325	201	124	
Lymph node dissection				0.646
absence	105	71	34	
presence	394	257	137	
Extent of liver resection				0.018
Non-extended liver resection	290	203	87	
Extended liver resection	209	125	84	
Peroperative bleeding (ml)	300 (5,5000)			0.930
PLR	123.2 (4.2,556.4)			0.582
NLR	2.8 (0.4,34.4)			0.649
AST	29.0 (10,711.0)			0.862
ALT	27.0 (5.0,579.0)			0.847
GGT	69.0 (9.0,1971.0)			0.656
PT	12.5 (10.0,28.5)			0.124
TBIL (umol/L)	13.7 (1.0,422.3)			0.972
ALB (g/L)	40.5 (11.5,51.3)			0.131

Variables	Overall (n=499)	Training set (n=328)	test set (n=171)	P
AFP (u/ml)				0.777
<=7.00	414	271	143	
> 7.00	85	57	28	
CEA (u/ml)				0.358
<=5.00	391	253	138	
> 5.00	108	75	33	
CA199 (u/ml)				0.682
<=30.00	190	127	63	
> 30.00	309	201	108	
Tumor differentiation				0.889
low	239	157	82	
moderate	245	162	83	
high	15	9	6	
MVI				0.892
absence	323	213	110	
presence	176	115	61	
Nerve invasion				0.629
absence	417	276	141	
presence	82	52	30	
Liver capsule invasion				0.925
absence	197	129	68	
presence	302	199	103	
Tumor number				0.963
1	412	271	141	
>=2	87	57	30	
Tumor size (cm)				0.637
<=5	232	150	82	
>5	267	178	89	
AJCC stage				0.037
I stage	138	95	43	
II stage	113	83	30	
III stage	248	150	98	
Lymphatic metastasis				0.273
negative	302	205	97	
x	78	52	26	
positive	119	71	48	

Table S2. model performence vs. number of features

n_features	Logistic Regression	XGBoost	Random Forest	SVM	LightGBM	Transformer	Transformer-MLP	MLP	RNN	LSTM	GRU
1	0.5550	0.5550	0.5550	0.5550	0.5550	0.5398	0.5664	0.5615	0.5550	0.5550	0.5550
2	0.6011	0.6011	0.6011	0.4603	0.6011	0.5999	0.6010	0.5979	0.5089	0.5089	0.5990
3	0.6174	0.5581	0.4930	0.5671	0.5773	0.6016	0.5639	0.6207	0.6249	0.6185	0.6243
4	0.6468	0.6013	0.5701	0.6194	0.5921	0.5931	0.6675	0.6571	0.5757	0.6525	0.6016
5	0.6459	0.5662	0.5250	0.5480	0.5724	0.6253	0.6291	0.6574	0.6183	0.6149	0.5770
6	0.6522	0.5571	0.5390	0.5814	0.5826	0.6601	0.6539	0.5823	0.6367	0.6621	0.6458
7	0.6635	0.5577	0.5578	0.6195	0.5760	0.6650	0.6186	0.6686	0.6699	0.6159	0.6636
8	0.6800	0.5990	0.6100	0.6390	0.6028	0.6501	0.6751	0.6511	0.6168	0.6393	0.5578
9	0.6705	0.5784	0.5861	0.6323	0.5969	0.6754	0.6841	0.7027	0.5946	0.7228	0.6795
10	0.6638	0.5988	0.6220	0.6477	0.6025	0.6590	0.6794	0.6490	0.6118	0.6821	0.6545
11	0.6620	0.6025	0.5893	0.6519	0.5966	0.6931	0.6896	0.5815	0.6362	0.6754	0.6104
12	0.6611	0.6046	0.5874	0.6619	0.5832	0.6429	0.6512	0.5890	0.6499	0.6918	0.5715
13	0.6539	0.5915	0.5895	0.6656	0.5790	0.6912	0.6450	0.6209	0.5657	0.6295	0.6551
14	0.6618	0.6091	0.6026	0.6484	0.5879	0.6456	0.6866	0.6100	0.5883	0.6247	0.6139
15	0.6636	0.6021	0.6097	0.6652	0.5729	0.7206	0.6599	0.6490	0.6118	0.6109	0.6613
16	0.6644	0.5957	0.6293	0.6610	0.5984	0.6984	0.6867	0.6590	0.5858	0.5845	0.6247
17	0.6733	0.6428	0.6455	0.6686	0.5975	0.7307	0.7083	0.5550	0.5878	0.6665	0.6490
18	0.6715	0.6252	0.6221	0.6677	0.5838	0.6895	0.7190	0.6683	0.6337	0.6739	0.6557
19	0.6702	0.6310	0.6393	0.6558	0.5966	0.6689	0.6963	0.7022	0.6042	0.6624	0.6468
20	0.6729	0.6277	0.6288	0.6537	0.6267	0.6842	0.7203	0.6240	0.6119	0.6322	0.6614
21	0.6687	0.6119	0.6000	0.6539	0.6173	0.6461	0.6984	0.6230	0.6347	0.6255	0.6872
22	0.6692	0.6115	0.6345	0.6559	0.5939	0.6672	0.6713	0.6554	0.6090	0.6580	0.6412
23	0.6766	0.6437	0.6229	0.6604	0.6086	0.6996	0.6765	0.6326	0.6425	0.6048	0.6641
24	0.6773	0.6109	0.6050	0.6650	0.6049	0.6859	0.7022	0.6206	0.6465	0.6887	0.6705
25	0.6753	0.6069	0.6390	0.6645	0.5925	0.6716	0.7223	0.6478	0.6718	0.6694	0.6528
26	0.6733	0.6204	0.6139	0.6651	0.5975	0.6900	0.7113	0.6759	0.6404	0.6435	0.6886
27	0.6750	0.6139	0.6083	0.6643	0.5940	0.7104	0.7476	0.6891	0.6268	0.6633	0.6483
28	0.6747	0.6118	0.6390	0.6627	0.6063	0.7230	0.7379	0.6458	0.6025	0.6720	0.6468
29	0.6754	0.5820	0.6214	0.6494	0.6136	0.7190	0.7268	0.6818	0.6443	0.6441	0.6686
30	0.6729	0.6085	0.6080	0.6534	0.6507	0.7254	0.7239	0.6732	0.6817	0.6404	0.6306
31	0.6715	0.6338	0.6262	0.6666	0.6219	0.7203	0.7435	0.6605	0.6599	0.6843	0.6762

