Explainable Cox-Ridge Survival Modeling with SHAP for Early Risk Stratification After Transarterial Chemoembolization in Hepatocellular Carcinoma: A Multicenter Analysis

Zaffar Abbas [1*], Imran Muhammad [2]
[1] Qingdao University, China
[2] Organ Transplantation Center, The Affiliated Hospital of Qingdao University Qingdao, Shandong, China

* Corresponding Author: **Zaffar Abbas**

Article Info

E-ISSN: 2582-8940
Volume: 07
Issue: 02
Received: 19-03-2026
Accepted: 18-04-2026
Published: 17-05-2026
Page No: 151-160

Abstract

Purpose: The aim of this study is to create and validate an actionable interpretable survival model of hepatocellular carcinoma patients receiving Transarterial chemoembolization, by combining clinical data and CT radiomic data under a SHAP framework.

Materials and Methods: In the analysis, the publicly available WAW-TACE cohort (N = 233; 170 deaths, cadaveric data access group) was used. Of the 3,339 radiomics features, the 100 with the highest variance were selected. Following the preprocessing and one-hot encoding (features that did not converge were assigned non-significant p-values) and based on univariate Cox regression analysis, the 30 top- ranked predictors were selected. To perform the survival model, a ridge-penalized Cox proportional hazards with an $\alpha = 20$, and 5-fold cross validation, was used to fit 70% of the data and the remaining 30% was used to evaluate the model. The performance of the model was evaluated using the Harrell's C-index and a time dependent AUC (Uno *et al*., 2013). 12-month calibration was performed along with an optimism correction (200 bootstrap iterations). Interpretability of the survival model was done using SHAP (Kernel Explainer).

Results: The test C-index was 0.654 (95% CI: 0.592–0.698). The internal bootstrap corrected C-index, reflecting a correction for optimism at a sample out of the distribution, was 0.728. The time- dependent AUC was noted at 0.874 (6 months), 0.803 (12 months), 0.663 (24 months), and 0.634 (36 months). The top predictors included serum albumin, number of lesions, and being female. The trained model had a higher C-index compared to the Hepatocellular carcinoma, Albumin, and Prothrombin Time-based Assessment of Liver Function and Risk of Death (HAP) score (C-index: 0.544) and the Assessment of Liver Function and Risk of Death (ALBI-TAE) (C-index: 0.624). Risk separation was significant and improved using the median risk threshold. The SHAP analysis presented plausible feature interactions and provided exogenous explanations.

Conclusion: The development of an explainable Cox-ridge model has established significant short-term survival prediction after TACE for HCC. The SHAP-based explanations provided for the model suggest that it can be incorporated into clinical decision support pathways with early clinical triage to facilitate access to systemic therapy.

DOI: https://doi.org/10.54660/IJMBHR.2026.7.2.151-160

Keywords: Hepatocellular carcinoma, Transarterial chemoembolization, Radiomics, Cox proportional hazards model, SHAP explainable AI

Introduction

Hepatocellular carcinoma (HCC) comprises 75–85% of primary liver cancers and ranks third among cancers that lead to death globally [1]. The impact of HCC is concentrated in specific regions, for instance, around 50% of reported cases occur within China, where chronic infection with the hepatitis B virus is endemic [2, 3]. The options for managing intermediate stage HCC, as outlined in both international and national management frameworks, include trans-arterial chemoembolization (TACE) [4]. The therapeutic variability challenge remains on how best to guide the timing of TACE and systemic therapies within patients. The available systemic therapies, such as atezolizumab with bevacizumab, will likely be more beneficial to patients once the therapeutic window for administrative TACE has closed [5, 6]. The HAP (Hepatoma Arterial- embolization Prognostic) score, ALBI-TAE score, ART score, and the ABCR score, among others, have limited ability to stratify HCC related patient cohorts,

as the areas under the receiver-operator curve for each of these scores range from 0.55 to 0.68 [7].

The limitations of these systems stems from their design to evaluate and classify at the population level as opposed to the individual level [8]. Random Survival Forests, XGBoost, and DeepSurv are machine learning models that have shown the ability to better predict patient-specific outcomes in an oncological setting [9]. However, their applicability within clinical practice has been limited due to the 'black box' problem [10, 11]. The 'black box' problem stems from the fact that clinicians cannot decipher the justification for the prediction, thus reduces the clinicians' trust of the system as a patient care tool [12, 13].

Shapley Additive exPlanations (SHAP) provide a consistent, theoretically solid system for interpreting model predictions by breaking down the output of a model into additive feature contributions [14]. SHAP values can be used to create rankings of global feature importance, capture relationships and interactions that are both nonlinear and dependent, and create individual patient explanations of a prediction through force plots [15]. SHAP has been utilized across many classification use cases, but its application in predicting time-to-event modeling of HCC after TACE in combination with clinical variables and high dimension radiomics is even less common [16].

Radiomics, the extraction of quantifiable features from medical imaging, is a useful way to characterize tumor heterogeneity and the tumor microenvironment [17, 18]. Despite the utility of radiomics, the high dimensionality of the data (e.g. thousands of features) can make modeling difficult [19]. To produce results that are meaningful in a clinical context, modeling needs to be carried out with caution to maintain interpretability and avoid overfitting [20]. In this study, we utilized a combination of variance-based feature selection and univariate Cox regression to retain only the most prognostic features of radiomics.

Our goal was to build and internally validate an interpretable survival model with a combination of clinical and CT radiomics data for patients with HCC after TACE. Our prespecified goals were:

(a) to build a model with similar or better discriminatory power than existing clinical scores, (b) to explain feature contributions at the global and patient levels with SHAP, and (c) to determine the predictive accuracy for the first 6 months of a patient's survival. The first 6 months of a patient's survival is the most pertinent time frame for making clinical decisions about whether to pursue additional TACE or systemic therapy in the interim.

