





OPEN

## Comparative analysis of radiomics and deep-learning algorithms for survival prediction in hepatocellular carcinoma

Felix Schön<sup>1,6</sup>, Aaron Kieslich<sup>2,6</sup>, Heiner Nebelung<sup>1</sup>, Carina Riediger<sup>3</sup>, Ralf-Thorsten Hoffmann<sup>1</sup>, Alex Zwanenburg<sup>2,4,5</sup>, Steffen Löck<sup>2,7</sup> & Jens-Peter Kühn<sup>1,7</sup>

To examine the comparative robustness of computed tomography (CT)-based conventional radiomics and deep-learning convolutional neural networks (CNN) to predict overall survival (OS) in HCC patients. Retrospectively, 114 HCC patients with pretherapeutic CT of the liver were randomized into a development (n = 85) and a validation (n = 29) cohort, including patients of all tumor stages and several applied therapies. In addition to clinical parameters, image annotations of the liver parenchyma and of tumor findings on CT were available. Cox-regression based on radiomics features and CNN models were established and combined with clinical parameters to predict OS. Model performance was assessed using the concordance index (C-index). Log-rank tests were used to test model-based patient stratification into high/low-risk groups. The clinical Cox-regression model achieved the best validation performance for OS (C-index [95% confidence interval (CI)] 0.74 [0.57–0.86]) with a significant difference between the risk groups (p = 0.03). In image analysis, the CNN models (lowest C-index [CI] 0.63 [0.39–0.83]; highest C-index [CI] 0.71 [0.49–0.88]) were superior to the corresponding radiomics models (lowest C-index [CI] 0.51 [0.30–0.73]; highest C-index [CI] 0.66 [0.48–0.79]). A significant risk stratification was not possible (p > 0.05). Under clinical conditions, CNN-algorithms demonstrate superior prognostic potential to predict OS in HCC patients compared to conventional radiomics approaches and could therefore provide important information in the clinical setting, especially when clinical data is limited.

Hepatocellular carcinoma (HCC) is the most common primary malignant liver tumor, accounting for approximately 75% in total<sup>1</sup>. Overall, primary liver tumors are the second leading cause of cancer deaths worldwide with a 5-year survival rate of 18.1%<sup>2</sup>. The Barcelona Clinic Liver Cancer (BCLC) staging system is the most widely used algorithm in western countries to recommend prognostic prediction and first-line treatment based on tumor burden, liver function and health status of the patient<sup>3</sup>. Nevertheless, BCLC classification remains controversial and has limited predictive power<sup>4,5</sup>.

In recent years, the focus of medical research and clinical practice has shifted towards individualized medicine. Therefore, prediction of overall survival (OS) in HCC patients is of increasing importance to individually adapt potential therapy patterns and their influence on OS. Rapid advances in technology have made conventional, feature-based radiomics and deep-learning-based approaches particularly suitable for attaining these goals. Previous studies reported positive results for predicting OS of HCC patients using conventional radiomics and deep-learning algorithms<sup>6,7</sup>. Nevertheless, the suitability for clinical routine is questionable and despite the great potential of these technologies, a prospective transfer into clinical routine remains challenging. Patients may

<sup>1</sup>Institute and Polyclinic for Diagnostic and Interventional Radiology, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany. <sup>2</sup>OncoRay-National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden, Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany. <sup>3</sup>Department of Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany. <sup>4</sup>National Center for Tumor Diseases (NCT/UCC) Dresden, Dresden, Germany. <sup>5</sup>German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>6</sup>These authors contributed equally: Felix Schön and Aaron Kieslich. <sup>7</sup>These authors jointly supervised this work: Steffen Löck and Jens-Peter Kühn. ✉email: felix.schoen@uniklinikum-dresden.de; aaron.kieslich@oncoray.de

have received imaging for initial tumor staging in different medical centers, resulting in a large heterogeneity of acquisition parameters. Currently, there are no models for predicting OS across all HCC tumor stages and therapies. Moreover, to the best of our knowledge, it is uncertain which modelling approach, conventional radiomics or deep-learning-based approaches, is robust against the heterogeneity often encountered in clinical settings. There is evidence that conventional radiomics approaches seem more susceptible to interference, while deep-learning approaches might be more robust<sup>8</sup>. In our exploratory study, we therefore aimed to examine the comparative robustness of computed tomography (CT)-based conventional radiomics and deep-learning convolutional neural networks (CNN) algorithms to predict OS in HCC patients against two important sources of heterogeneity in real-world clinical settings: varied acquisition parameters and diverse tumor stages and treatments.

## Materials and methods

### Ethical aspects

The study was approved by the local ethics committee (EK 39012022) and conforms to the Declaration of Helsinki. The informed consent was waived by the ethics committee due to the retrospective nature of the study.

### Study population

A total of 343 patients with initial diagnosis of HCC were discussed between January 2010 and October 2021 in the tumor board of our University Hospital. Subsequently, patients were selected according to the following inclusion criteria:

(1) HCC patients who received a contrast-enhanced CT scan of the liver (consisting of at least an arterial and venous contrast phase) before therapy initiation; (2) the diagnosis of HCC had to be confirmed by a second imaging modality (e.g. ultrasound or magnetic resonance imaging) showing typical HCC changes, or by histopathological findings, according to the German HCC guideline<sup>9</sup>; (3) initial therapy and at least one follow-up imaging was carried out at our hospital.