Table S3. C-index of all 11 prediction models

Model	C-index	95% CI
Logistic Regression	0.6492	(0.5965, 0.7015)
XGBoost	0.6048	(0.5517, 0.6561)
Random Forest	0.6237	(0.5666, 0.6799)
SVM	0.6498	(0.5975, 0.7012)
LightGBM	0.5817	(0.5298, 0.6365)
MLP	0.6591	(0.6009, 0.7140)
Transformer-MLP	0.7462	(0.6898, 0.7950)
Transformer	0.6783	(0.6276, 0.7266)
LSTM	0.6468	(0.5878, 0.7011)
GRU	0.6483	(0.5927, 0.7014)
RNN	0.6443	(0.5897, 0.6959)

This table presents the discriminative performance (evaluated by C-index) and corresponding 95% CIs of 5 traditional machine learning models (Logistic Regression, XGBoost, Random Forest, SVM, LightGBM) and 6 deep learning models (MLP, Transformer-MLP, Transformer, LSTM, GRU, RNN) on the independent test set, reflecting the ability of each model to distinguish ICC postoperative survival outcomes.

Table S4a. Comprehensive performance metrics of all 11 prediction models on the training set for ICC postoperative survival prediction. This table displays multiple performance indicators (AUC, Accuracy, Sensitivity, Specificity, PPV, NPV, Brier Score) of all 11 models (5 traditional machine learning + 6 deep learning) on the training set, illustrating the fitting effect and discriminative ability of each model during the training phase.

Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	Brier Score
Logistic Regression	0.7587	0.6860	0.4310	0.8255	0.5747	0.7261	0.1878
XGBoost	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0010
Random Forest	0.9993	0.9634	0.8966	1.0000	1.0000	0.9464	0.0897
SVM	0.8969	0.7348	0.3103	0.9670	0.8372	0.7193	0.1776
LightGBM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0115
MLP	0.8635	0.7481	0.4301	0.9231	0.7547	0.7464	0.1582
Transformer	0.9168	0.8397	0.7204	0.9053	0.8072	0.8547	0.1142
Transformer-MLP	0.9627	0.8626	0.9570	0.8107	0.7355	0.9716	0.1386
LSTM	0.7494	0.6947	0.5376	0.7811	0.5747	0.7543	0.2177
GRU	0.8112	0.7252	0.4839	0.8580	0.6522	0.7513	0.1766
RNN	0.7638	0.6985	0.5161	0.7988	0.5854	0.7500	0.1915

Table S4b. Comprehensive performance metrics of all 11 prediction models on the independent test set for ICC postoperative survival prediction. This table presents multiple performance indicators (AUC, Accuracy, Sensitivity, Specificity, PPV, NPV, Brier Score) of all 11 models on the independent test set, facilitating systematic comparison of generalization ability among different model types and highlighting the superior performance of the Transformer-MLP model in practical clinical application scenarios.

Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	Brier Score
Logistic Regression	0.6835	0.6608	0.3607	0.8273	0.5366	0.7000	0.2138
XGBoost	0.6177	0.6140	0.2951	0.7909	0.4390	0.6692	0.2876
Random Forest	0.6373	0.6608	0.2131	0.9091	0.5652	0.6757	0.2197
SVM	0.6867	0.6374	0.1148	**0.9273**	0.4667	0.6538	0.2152
LightGBM	0.5969	0.6140	0.3115	0.7818	0.4419	0.6719	0.2754
MLP	0.7317	0.7310	0.4262	0.9000	**0.7027**	0.7388	0.1974
Transformer	0.7578	**0.7427**	0.6066	0.8182	0.6491	0.7895	**0.1917**
Transformer-MLP	**0.8131**	0.6784	**0.8361**	0.5909	0.5312	**0.8667**	0.1964
LSTM	0.7073	0.7018	0.5246	0.8000	0.5926	0.7521	0.2246
GRU	0.7004	0.6901	0.4590	0.8182	0.5833	0.7317	0.2065
RNN	0.6981	0.6550	0.4590	0.7636	0.5185	0.7179	0.2088

Table S5. Five-fold cross-validation performance of Transformer-MLP model

Fold	cohort	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	Brier Score
1	Train	0.9300	0.8680	0.7944	0.9082	0.8252	0.8900	0.1175
	Validation	0.7294	0.7105	0.4815	0.8367	0.6190	0.7455	0.2018
2	Train	0.8136	0.7591	0.4206	0.9439	0.8036	0.7490	0.1700
	Validation	0.6999	0.6316	0.2593	0.8367	0.4667	0.6721	0.2146
3	Train	0.9342	0.8482	0.6636	0.9490	0.8765	0.8378	0.1174
	Validation	0.7181	0.7105	0.2593	0.9592	0.7778	0.7015	0.1951
4	Train	0.8678	0.7921	0.5981	0.8980	0.7619	0.8037	0.1464
	Validation	0.6810	0.6316	0.3333	0.7959	0.4737	0.6842	0.2130
5	Train	0.9195	0.8355	0.7407	0.8878	0.7843	0.8614	0.1172
	Validation	0.6162	0.6267	0.4231	0.7347	0.4583	0.7059	0.2473
Mean	**Train**	**0.8929**	**0.8206**	**0.6435**	**0.9174**	**0.8103**	**0.8284**	**0.1337**
	Validation	**0.6885**	**0.6622**	**0.3513**	**0.8326**	**0.5591**	**0.7018**	**0.2144**
Std	**Train**	**0.0510**	**0.0437**	**0.1465**	**0.0269**	**0.0406**	**0.0580**	**0.0235**
	Validation	**0.0408**	**0.0408**	**0.0954**	**0.0898**	**0.1360**	**0.0267**	**0.0189**

The Transformer-MLP model demonstrated a good fit on the training set, with an average AUC of 0.8929 ± 0.0510. On the validation set, the average AUC was 0.6885 ± 0.0408. The performance gap indicates a degree of overfitting, although validation set performance remained substantially better than chance. Performance across validation folds showed reasonable variation (AUC range: 0.6162 - 0.7294), reflecting the influence of data partitioning.

Fold	cohort	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV	Brier Score	C-index
1	Test	0.7127	0.7018	0.5738	0.7727	0.5833	0.7658	0.2063	0.6469
2	Test	0.7048	0.6959	0.3607	0.8818	0.6286	0.7132	0.2043	0.6696
3	Test	0.7303	0.7076	0.3934	0.8818	0.6486	0.7239	0.1958	0.6676
4	Test	0.6914	0.6667	0.3934	0.8182	0.5455	0.7087	0.2094	0.6518
5	Test	0.6821	0.6550	0.5082	0.7364	0.5167	0.7297	0.2252	0.6610
Mean		**0.7043**	**0.6854**	**0.4459**	**0.8182**	**0.5845**	**0.7283**	**0.2082**	**0.6594**
Std		**0.0176**	**0.0224**	**0.0895**	**0.0578**	**0.0516**	**0.0224**	**0.0108**	**0.0093**

On the independent test set, the Transformer-MLP model achieved an average AUC of 0.7043 and an average C-index of 0.6594. Performance was stable across folds (AUC SD: 0.0176; C-index SD: 0.0093). Sensitivity varied moderately (range: 0.3607 - 0.5738), suggesting an influence of training sample composition on high-risk patient identification, while specificity remained relatively stable (mean: 0.8182), indicating reliable discrimination of low-risk patients.