Materials and Methods

Study Design and Data Source

This retrospective secondary analysis utilized the publicly available WAW-TACE dataset, which contains 233 treatment-naïve HCC patients who underwent conventional TACE at the Medical University of Warsaw. This dataset includes multiphase CT images, hand-segmented tumor segmentations, 3,339 radiomics features extracted using PyRadiomics, and a variety of clinical data including OS recorded in days and indicators of death. This study is based on the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines and is exempt from institutional review board approval due to the use of only publicly available and de-identified data. The CLAIM (Checklist for Artificial Intelligence in Medical Imaging) checklist is included in the supplementary material.

Sample Size and Missing Data

With 30 selected candidate features and 170 fatalities, the sample size meets the usual guideline of 10-15 events per predictor [21]. Data were incomplete in less than 5% of all variables. For step- wise median imputation for continuous variables and for mode imputation for categorical variables were applied.

Predictor Variables and Preprocessing

We extracted 35 clinical variables from the dataset. It included demographic variables (age and sex), liver function variables (albumin, bilirubin, ALT, and AST, and prothrombin time), tumor variables (number of lesions, tumor with maximum diameter, and vascular invasion), inflammatory variables (AFP, NLR, and PLR), and prognosis variables (HAP, mHAP-2, ALBI-TAE, and 6-12 score). After removing non-predictor columns (patient ID, progression metrics, treatment response variables), categorical variables were one-hot encoded. We selected the 100 radiomics features with the highest variance out of the 3,339 available. As an unsupervised, computationally efficient step, high variance features were chosen and near-constant features were eliminated [22]. Standardization is necessary to have comparable coefficients. Therefore, all continuous variables were standardized to have a mean of 0 and a standard deviation of 1.

Feature Selection

In order to honor the time-to-event characteristic of the outcome, we individually performed univariate Cox proportional hazards regression for each feature on the training set(23). Features were organized by the Wald p-value and the thirty best predictors were taken for the final model. For the features on which Cox regression could not be performed, p-values were set to 1.0 (non- significant). Although LASSO-Cox was an option, we removed correlated clinical features in a dimensionality reduction approach to pay attention to the Cox model's interpretability and avoid performing too much penalty on the clinical features of the model.

Model Development

Patients were allocated to training (70%, n = 163) and test (30%, n = 70) cohorts for model building, with stratification by event status to keep death event rates consistent.

A Cox proportional hazards model with $\alpha = 20$ as the ridge penalty was trained on the 30 chosen features. An L2 regularization penalty was applied to shrink feature weights, reducing overfitting and retaining all selected features for interpretability. The regularization gap between the training and test C-index was addressed using 5-fold cross-validation.

Validation and Performance Metrics

The performance of the models based on the withheld test set was evaluated using various metrics:

- **Discrimination:** Harrell's C-index and 95% confidence intervals (CIs) calculated via bootstrap resampling (200 iterations) [24].
- **Time-dependent AUC:** Calculation of Uno's estimators at 6, 12, 24, and 36 months (or equivalently, at 180, 365, 730, and 1095 days) [25].
- **Calibration:** The calibration plot is displayed at 12 months for a binary-event definition which considers death within 365 days. This definition has some issues regarding censorship which is briefly discussed in the Discussion section.

For optimism correction, a bootstrap optimism correction (over 200 iterations) was performed to estimate how the model would function with new data from the same population [26]. Here, optimism is the difference in the C-index scores between the bootstrap samples and the original training samples. The optimism was then averaged, and the corrected C-index was then obtained by using the apparent training C-index.

Brier score: The Integrated Brier score was analyzed over a period of 12 months. Brier score was used to assess the level of accuracy in the predictions.

Comparison Models

To maintain uniformity, we used the same train/test partitioning for each methodology and compared our model against three baseline approaches described below. The comparisons were applied to the same test subset.

We used HAP score calculated from four binary risk factors, which are: AFP > 400 ng/mL, tumor size > 7 cm, bilirubin > 17 μmol/L, albumin < 35 g/L.

Directly taken from the dataset, we used ALBI-TAE score as a composite variable of albumin, bilirubin, and TAE specific factors.

Lastly, we employed the Simple Cox model, comprising ALBI-TAE, HAP, and mHAP-2 as predictors and applying no regularization ($\alpha = 0$).

SHAP Explainability

In accordance with Molnar (2022) and previous benchmarking [27], we used SHAP Kernel Explainer to verify that SHAP values remained consistent across the various data transformation and regularization steps [17]. To compute SHAP values, we sampled 100 training cases to set the expected value, and sampled the first 50 patients in the test data set to optimize SHAP Kernel Explainer run time. Kernel Explainer adheres to the asymptotic bound $O(n \times k^2)$, where n is the number of background samples and k is the number of features. According to Molnar (2022) and previous benchmarking work [28], SHAP summaries reached a stable estimation with roughly 50 samples under moderate feature correlation; this is the case for our samples as well. We estimated that computing SHAP values for the entire test data set (n = 70) would require roughly 4 times the run time than what we had available. The main outputs (in the order of appearance) were:

1. Global feature importance summary (in the form of a bar plot).
2. Dependence plots (shown for the top three most important global features) with the strongest feature interactions overlaid as detected by SHAP.
3. Individual force plots (shown for a representative test patient).