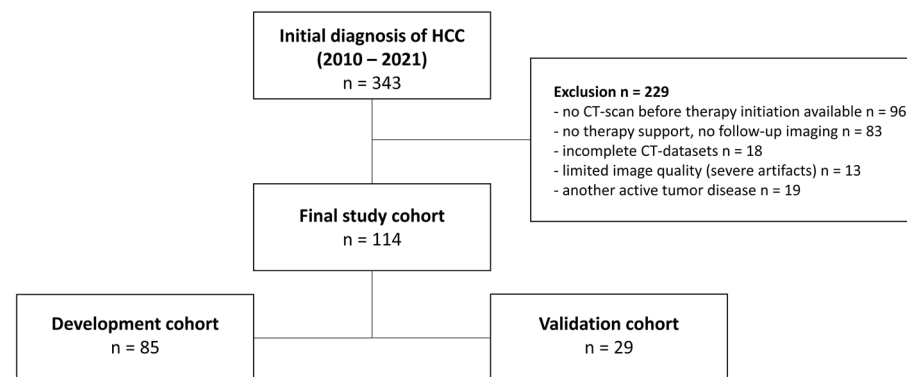
The exclusion criteria were: (1) incomplete CT scans (missing arterial or venous contrast phase); (2) CT scans with severe artifacts; (3) patients with another active tumor disease (defined as tumor diagnosis or therapy within 2 years prior to inclusion in the present study).

Based on these criteria (Fig. 1), a total of 114 patients were retrospectively enrolled and divided into a development (n = 85) and a validation (n = 29) cohort by stratified randomization, with stratification being performed on the initial therapy concept. Overall, the diagnosis of HCC was confirmed histopathologically in 60/114 patients, with the remaining 54/114 HCCs confirmed by imaging patterns.

### Clinical variables and radiological characteristics

Demographic data and routine lab tests were obtained for all patients. This included age, gender, time to death or follow-up time, as well as the serological parameters [alpha-fetoprotein (AFP), alanine-aminotransferase (ALAT), aspartate-aminotransferase (ASAT), albumin, total bilirubin, creatinine, gamma-glutamyltransferase (GGT) and International Normalized Ratio (INR)].

Tumor characteristics (number of lesions, presence of metastases, volume and density values of the largest HCC lesion), imaging features (status of liver cirrhosis and ascites) and initial therapy concepts [surgical resection (RES), radiofrequency ablation (RFA), liver transplantation (TRANS), transarterial chemoembolization (TACE), radiotherapy (RT), systemic therapy (ST) and best supportive care (BSC)] were recorded in addition. The ALBI-, Child-Pugh-, MELD-Score and the BCLC stage were calculated using the respective established formulae and flow charts<sup>3,10–13</sup>. The Child-Pugh Score was only evaluated in patients with suspected cirrhosis. Status of liver cirrhosis (present/absent) was assessed on CT by two residents (2 and 3 years of experience in liver imaging) analogously to Nebelung et al.<sup>14</sup> using the following criteria: hypertrophy of the liver segments I/II/III with concomitant atrophy of the segments VI/VII, surface and parenchymal nodularity of the liver, heterogeneous density values, portal vein enlargement, and ascites. Ascites was classified as absent, mild or moderate. In case of disagreement, the final decision was made by a senior radiologist with more than 15 years of experience



**Figure 1.** Study population. After applying inclusion and exclusion criteria, 114 patients were included and divided into a development and validation cohort by stratified randomization.

in liver imaging. Hepatic encephalopathy was not considered since its assessment is subjective and was not adequately documented.

### Overall survival

Overall survival was calculated as the period from the initial CT scan to the time of either death or last contact to our hospital (e.g. follow-up examination or discharge from inpatient stay).

### Imaging protocol and annotation

The CT scans were acquired on a total of 24 different scanners at 21 medical centers. Seventy-eight patients (68%) received their initial CT at our hospital. External CT scans of the remaining 36 patients (32%) were transmitted to our institution as part of routine clinical practice. Common contrast media methods (for arterial contrast: bolus tracking or approximately 25 to 35 s after contrast agent injection; for venous contrast: approximately 60 to 70 s after contrast agent injection) were applied for image acquisition. See Supplementary Table S1 for the variability of more scan parameters.

A resident with two years of experience in liver imaging contoured the liver parenchyma and the largest HCC-lesion in both contrast phases using the open-source software 3D Slicer (<http://www.slicer.org>)<sup>15</sup>. All segmentations were verified by the same resident 4 weeks later. An example of segmentation is shown in Supplementary Fig. S1.

### Standardization of the CT datasets

Variations in the circulatory capacity of patients, contrast medium injection parameters, and imaging time contribute to interindividual flood points of contrast agent<sup>16</sup>. The resulting differences across patients may influence the radiomics data derived from these scans<sup>17</sup>. To address this, a self-developed standardization procedure was performed.

For each patient, the mean CT number ( $CTN_{mean}$ ) in a circular segmentation within the aorta at the level of the coeliac trunk in the arterial and venous phase was recorded. The mean CTN for each phase ( $CTN_{mean,cohort}$ ) was used to scale the CTN of each patient ( $CTN_{mean,i}$ ) using the formula  $CTN_{new,i} = CTN_{old,i} \times \frac{CTN_{mean,cohort}}{CTN_{mean,i}}$ .

To compensate for different slice thicknesses, all CT images were interpolated to an isotropic voxel size of 1.0 mm<sup>3</sup>. An anti-aliasing filter was applied, and contours were re-segmented to density values between – 200 and 500 Hounsfield units (HU). Details see Supplementary Table S2.

### Conventional radiomics risk modelling

Radiomics features were extracted from the segmentations of the liver and HCC in the arterial (*\_art*) and venous (*\_ven*) phase. The extraction was implemented according to the recommendations by the Image Biomarker Standardization Initiative (IBSI) using the publicly available open-source Medical Image Radiomics Processor (MIRP)<sup>18,19</sup>. Feature values obtained from the venous phase images were subtracted from the corresponding arterial phase values to quantify differences between both phases (*\_diff*). In summary, six feature subgroups were extracted (*HCC\_art*, *HCC\_ven*, *HCC\_diff*, *liver\_art*, *liver\_ven*, *liver\_diff*), resulting in 1146 imaging features per patient.