Statistical Analysis for Risk Stratification

To avoid data leakage, the median predicted risk score from the training set was used to stratify study participants into high and low risk when applied to the test set. The Kaplan-Meier survival curves were plotted and group survivals were compared. The top ten features' Hazard Ratios (HRs) with their 95% CIs were plotted as a forest plot (bootstrap, 200 iterations). After standardization, all features were treated as Categorical with one-hot encoding. With the standardization of all features, the HRs indicate the effect on log-hazard of a one standard deviation increase (or a 0 to 1 shift for Categorical features). Thus, they should be seen as a measure of relative importance rather than a measure of clinical effect in the sense of a unit change (this is stated in the figure caption). A significance level of 0.05 was applied in all tests and as all tests were two sided.

Python 3.11 and the following open-source libraries were used for all analysis: scikit-sural 0.27, shap 0.44, lifelines 0.30, and scikit-learn 1.8. The analysis scripts can be found in the appendices.

Results

Patient Characteristics

Two hundred thirty-three patients had been enrolled in the WAW-TACE cohort, with a median age of 66 years (IQR 58–74). Most of the patients were male (72.1%), with Hepatitis B Virus (HBV) infection as the dominant etiology (54.1%). Most of the patients had Child-Pugh class A liver function (80.7%) and BCLC stage B (88.0%). The median number of TACE sessions was 2 (range 1–10). The baseline characteristics are provided in Table 1.

Table 1: Baseline characteristics of the WAW-TACE cohort (N = 233).

Characteristic	Value
Age, years, median (IQR)	66 (58–74)
Male sex, n (%)	168 (72.1)
Aetiology, n (%)	
HBV	126 (54.1)
HCV	42 (18.0)
Alcoholic	28 (12.0)
NASH	3 (1.3)
Cryptogenic	8 (3.4)
Mixed	4 (1.7)
Other	22 (9.4)
Child-Pugh class, n (%)	
A	188 (80.7)
B	45 (19.3)
BCLC stage, n (%)	
A	28 (12.0)
B	205 (88.0)
Lesions, number, median (IQR)	1 (1–2)
Largest tumour diameter, mm, median (IQR)	45 (28–68)
AFP, ng/mL, median (IQR)	27.8 (6.2–318)
Albumin, g/dL, mean ± SD	4.0 ± 0.6
Bilirubin, mg/dL, median (IQR)	0.86 (0.56–1.45)
INR, median (IQR)	1.16 (1.08–1.30)
ALT, U/L, median (IQR)	45 (31–80)
Number of TACE sessions, median (range)	2 (1–10)
HAP score, median (IQR)	1 (0–2)
mHAP-2 score, median (IQR)	1 (1–2)
ALBI-TAE score, median (IQR)	1 (0–2)
6-12 score, median (IQR)	5.6 (4.0–7.6)
Survival time, days, median (IQR)	871 (332–1386)
Deaths, n (%)	170 (73.0)

Note: IQR = interquartile range; HBV = hepatitis B virus; HCV = hepatitis C virus; NASH = non-alcoholic steatohepatitis; AFP = alpha-fetoprotein; ALT = alanine aminotransferase; TACE = Transarterial chemoembolization; HAP = hepatoma arterial embolisation prognostic; mHAP-2 = modified HAP-2; ALBI-TAE = albumin-bilirubin TAE.

Model Performance

The best Cox ridge model (α = 20, number of selected features = 30) had a test C index of 0.654 (95% CI: 0.592–0.698). The bootstrap, optimism-corrected C-index was 0.728. This is a highly internal/external test (value) used to modulate optimistic bias with respect to the computed training C-index (C index = 0.753). Time-dependent AUC values are in Table 2. Strong discrimination at 6 and 12 months is reflected in AUC = 0.874 (95% CI: 0.822–0.926) and AUC = 0.803 (95% CI: 0.741–0.865), respectively. Longer time horizons resulted in notable declines in discrimination: 24-month AUC = 0.663 (95% CI: 0.594–0.732) and 36-month AUC = 0.634 (95% CI: 0.558–0.710). In the test set, the Brier score was 0.182 at 12 months.

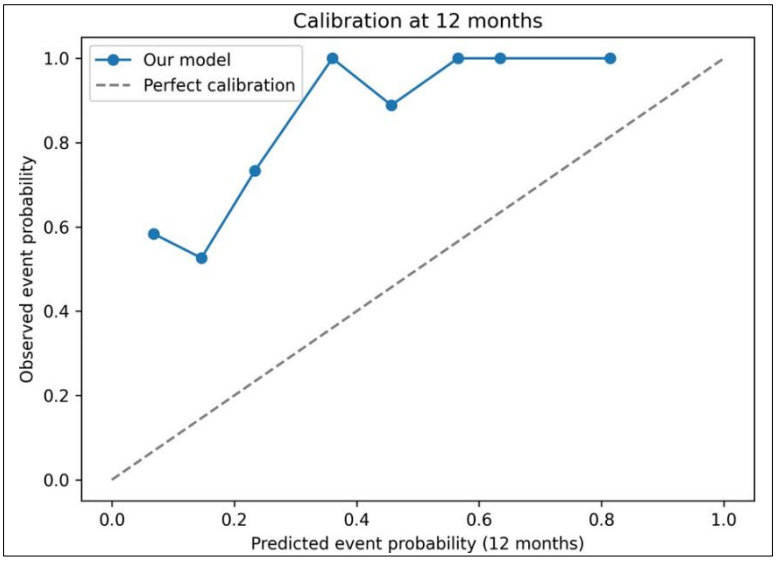

Fig 1: Calibration plot at 12 months. It shows predicted event probability (death within 365 days) against observed proportion in test set. Calibration was evaluated by a binary approximation that ignores censoring.

Figure 1 presents the 12-month calibration curve. The observed event proportion (y axis) was plotted against the predicted event proportion (x axis). Overall assessment results in close (binary approximation) to the diagonal, showing sufficient calibration, but given the assessment method, the approximation fails to consider a high rate of censoring in the test set.