To develop conventional radiomics models, the “Fully Automated Machine Learning with Interpretable Analysis of Results” (FAMILIAR, version 1.2.0) framework (<https://github.com/alexzwanenburg/familiar>) was used<sup>20</sup>. The utilized settings for feature extraction and model building can be found in Supplementary Table S3. Three primary models were constructed to predict OS, consisting of a clinical model, an image-based radiomics model and a combined model of clinical and imaging data. Four supplementary models analyzing the imaging segmentations separately were additionally created to compare the predictive power for OS in the different contrast phases and imaging components (whole liver parenchyma vs. HCC).

For each model, feature importance was evaluated using a 15-times repeated threefold cross-validation scheme, resulting in 45 internal models in total. In each iteration, multiple feature processing steps were applied: missing value imputation, feature transformation, filtering and clustering. The overall importance of a feature was assessed by its occurrence within the top five highest ranked features in all 45 internal models. The signature size was assigned as the median signature size of all 45 internal models. The features with the highest importance were used to create a Cox proportional hazards model for the prediction of OS. Subsequently, the models were validated on the validation cohort. Details of feature processing and model development are given in Supplementary Table S3.

### Deep-learning-based risk modelling

All segmentations of the liver and the HCC were considered. To accommodate for the large range of sizes observed for liver and HCC lesions across the image datasets, a cropping procedure was applied: all images were cropped to the 95th percentile of the distribution of liver or HCC sizes in each direction. In addition, all images were resampled to a voxel size of 2 × 2 × 2 mm<sup>3</sup>. The resulting image dimensions were 64 × 64 × 64 and 132 × 144 × 132 voxels for the HCC and the liver segmentations, respectively. The voxel intensities were rescaled to the interval [0, 1]. To avoid overfitting on characteristics outside of the ROI, these regions were masked by setting all voxel intensities outside the ROI to zero.

For clinical data, missing values were imputed using the median value over all patients. If the percentage of missing features for one patient exceeded 30%, the patient was excluded. All clinical data were converted to a numerical scale. The features were transformed using Yeo-Johnson normalization and Z-standardization and

mapped linearly to the interval 0 to 1 based on the development cohort. Transformation parameters were applied to the validation cohort unchanged.

Four primary deep-learning-based models were constructed to predict OS, consisting of a clinical model, two image-based models (based on the HCC and liver segmentations) and a combined model of clinical and HCC imaging data. As deep-learning algorithms require significantly more computing power, it was not possible to create an imaging model consisting of all CT data as in the conventional radiomics approach. Four supplementary models analyzing the imaging segmentations separately were additionally created to compare the predictive power for OS in the different contrast phases and imaging components (whole liver parenchyma vs. HCC).

All models were implemented using the Python-based deep-learning library Pytorch<sup>21</sup>. The general architecture of the networks was designed after Hosny et al., Starke et al., and Nie et al. and is illustrated in Supplementary Fig. S2<sup>22–24</sup>. For example, the proposed image-based model consists of four convolutional layers and three fully connected layers. To regulate the model, batch normalization and dropout layers are incorporated. CT images of both arterial and venous phases form the input of the model. They are processed through the convolutional layers before being concatenated and further processed by the fully connected layers. According to Katzman et al., the loss function is set to the negative log of the Cox partial likelihood with regularizations<sup>25</sup>. Therefore, the final output is a single value representing the predicted hazard of the model. Details regarding the utilized hyperparameters can be found in Supplementary Table S4.

The number of training epochs was determined through a 15-times repeated threefold cross-validation, resulting in 45 internal models in total. Each model was trained for 500 epochs on the training fold and monitored for testing fold performance after every epoch. Model performance was assessed by the average performance of the last five epochs to reduce statistical noise. Finally, for validation, 45 models were trained on the entire development cohort using the number of epochs with the highest cross-validation performance. The final prediction for a patient was established by taking the average prediction of all 45 models.

### Evaluation of prognostic performance

Prognostic performance was evaluated by the concordance index (C-index) and the ability to stratify patients into risk groups based on the model predictions. The C-index measures the agreement between the actual OS and the model predictions. A C-index of 0.5 indicates no prognostic value, while a value close to 1 indicates perfect prediction. Patients were allocated into a low- or high-risk group for death based on the hazard values predicted by the models. The median value of these predictions was used as a cutoff on the development cohort. Patients with a predicted hazard exceeding the cutoff were assigned to the high-risk group. The difference between the low- and high-risk group was assessed using the log-rank test. The significance level was set to  $\alpha = 0.05$ . The confidence intervals (CI) for the internal cross-validation were calculated by analyzing the distribution of the 45 model performances. To estimate the CIs for the validation, the percentile bootstrap method was performed<sup>26</sup>. To compare the prognostic performance of two models, a two-sample bootstrap test was employed: The difference in C-indices was computed for 1000 bootstrap samples of the validation cohort. The smaller proportion of bootstrap samples in which the C-index difference was either greater than 0 or less than 0 was multiplied by 2 to obtain the p-value.

## Results

### Study population

Development and validation cohort were balanced in terms of clinical parameters and baseline demographics ( $p > 0.05$ ; Table 1).

### Conventional radiomics approach

Three primary models were developed: a clinical model, a radiomics model including all imaging features and a model combining clinical and imaging data. In addition, four supplementary radiomics models (*HCC\_art*; *HCC\_ven*; *liver\_art*; *liver\_ven*) were developed based on the individual image segmentations.

The median signature sizes were three, six and seven for the clinical, image-based, and combined model, respectively, and ranged between two and five for the supplementary models. For the clinical model, six patients (four in development cohort and two in validation cohort) were excluded from the analysis due to missing values  $> 30\%$ . The final Cox-regression models are reported in Table 2 for the primary analyses and in Supplementary Table S5 for the supplementary analyses. The results of the internal cross-validation are shown in Supplementary Table S6.

In independent validation (Table 3), the clinical model showed the best result with a C-index [95% CI] of 0.74 [0.57–0.86], outperforming the image-based and combined model significantly ( $p = 0.016$  and  $p = 0.034$ , respectively). The best supplementary model (*HCC\_art*) clearly outperformed the primary image-based model, which showed a performance close to random prediction (C-index [95% CI] 0.66 [0.48–0.79] vs. C-index [95% CI] 0.51 [0.30–0.73]). The risk stratification into high- and low-risk groups showed significant differences in OS only for the clinical model ( $p = 0.031$ ; Fig. 2).

### Deep-learning approach

Four primary models were developed: a clinical model, two image-based models (*HCC\_art* + *HCC\_ven* and *liver\_art* + *liver\_ven*) and a model combining clinical data and HCC imaging data. Additionally, four supplementary image-based models were established (*HCC\_art*; *HCC\_ven*; *liver\_art*; *liver\_ven*). The results of the internal cross-validation for all primary and supplementary models are shown in Supplementary Table S7. The obtained number of epochs was 430, 39, 14 and 383 for the clinical model, the HCC-based model, the liver-based model and the combined model, respectively.

Variable	Development cohort (n = 85)			Validation cohort (n = 29)			p-value
	Median	Range	Missing (%)	Median	Range	Missing (%)	
Time to death of dead patients, years	1.65	0.01–6.20	n/a	1.83	0.29–3.70	n/a	0.36
Follow up time of patients alive, years	1.63	0.32–9.54	n/a	1.99	0.41–9.53	n/a	0.62
Age, years	71.11	48.17–82.12	0 (0)	70.08	39.60–84.27	0 (0)	0.84
AFP, ng/ml	8.6	0.9–707,760.0	10 (12)	5.6	1.0–22,169.6	3 (10)	0.15
ALAT, $\mu\text{mol/s L}$	0.58	0.19–4.89	3 (4)	0.57	0.27–1.78	1 (3)	0.47
ASAT, $\mu\text{mol/s L}$	0.78	0.19–3.17	5 (6)	0.71	0.38–3.33	1 (3)	0.12
GGT, $\mu\text{mol/s L}$	3.04	0.29–28.17	3 (4)	2.28	0.53–14.58	1 (3)	0.40
Albumin, g/L	39.9	22.7–48.2	7 (8)	39.9	29.3–47.0	4 (14)	0.68
Bilirubin, $\mu\text{mol/L}$	15.2	4.5–122.1	5 (6)	12.3	2.5–81.6	1 (3)	0.37
INR	1.16	0.91–2.91	4 (5)	1.14	0.93–3.15	2 (7)	0.56
Creatinine, $\mu\text{mol/L}$	79	45–162	3 (4)	87	53–141	1 (3)	0.28
MELD Score	10	7–20	7 (8)	10	7–21	2 (7)	0.88
HCC Volume arterial phase, $\text{cm}^3$	18.9	0.6–1339.2	0 (0)	12.8	0.8–1473.9	0 (0)	0.83
HCC Volume venous phase, $\text{cm}^3$	18.3	0.5–1246.8	0 (0)	14.0	1.0–2000.8	0 (0)	0.88
HCC Mean CTN arterial phase, HU	69.5	32.1–154.2	0 (0)	63.1	35.9–111.1	0 (0)	0.37
HCC Mean CTN venous phase, HU	74.9	34.8–124.6	0 (0)	74.5	43.1–115.5	0 (0)	0.49
HCC Variance CTN arterial phase, $\text{HU}^2$	365	84–1881	0 (0)	358	127–1748	0 (0)	0.98
HCC Variance CTN venous phase, $\text{HU}^2$	318	103–974	0 (0)	292	123–714	0 (0)	0.74
	Number of patients (%)		Missing (%)	Number of patients (%)		Missing (%)	
Patients lost in follow-up	40 (47)		n/a	17 (59)		n/a	0.45
Died from treatment-related causes	1 (1)		n/a	1 (3)		n/a	0.43
Sex, male/female	67/18 (79/21)		0 (0)	27/2 (93/7)		0 (0)	0.14
Therapy concept, curative/palliative	49/36 (58/42)		0 (0)	17/12 (59/41)		0 (0)	1.00
Therapy, BSC/ST/RES/RFA/RT/TACE/TRANS	5/8/25/26/6/14/1 (6/9/29/31/7/16/1)		0 (0)	2/3/8/9/2/5/0 (7/10/28/31/7/17/0)		0 (0)	1.00
Liver Cirrhosis, present/absent	60/25 (71/29)		0 (0)	21/8 (72/28)		0 (0)	1.00
Ascites, absent/mild/moderate	61/13/11 (72/15/13)		0 (0)	19/8/2 (66/28/7)		0 (0)	0.34
Child–Pugh score, 0/A/B/C	25/35/15/3 (29/31/18/4)		7 (8)	8/12/5/1 (28/41/17/3)		3 (10)	1.00
BCLC stage, A/B/C/D	29/36/17/3 (34/42/20/4)		0 (0)	12/6/10/1 (41/21/34/3)		0 (0)	0.17
ALBI score, 1/2/3	44/30/4 (52/35/5)		7 (8)	15/10/0 (52/34/0)		4 (14)	0.51
Number of HCC lesions, 1/2/3/ > 3	53/13/3/16 (62/15/4/19)		0 (0)	17/5/4/3 (59/17/14/10)		0 (0)	0.19
Lymph nodes metastases, present/absent	1/84 (1/99)		0 (0)	1/28 (3/97)		0 (0)	1.00
Distant metastases, present/absent	2/83 (2/98)		0 (0)	2/27 (7/93)		0 (0)	0.57