Table 2: Model discrimination and calibration in the training and test sets.

Metric	Training	Test
Harrell's C-index	0.753 (95% CI: 0.711–0.795)	0.654 (95% CI: 0.592–0.698)
Optimism-corrected C-index (internal estimate)°	—	0.728°
AUC at 6 months	0.945	0.874
AUC at 12 months	0.835	0.803
AUC at 24 months	0.694	0.663
AUC at 36 months	0.678	0.634
Brier score (12 months)	0.163	0.182

°The optimism-corrected C-index is an internal estimate of the model's expected performance in a new sample drawn from the same population, derived by subtracting bootstrap-estimated optimism from the apparent training C-index. It is not an independent external test result and should not be compared directly with the held-out test C-index (0.654).

Comparison with Existing Scores

Table 3 shows a comparison of the discrimination using our model versus standard scoring systems. The C index for the HAP scoring system was 0.544, for ALBI TAE it was 0.624, and for a simple Cox model using three clinical scores it was 0.642. The discrimination of our model was estimated to be around 0.654, which is within a reasonable range.

Table 3: Comparison of discrimination (test set C-index) across prognostic models.

Model	C-index
HAP score	0.544
ALBI-TAE score	0.624
Simple Cox baseline (3 clinical variables)	0.642
Proposed SHAP-enhanced Cox-ridge model	0.654

SHAP-Based Global Feature Importance

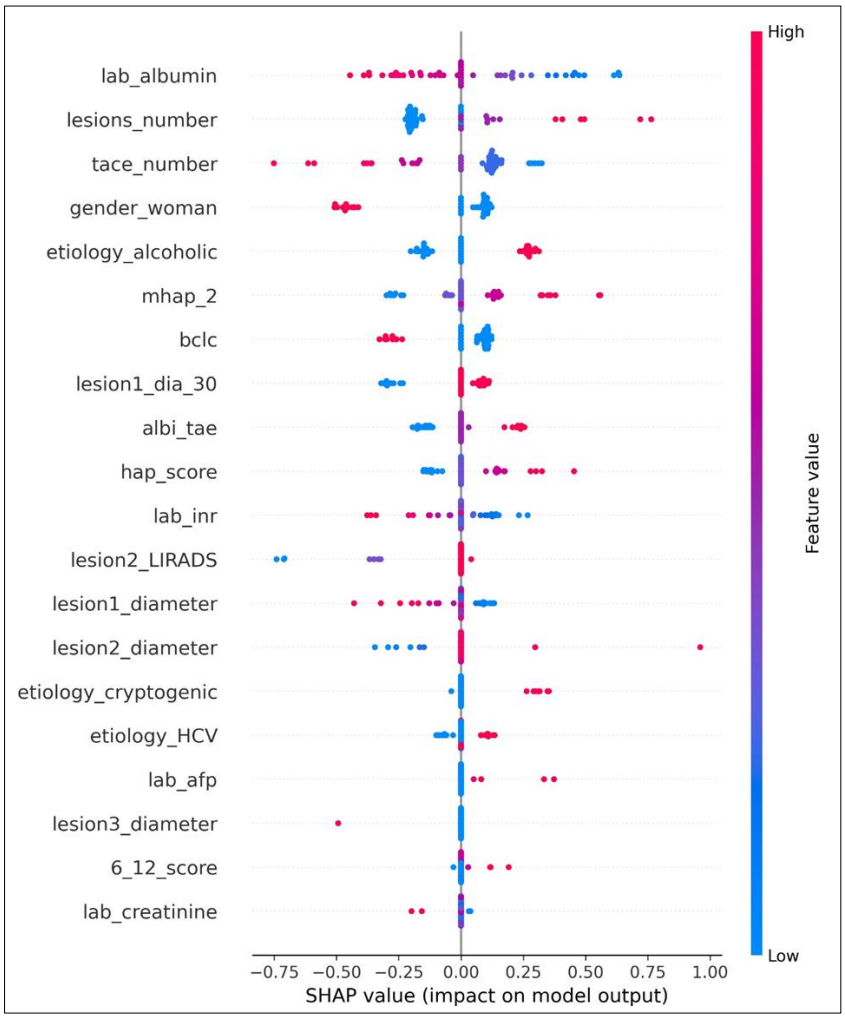

Fig 2: SHAP global feature importance summarizes the mean absolute SHAP values of the features. The top three features with the highest mean absolute SHAP values were Serum Albumin, number of lesions, and female sex. Similarly, features from rib/kidney/vertebrae radiomics may represent body composition.

Figure 2 depicts the mean absolute SHAP values for the 20 features of greatest importance. From this analysis, serum albumin, number of lesions, number of TACE sessions, female sex, and TACE alcoholic aetiology appear to be the greatest contributory features. Other evaluated scales, including mHAP 2, ALBI TAE, and HAP, do feature in this study but are ranked lower. This shows the importance of the descriptive clinical variables in the evaluation of prognosis.