**Table 1.** Patient characteristics of the development and validation cohort. The variables describing the CTN, and volume refer to the largest HCC lesion. P-values were obtained by using Chi-square homogeneity tests and two-sided Mann–Whitney *U* tests for categorical and numerical variables, respectively. *AFP* alpha-fetoprotein, *ALAT* alanine-aminotransferase, *ASAT* aspartate-aminotransferase, *BCLC* Barcelona clinic liver cancer; *BSC* best supportive care, *CTN* computed tomography number, *GGT* gamma-glutamyltransferase, *HCC* hepatocellular carcinoma, *INR* international normalized ratio, *RES* surgical resection, *RFA* radiofrequency ablation, *RT* radiotherapy, *ST* systemic therapy, *TACE* transarterial chemoembolization, *TRANS* liver transplantation.

In validation (Table 4), the image-based HCC model showed the best performance of all primary models (C-index [95% CI] 0.69 [0.44–0.84]), while the clinical model was the worst (C-index [95% CI] 0.58 [0.40–0.76]). Overall, the supplementary image-based model *HCC\_art* performed even slightly better than the primary HCC model (C-index [95% CI] 0.71 [0.49–0.88]). Risk stratification in groups at high and low risk of death revealed no significant results in the validation cohort ( $p > 0.05$ ).

### Comparison of the conventional radiomics and the deep-learning approach

The deep-learning approach outperformed the conventional radiomics approach, with a significant improvement for the *liver\_ven* model ( $p = 0.032$ ). Figure 3 highlights the differences of the C-indices in the validation cohort between the conventional radiomics and deep-learning models. Figure S3 shows the calibration plots of the best performing image-based models from both the conventional radiomics and deep-learning approaches.

Primary conventional models	Variables	Hazard ratio [95% CI]	p-value
Clinical model	GGT	1.61 [1.15–2.27]	0.01*
	AFP	1.24 [0.92–1.67]	0.15
	volume HCC_ven	1.17 [0.81–1.68]	0.41
Image-based model	stat_rms_liver_ven	0.72 [0.52–0.99]	0.05
	szm_szhge_3d_fbn_n32_liver_diff	1.71 [1.17–2.52]	0.01*
	morph_moran_i_liver_diff	0.69 [0.49–0.96]	0.03*
	dzm_lde_3d_fbn_n32_liver_diff	1.26 [0.90–1.78]	0.18
	morph_pca_least_axis_liver_art	1.66 [1.13–2.45]	0.01*
	morph_pca_least_axis_liver_diff	1.31 [0.92–1.87]	0.14
Combined model	GGT	1.58 [1.11–2.23]	0.01*
	stat_p90_liver_ven	0.88 [0.61–1.28]	0.51
	szm_szhge_3d_fbn_n32_liver_diff	1.70 [1.11–2.60]	0.02*
	dzm_lde_3d_fbn_n32_liver_diff	1.18 [0.83–1.67]	0.35
	stat_max_HCC_diff	1.30 [0.92–1.83]	0.14
	ih_min_grad_fbn_n32_liver_ven	0.92 [0.62–1.36]	0.56
	morph_pca_least_axis_liver_diff	1.17 [0.80–1.72]	0.42

**Table 2.** Final signatures of the primary clinical, image-based, and combined multivariate Cox-regression models and their respective parameters. The hazard ratio (HR) [95% CI] and the corresponding p-values of the regression are shown based on the development cohort. *AFP* alpha-fetoprotein, *art* arterial phase, *diff* difference, *CI* confidence interval, *GGT* gamma-glutamyltransferase, *HCC* hepatocellular carcinoma, *ven* venous phase. \*Statistically significant ( $p < 0.05$ ).

		Development cohort		Validation cohort	
		C-index [95% CI]	p-value	C-index [95% CI]	p-value
Primary conventional models	Clinical model	0.69 [0.58–0.78]	0.045*	0.74 [0.57–0.86]	0.031*
	Image-based model	0.75 [0.68–0.81]	<0.001*	0.51 [0.30–0.73]	0.66
	Combined model	0.75 [0.65–0.83]	<0.001*	0.55 [0.40–0.69]	0.89
Supplementary conventional models	<i>HCC_art</i>	0.63 [0.53–0.73]	0.58	0.66 [0.48–0.79]	0.69
	<i>HCC_ven</i>	0.62 [0.52–0.71]	0.24	0.53 [0.36–0.71]	0.57
	<i>liver_art</i>	0.65 [0.54–0.76]	0.14	0.54 [0.30–0.74]	0.46
	<i>liver_ven</i>	0.71 [0.61–0.80]	0.013*	0.46 [0.23–0.67]	0.62

**Table 3.** Final performance of the Cox-regression models in development and independent validation: C-indices [95% CI] and p-values for risk stratification. *art* arterial phase, *CI* confidence interval, *HCC* hepatocellular carcinoma, *ven* venous phase. \*Statistically significant ( $p < 0.05$ ).

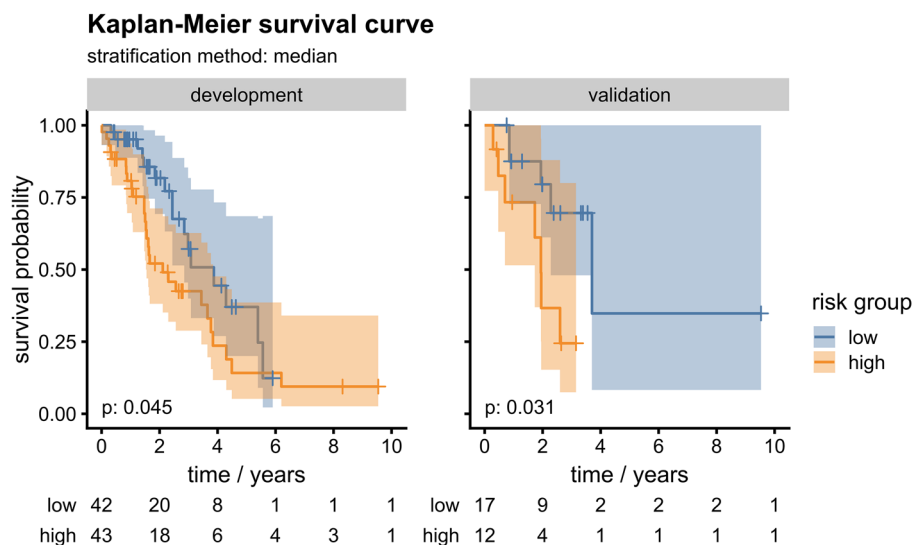
## Discussion

In the present study, we investigated whether conventional radiomics and deep-learning algorithms can predict OS in HCC patients based on CT data regardless of tumor stage or applied therapy and compared both methods for superiority. Overall, deep-learning algorithms outperformed conventional radiomics features and could help to predict OS. Still, the clinical Cox-regression model showed the best performance in the presented setting.

To the best of our knowledge, our study is the first radiomics analysis of CT scans for OS of HCC patients across all tumor stages and common therapies. Previous studies have focused on specific therapies or tumor stages. In addition, analyses were often based on only one contrast phase and rarely used the combination of HCC and liver parenchyma.

To date, the predictive power of deep-learning algorithms on CT images for predicting OS of HCC patients has not been comprehensively evaluated. Wang et al. reported a C-index of 0.58 for patients undergoing stereotactic radiotherapy<sup>27</sup>. Better results were observed in patients who received a TACE alone (C-indices = 0.65 and 0.73) or a combination of TACE and sorafenib (C-index = 0.72)<sup>28–30</sup>. The C-indices of 0.63–0.71 obtained in the present study are in line with the listed values and thus show the potential for outcome prediction even in HCC patients receiving different therapeutic approaches, although no significant risk stratification was possible.

Conventional radiomics models for predicting OS of HCC patients have been evaluated more commonly so far. C-indices from literature range between 0.63–0.78 and 0.60–0.67 for HCC patients undergoing surgery or TACE, respectively<sup>31–35</sup>. However, validation on holdout or external datasets was not always performed and risk stratification was not always possible. Here, the majority of our conventional radiomics models showed a performance close to random prediction. With a C-index of 0.66, only the best model (*HCC\_art*) showed a value comparable to the literature.

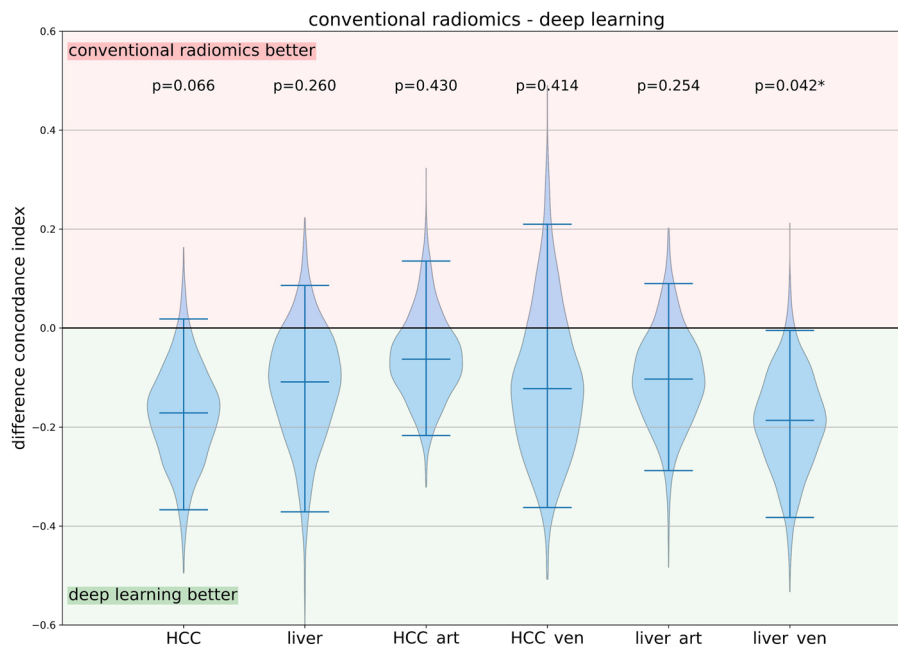


**Figure 2.** Kaplan–Meier survival curves of patients stratified into risk groups (cutoff value = 1.024 years) by the clinical model in the development and validation cohort. Differences in OS between low- and high-risk groups were evaluated by the log-rank test.