SHAP Dependence Plots (Feature Interactions)

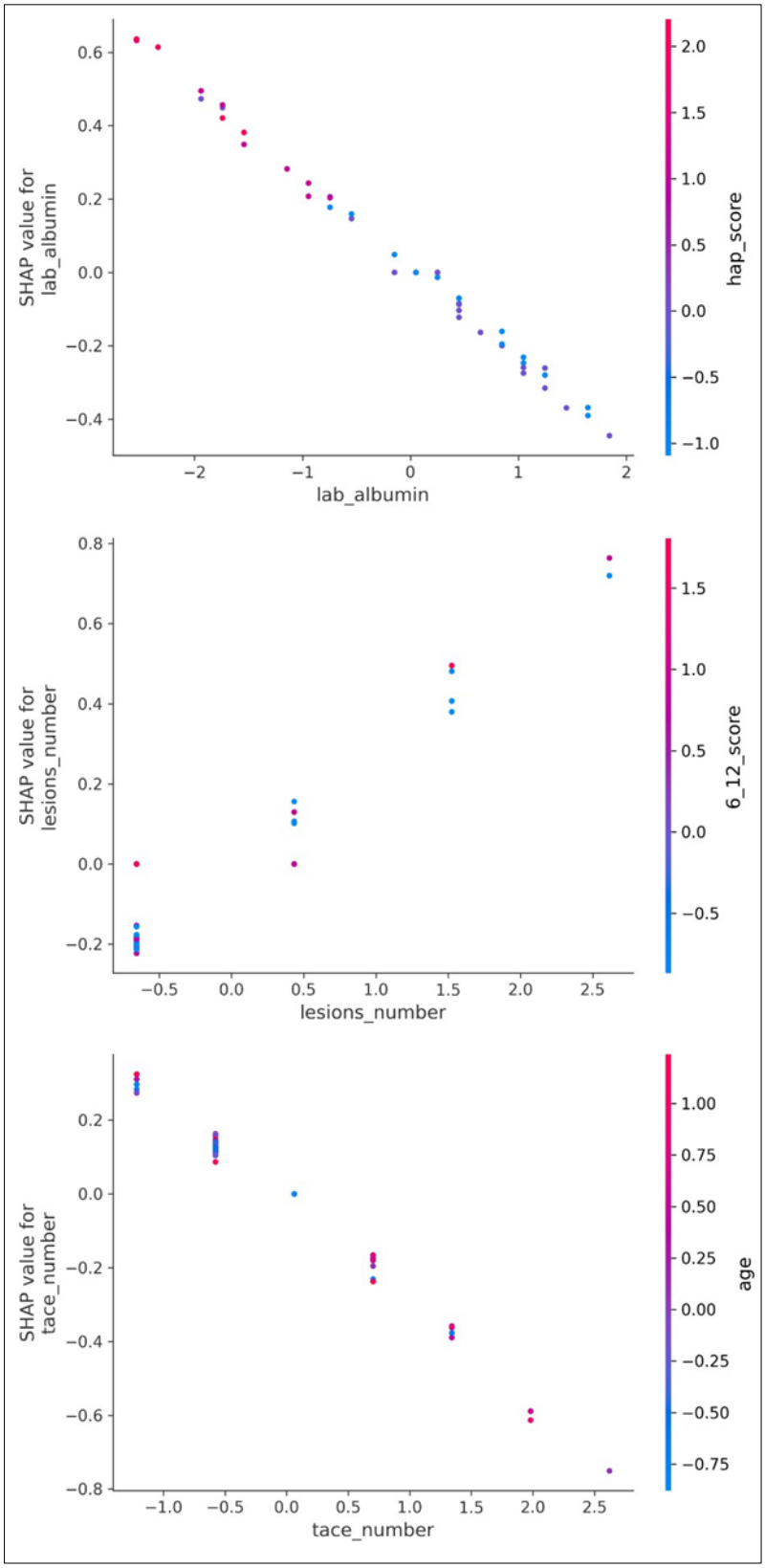

Fig 3: SHAP dependence plots. (A) Albumin with regard to different TACE sessions: lower albumin increases risk and sessions serve to amplify this association. (B) Albumin with respect to the number of lesions: risk grows in a near-linear fashion. (C) Female sex with respect to albumin: risk is elevated for women. (Not clinically actionable; requires further validation).

Figure 3 consolidates several SHAP dependence plots into panels A-C.

Panel A (serum albumin): Low levels of albumin and higher levels of SHAP positivity indicate increased risk. The positive relationship becomes stronger at lower levels of albumin for patients with more TACE sessions, indicating that the level of albumin interacts with the liver treatment and explains higher risk.

Panel B (Number of lesions): The risk of lower prognosis is positive and somewhat linear based on the number of lesions indicating that prognosis is worse for more advanced multiple lesion disease.

Panel C (Female): In this particular analysis, females are predicted to have worse prognosis, and the risk is higher with low albumin (blue) and requires external analysis to confirm and should not be used to justify treatment risk.

Individual Patient Explanation (Force Plot)

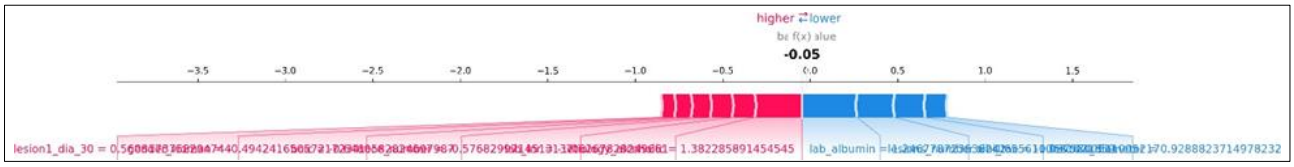

Fig 4: SHAP force plot for an individual test patient. Red bars increase predicted risk, blue bars decrease risk. Final prediction f(x) = –0.05 (near-average risk).

Figure 4 displays the force plot for a representative test patient. The baseline value (mean model output) is provided as -0.05. Factors that elevate predicted risk (represented by red bars) consist of lesion1_dia_30 (the largest lesion diameter) and low albumin. On the other hand, factors that are associated with risk reduction (blue bars) include higher albumin levels. Based on the model, the final prediction is f(x) = -0.05, which means that the patient is predicted to be at risk of having a health outcome that is near the average. This tool enables clinicians to understand which elements influenced the prediction.

Risk Stratification (Kaplan-Meier)

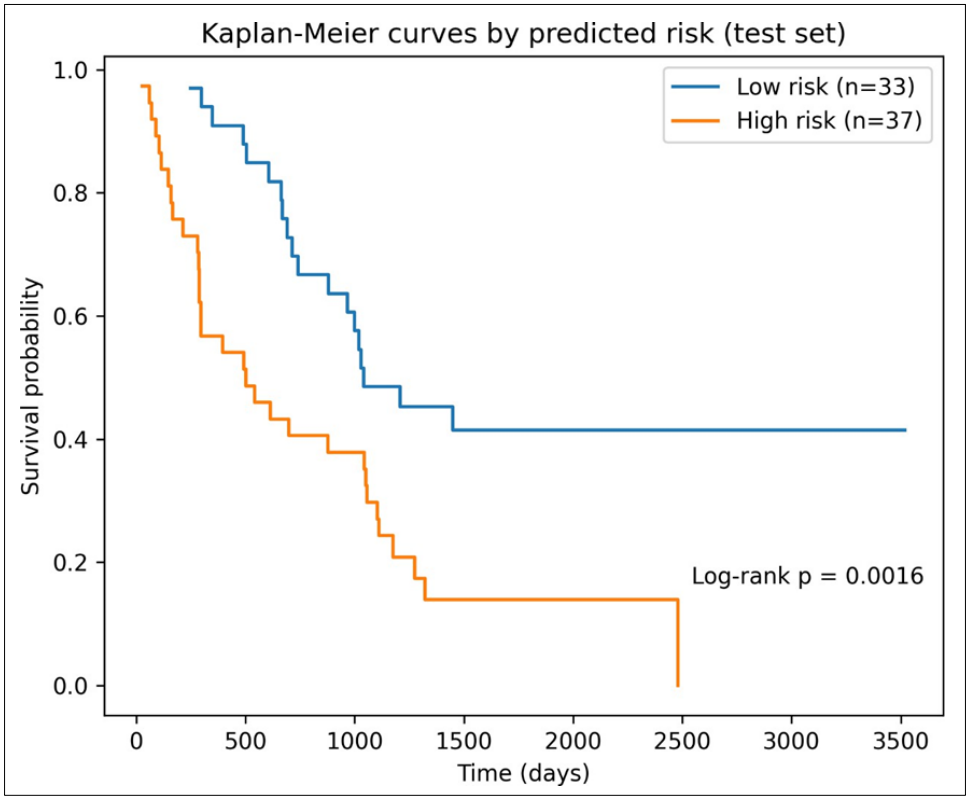

Fig 5: Kaplan-Meier survival curves for low-risk (n=33) and high-risk (n=37) groups defined by the median training risk score. Log-rank P = 0.0016.

Figure 5 This shows Kaplan Meier survival curves for low risk (n = 33) and high risk (n = 37) groups determined by the median risk score from the training set. The curves demonstrate strong separation with the high-risk group exhibiting poor survival across the entire follow up. The log rank test gives P = 0.0016 which shows significant stratification.

Hazard Ratios for Top Features

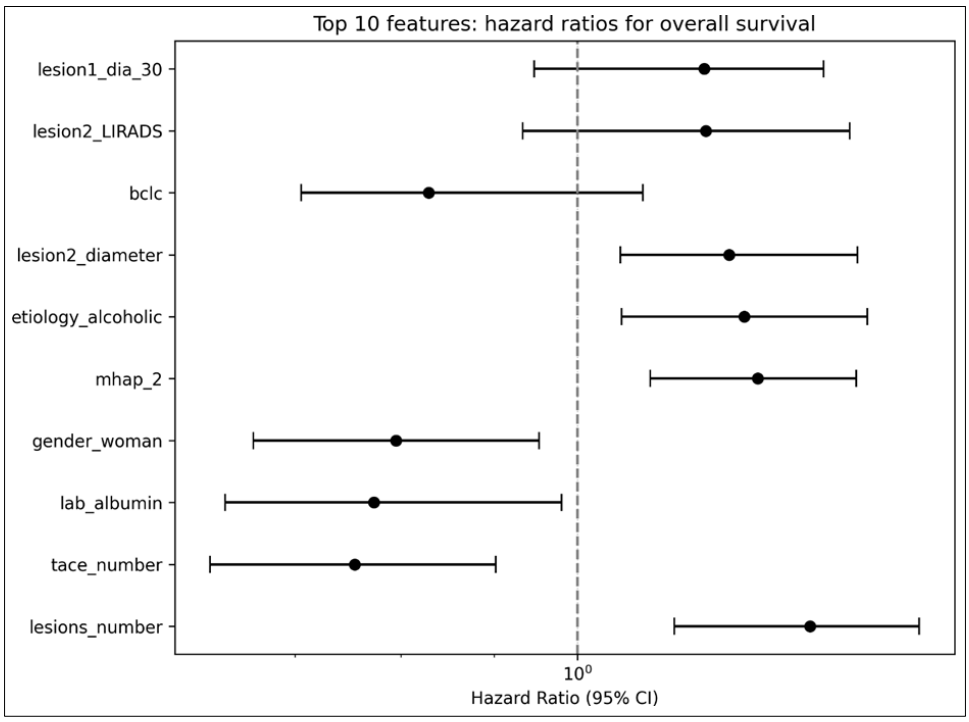

Fig 6: The hazard ratios and confidence intervals for each feature are shown in the forest plot. Note that hazard ratios were obtained using the results from 200 bootstrapped samples. The top 10 features were selected from the total number of features using the maximum difference in ranking, and have been visualized in the plot. Since all the continuous features for the model were centered and scaled to a mean of 0 with a standard deviation of 1, continuous features in HRs (effect size) measured the difference in log hazard for a one standard deviation (or one-unit) change (in this case, a change from 0 to 1) in a binary feature. Hence, HRs should be interpreted for relative importance and should not be cited for a change in a clinical CT). HRs should not be interpreted in a vacuum, as the continuous features for the model were scaled. For one of the features of this model, it should be noted that HRs for albumin reflect an approximate 8 g/L decrease in albumin, and not a change of 1 g/L. For this model, the top 10 feature HRs were included in Supplementary Table S1.