		Development cohort		Validation cohort	
		C-index [95% CI]	p-value	C-index [95% CI]	p-value
Primary deep-learning models	Clinical model	0.74 [0.69–0.81]	< 0.001*	0.58 [0.40–0.76]	0.92
	Image-based model HCC ( <i>HCC_art</i> + <i>HCC_ven</i> )	0.60 [0.50–0.69]	0.45	0.69 [0.44–0.84]	0.42
	Image-based model liver ( <i>liver_art</i> + <i>liver_ven</i> )	0.72 [0.63–0.80]	< 0.001*	0.65 [0.37–0.86]	0.40
	Combined model	0.65 [0.57–0.72]	0.029*	0.62 [0.41–0.81]	0.18
Supplementary deep-learning models	<i>HCC_art</i>	0.60 [0.50–0.70]	0.54	0.71 [0.49–0.88]	0.42
	<i>HCC_ven</i>	0.66 [0.57–0.75]	0.078	0.63 [0.39–0.83]	0.18
	<i>liver_art</i>	0.73 [0.62–0.81]	< 0.001*	0.63 [0.43–0.79]	0.86
	<i>liver_ven</i>	0.68 [0.58–0.80]	0.037*	0.65 [0.39–0.80]	0.17

**Table 4.** Final performance of the deep-learning-based models in development and independent validation: C-Indices [95% CI] and p-values for risk stratification. *art* arterial phase, *CI* confidence interval, *HCC* hepatocellular carcinoma, *ven* venous phase. \*Statistically significant ( $p < 0.05$ ).

Each deep-learning image-based model outperformed its conventional radiomics counterpart with statistically significance for the venous liver model (*liver\_ven*), leading us to the conclusion that deep-learning may offer an enhanced prognostic utility. The main reason for the limited performance of the conventional radiomics approach may be the heterogeneous study cohort. Previous studies reported lack of reproducibility of hand-crafted radiomics features between different CT scanners, acquisition and reconstruction parameters<sup>8,36–39</sup>. As we used CT data from 24 different scanners, acquisition parameters were heterogeneous, which may have negatively affected reproducibility of radiomics features. In contrast, features extracted from deep-learning may be more robust<sup>40</sup>. This observation aligns with findings from a comparative study on head and neck cancer OS prediction, which demonstrated that deep learning models exhibit superior generalizability across different institutions compared to conventional radiomics approaches<sup>41</sup>. Overall, the clinical model based on Cox-regression was superior to all imaging approaches with significantly different OS between the stratified risk groups suggesting a high importance of clinical factors for generalized prediction models. The parameters of the final signature, consisting of GGT, AFP and the HCC volume, have a known impact on the prognosis of affected patients. Elevated GGT levels may indicate liver damage, such as chronic hepatic parenchymal remodeling or HCC<sup>42</sup>, and are associated with OS in HCC<sup>43,44</sup>. AFP is the most common serum marker in HCC, with higher AFP levels associated with poorer OS<sup>45,46</sup>. HCC volume is associated with tumor malignancy and infiltrative behavior<sup>47</sup>. Wu et al. point out that tumor size at diagnosis is an independent prognostic factor for OS, irrespective of tumor grade, stage, or treatment selected<sup>48</sup>. Our results support these findings. However, other parameters, such as the MELD-Score were not included in the clinical model. Although this factor has been identified as predictor of HCC prognosis<sup>49</sup>, within the scope of our multi-step machine learning workflow and the heterogeneity of our patient cohort, the inclusion of additional clinical parameters did not give benefit to the predictive performance



**Figure 3.** Comparison of C-indices between the conventional radiomics and the deep-learning models in the validation cohort. Positive values indicate better performance of the conventional approach, whereas negative values indicate better performance of the deep-learning approach. The whiskers represent the 95% confidence interval. The horizontal line within the distributions illustrates the median value. For the comparison of the HCC and liver models, the primary radiomics model was used. \*Statistically significant ( $p < 0.05$ ).

and its generalizability. As lack of clinical data continues to be a non-negligible problem in patient care in some cases, the development of image parameter-based deep-learning and conventional radiomics models is essential. Therefore, further studies in larger patient groups are essential to further explore the comparative potential between image-based algorithms and clinical models.

Interestingly, the clinical deep-learning model was outperformed by the clinical Cox-regression model, although the difference was not statistically significant. One potential explanation for this finding is that the clinical deep-learning model may not have been fully optimized. Specifically, overfitting on the development cohort was observed, which suggests that further refinement and optimization of the model hyperparameters may lead to a better performance on future datasets.

The complexity of deep-learning models raises questions about their value for OS prediction. To increase the value of deep-learning models for OS prediction, the models should be interpretable and easily to understand and rely on for physicians. To improve the interpretability of the deep-learning models, their output was correlated with the conventional radiomics features. For the HCC model, a high correlation with the HCC volume was observed (Spearman  $R = 0.94$ ), indicating that an increase in HCC volume corresponds to lower OS. This finding is consistent with the results of the clinical Cox-regression model of this study. Similarly, predictions of the liver model were associated with liver volume ( $R = 0.86$ ), suggesting that an increase in liver volume corresponds to lower OS. This finding contradicts the expectation that progressing cirrhosis is associated with a decreasing liver volume leading to lower OS<sup>50</sup>. Future research should aim to better understand the complex patterns that deep-learning algorithms can detect.

There are some limitations of this study. First, it was a retrospective study with a small sample size of 114 patients with limited follow-up duration. Especially for CNNs, the limited sample size increases the risk of overfitting and reduces result reliability. However, multiple strategies were employed to minimize the risk of overfitting despite the limited sample size: (i) early stopping of the training process using cross-validation; (ii) masking the CT image to the ROI to prevent overfitting on surrounding anatomical structures; (iii) architectural considerations like batch normalization, the use of dropout layers and pooling layers and data augmentation; (iv) ensemble prediction by averaging the output from 45 individually trained models for the final prediction. The small resulting differences of CNN performances between the development and validation cohort suggest that overfitting is unlikely, despite the limited sample size. Second, the study population has high heterogeneity in terms of various factors such as applied treatment, CT acquisition protocols, and HCC tumor stages. While this heterogeneous group reflects everyday clinical practice, it may also limit the generalizability of the findings, as the specific distribution of clinical characteristics and treatment approaches may differ significantly across clinics. In addition, 1/85 and 1/29 patients in the development and validation cohort, respectively, were expected to die from treatment-related causes. Due to this minority, the impact on the developed models can be considered negligible. Third, not all clinical parameters were available and only the largest lesion was segmented in multifocal HCC. Whole tumor burden analysis may improve the efficiency of OS prediction, although all HCC lesions