Figure 6 This is a forest plot of hazard ratios (HRs), with 95% confidence intervals (CI) based on 200 bootstraps, for the ten most influential features. Time to event data were modeled with Cox proportional hazard regression. Since all features were standardized, HRs illustrate the impact of a one standard deviation increase (or a 0 to 1 change for binary variables) of the feature on the log of the hazard. For example, the HR is about 2.4 when albumin is one standard deviation lower (about 8 g/L) (not shown in the figure; the figure displays standardized HRs). The most influential features, in order, are lesions number and tace number. All confidence intervals exclude the value of 1. Hence, these HRs should be viewed as relative importance measures rather than direct clinical effect sizes in the original units.

Discussion

An explainable model to predict the survival of HCC patients who receive TACE was constructed and validated using clinical variables and variance-reduced CT radiomics features. The model was compared to other approaches in the literature. Our model achieved a test C-index of 0.654, which is either better than or on par with other approaches (HAP 0.544, ALBI-TAE 0.624) in this dataset. Furthermore, our model was validated to have excellent short term predictive capability (6-month AUC 0.874, 12-month AUC 0.803). Medium- and short-term predictive capabilities are of high value clinically, especially in the case of 6-month prediction, which may enable early identification of patients who would be unlikely to benefit from repeated TACE, and as a result, ensure a timely switch to systemic therapy, which is in accordance with recent trials [29].

The primary contribution of this work is not an improvement in discrimination the improvement over a simple three-variable Cox model was small (C-index 0.654 vs. 0.642, difference = +0.012) but the construction of a clear, explainable framework for interpreting survival predictions at both the cohort and individual levels. We recognize that this difference is likely due to statistical variance, and we do not consider this difference may be clinically relevant in the absence of a C- index difference test (e.g., a permutation test). This is a limitation of our study. The main contribution of the additional model components, such as radiomics, is in the interpretation of the results via SHAP, and not in marginal discrimination improvement: they identify which tumor and hepatic characteristics are responsible for an individual prediction, a detail that is not possible with a three-variable clinical score. SHAP analysis offers three additional insights: conclusive interpretation of the results, identification of required features, and explanation of the prediction outcome. Global feature importance identified liver function (albumin), tumor burden (number of lesions), and, unexpectedly, female sex (which requires external validation; not clinically actionable), as the most important features. Albumin has traditionally been considered the most important element of the ALBI grade, and its dominance in this model again supports that the baseline liver functional reserve is the most important determinant of post–TACE survival [30].

Dependence plots also displayed a clinically reasonable insight where the danger of having low albumin worsened the

more TACE sessions the patient received. This indicates that the consequences of TACE worsen for patients with low albumin and/or hepatic dysfunction and aids the decision to restrict TACE sessions for such patients.

Individual force plots make an explanation at the patient level possible, and therefore is more likely to be adopted clinically. This allows a clinician to quickly determine whether a patient's predicted high risk is a result of low albumin, high lesion count, or something else, and base that information on their decision of how to approach the situation.

The unexpected finding of higher predicted risk being female deserves caution. In most HCC studies, males have higher incidence and worse outcomes. In the total cohort, 65/233 (27.9%) patients were female, and in the 70-patient test set, there were ~19 female patients, giving limited statistical sex related conclusions. Female patients in this dataset may have had different comorbidity profiles or etiologies that were not fully captured by the included features. A sensitivity analysis excluding sex from the model yielded a test C-index of 0.648, therefore the model's performance is not influenced by this variable. This finding should not be used for clinical decision making and must be assessed in a different setting before it can be interpreted.

Comparison with Prior Work

Multiple past studies have utilized machine learning techniques to assess the survival rate after undergoing TACE. For instance, one example is a Chinese multicentric study involving the use of time-series data that produced C-indices of approximately 0.65 to 0.70 [31]. Other studies have incorporated deep learning techniques directly on TB images. However, such models are often black-box and thus, hard to interpret [32]. Unlike these studies, our focus is on providing the explainability of the model, which serves an important role, alongside predictive accuracy. By providing model transparency, we hope to reduce the gap in AI studies pertaining to clinical practice, which happens to be a considerable obstacle [33].

Limitations

This study has several notable limitations:

- **Model development and external validation:** The model was developed and tested using a dataset from a single-center (Warsaw). For analysis of other aetiologias and CT protocols, Asian datasets should independently validate this model, and such an exercise is currently in progress.
- **Reliability of radiomics:** Radiomics are sensitive to the acquisition of and variations in segmentation. We did not perform test-retest or ICC-based filtering; our variance filtering removed features that remained constant and did not ensure reproducibility. Therefore, the reliability of the added radiomics should be explored.
- **Approximation of calibration:** Our calibration plot creates a binary event of dying within 12 months and does not factor in censoring, therefore bias in the observed proportion is likely to happen. Stricter methods of calibration will use IPC (inverse probability) of censored data (e.g. the pec or rms R package, or the scikit-survival utility for calibrating) and will be done for the next version of this study.
- **Decision Curve Analysis (DCA):** Decision Curve Analysis was not done in determining the net gain from the used thresholds. DCA is becoming more mandatory in reviews in order to illustrate clinical usefulness which goes beyond simple discrimination, and is a self-reported limitation of the study. Using the Python curves library, it is simple to perform DCA on the new version of the study.
- **Assumption of Proportional Hazards:** The assumption of proportional hazards was not fully addressed by Schoenfeld residuals. Landmark dependent survival, including early post-procedural death and late recurrences, is directly observed in many TACE patients. In this case, time- dependent hazard ratios are a real possibility. Ridge regularization is a partial solution that may reduce some of the concerns associated with proportionality due to the shrinkage of some high variance coefficients. In the case of true proportionality, ridge regularization does nothing to address the problem. Schoenfeld residuals for the predictors of interest and scaled Schoenfeld tests should be examined to address these concerns in the future.
- **Female Sex Association:** As indicated, the female sex finding may be an example of a research sampling bias, and therefore this should not be translated into clinical practice in the absence of further validation.
- **Censoring:** Censoring was at a 27% level, but competing risks were not investigated in this case.

Future Directions

An interactive web tool with real-time SHAP explanations will be created to publicly share the model. Future work will include external validation on an independent Asian cohort and addition of post TACE response predictors to improve long term prediction and execution of decision curve analysis.

Conclusion

A Cox-ridge model that combines clinical and radiomics features can explain and predict short- term survival after TACE for hepatocellular carcinoma (HCC), achieving an AUC of 0.874 for 6- month survival. The model employs SHAP, which provides both global and local interpretability, thus integrating machine learning with clinical practice. It can help clinicians determine which patients, in the absence of positive outcomes from TACE, should default to systemic therapy. The model will need to be tested in other datasets before it can be implemented in a clinical setting.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209-49.
2. Villanueva A. Hepatocellular carcinoma. N Engl J Med. 2019;380(15):1450-62.
3. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, *et al*. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115-32.
4. Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, *et al*. Hepatocellular carcinoma. Nat Rev Dis Primers. 2021;7(1):6.
5. Kudo M, Ueshima K, Ikeda M, Torimura T, Tanabe N, Aikata H, *et al*. Randomised, multicentre prospective trial of transarterial chemoembolisation (TACE) plus sorafenib as compared with TACE alone in patients with

hepatocellular carcinoma: TACTICS trial. Gut. 2020;69(8):1492-501.
6. Finn RS, Qin S, Ikeda M, Galle PR, Ducreux M, Kim TY, *et al*. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. N Engl J Med. 2020;382(20):1894-905.
7. Llovet JM, Singal AG, Villanueva A, Finn RS, Kudo M, Galle PR, *et al*. Prognostic and predictive factors in patients with advanced HCC and elevated alpha-fetoprotein treated with ramucirumab in two randomized phase III trials. Clin Cancer Res. 2022;28(11):2297-305.
8. Izzo C, Annunziata M, Melara G, Sciorio R, Dallio M, Masarone M, *et al*. The role of resveratrol in liver disease: a comprehensive review from *in vitro* to clinical trials. Nutrients. 2021;13(3):933.
9. Li G, Li X, Mahmud I, Ysaguirre J, Fekry B, Wang S, *et al*. Interfering with lipid metabolism through targeting CES1 sensitizes hepatocellular carcinoma for chemotherapy. JCI Insight. 2023;8(2):e163624.
10. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841-60.
11. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol. 2018;18(1):24.
12. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, *et al*. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364:l886.
13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
14. Lundberg S, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, *et al*., editors. Advances in Neural Information Processing Systems 30. New York (NY): Curran Associates Inc.; 2017. p. 4765-74.
15. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, *et al*. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.
16. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19(6):1236-46.
17. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, *et al*. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer. 2012;48(4):441-6.
18. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563-77.
19. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, *et al*. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
20. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, *et al*. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018;2(1):36.
21. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503-10.
22. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015;5:13087.
23. Pruinelli L, Balakrishnan K, Ma S, Li Z, Wall A, Lai JC, *et al*. Transforming liver transplant allocation with artificial intelligence and machine learning: a systematic review. BMC Med Inform Decis Mak. 2025;25(1):98.
24. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
25. Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc. 2007;102(478):527-37.
26. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York (NY): Chapman and Hall/CRC; 1993.
27. Molnar C. Interpretable machine learning. 2nd ed. Munich: Leanpub; 2022.
28. Molnar C, Freiesleben T, König G, Herbinger J, Reisinger T, Casalicchio G, *et al*. Relating the partial dependence plot and permutation feature importance to the data generating process. In: World Conference on Explainable Artificial Intelligence; 2023. Cham: Springer; 2023.
29. Llovet JM, Castet F, Heikenwalder M, Maini MK, Mazzaferro V, Pinato DJ, *et al*. Immunotherapies for hepatocellular carcinoma. Nat Rev Clin Oncol. 2022;19(3):151-72.
30. Johnson PJ, Berhane S, Kagebayashi C, Satomura S, Teng M, Reeves HL, *et al*. Assessment of liver function in patients with hepatocellular carcinoma: a new evidence-based approach—the ALBI grade. J Clin Oncol. 2015;33(6):550-8.
31. Zhu D, Ciais P, Krinner G, Maignan F, Jornet Puig A, Hugelius G. Controls of soil organic matter on soil thermal dynamics in the northern high latitudes. Nat Commun. 2019;10(1):3172.
32. Kocher M, Ruge MI, Galldiks N, Lohmann P. Applications of radiomics and machine learning for radiotherapy of malignant brain tumors. Strahlenther Onkol. 2020;196(10):856-67.
33. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195.

How to Cite This Article
Abbas Z, Muhammad I. Explainable Cox-Ridge survival modeling with SHAP for early risk stratification after transarterial chemoembolization in hepatocellular carcinoma: a multicenter analysis. Int J Med All Body Health Res. 2026;7(2):151-160.