were included as the entire liver parenchyma was additionally segmented. Fourth, in the deep-learning models, the analysis was conducted on the masked CT images, which may introduce bias towards the volume of the regarded ROI and could potentially exclude areas with prognostic value in the peritumoral region<sup>51</sup>. However, the utilization of masked images may also result in the reduction of background information. This compels the model to prioritize important areas within the image. As a result, this approach has the potential to enhance both the reliability and quality of the model's outputs<sup>52</sup>. Moreover, this approach ensures that the conventional radiomics and deep-learning models are based on similar data, which enables a more equitable comparison. Fifth, all segmentations were derived manually by a single radiologist and subjective bias cannot be excluded. Therefore, further studies on a large sample size are needed to increase the reliability of the results.

In conclusion, deep-learning algorithms showed superiority over conventional radiomics for predicting OS in patients with HCC across a wide spectrum of therapies, tumor stages and CT acquisition protocols. In total, they showed comparable performance to previously presented models, which were, however, adjusted to therapy subgroups. The results advocate the development of deep-learning models in the clinical prognosis of HCC patient survival on a larger scale and may provide important information in the clinical setting, especially when clinical data is limited.

## Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Received: 16 July 2023; Accepted: 20 December 2023

Published online: 05 January 2024

## References

- Altekruse, S. F., Devesa, S. S., Dickie, L. A., McGlynn, K. A. & Kleiner, D. E. Histological classification of liver and intrahepatic bile duct cancers in SEER registries. *J. Regist. Manag.* **38**, 201–205 (2011).
- Jemal, A. *et al.* Annual report to the nation on the status of cancer, 1975–2014 featuring survival. *J. Natl. Cancer Inst.* **109**, djx030 (2017).
- Reig, M. *et al.* BCLC strategy for prognosis prediction and treatment recommendation: The 2022 update. *J. Hepatol.* **76**, 681–693 (2022).
- Vitale, A. *et al.* Survival benefit of liver resection for patients with hepatocellular carcinoma across different Barcelona clinic liver cancer stages: A multicentre study. *J. Hepatol.* **62**, 617–624 (2015).
- Huitzil-Melendez, F.-D. *et al.* Advanced hepatocellular carcinoma: Which staging systems best predict prognosis?. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **28**, 2889–2895 (2010).
- Wakabayashi, T. *et al.* Radiomics in hepatocellular carcinoma: A quantitative review. *Hepatol. Int.* **13**, 546–559 (2019).
- Ahn, J. C., Qureshi, T. A., Singal, A. G., Li, D. & Yang, J.-D. Deep learning in hepatocellular carcinoma: Current status and future perspectives. *World J. Hepatol.* **13**, 2039–2051 (2021).
- Denzler, S. *et al.* Impact of CT convolution kernel on robustness of radiomic features for different lung diseases and tissue types. *Br. J. Radiol.* **94**, 20200947 (2021).
- Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF): Diagnostik und Therapie des Hepatozellulären Karzinoms und biliärer Karzinome, Langversion 4.0, 2023, AWMF-Registernummer: 032–053OL. <https://www.leitlinienprogramm-onkologie.de/leitlinien/hcc-und-biliaere-karzinome>.
- Wiesner, R. H. *et al.* MELD and PELD: Application of survival models to liver allocation. *Liver Transplant. Off. Publ. Am. Assoc. Study Liver Dis. Int. Liver Transplant. Soc.* **7**, 567–580 (2001).
- Malinchoc, M. *et al.* A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* **31**, 864–871 (2000).
- Johnson, P. J. *et al.* Assessment of liver function in patients with hepatocellular carcinoma: A new evidence-based approach—the ALBI grade. *J. Clin. Oncol.* **33**, 550–558 (2015).
- Pugh, R. N., Murray-Lyon, I. M., Dawson, J. L., Pietroni, M. C. & Williams, R. Transection of the oesophagus for bleeding oesophageal varices. *Br. J. Surg.* **60**, 646–649 (1973).
- Nebelung, H. *et al.* Radioembolization versus portal vein embolization for contralateral liver lobe hypertrophy: Effect of cirrhosis. *Abdom. Radiol.* **46**, 4046–4055 (2021).
- Fedorov, A. *et al.* 3D slicer as an image computing platform for the quantitative imaging network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).
- Han, J. K. *et al.* Factors influencing vascular and hepatic enhancement at CT: Experimental study on injection protocol using a canine model. *J. Comput. Assist. Tomogr.* **24**, 400 (2000).
- Fiz, F. *et al.* Contrast administration impacts CT-based radiomics of colorectal liver metastases and non-tumoral liver parenchyma revealing the “radiological” tumour microenvironment. *Diagnostics* **11**, 1162 (2021).
- Zwanenburg, A., Leger, S., Vallières, M. & Löck, S. Image biomarker standardisation initiative. *Radiology* **295**, 328–338 (2020).
- Zwanenburg, A. *et al.* Assessing robustness of radiomic features by image perturbation. *Sci. Rep.* **9**, 614 (2019).
- Zwanenburg, A. & Löck, S. familiar: End-to-end automated machine learning and model evaluation (2021).
- Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* Vol. 32 (eds Paszke, A. *et al.*) (Curran Associates Inc, 2019).
- Hosny, A. *et al.* Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLoS Med.* **15**, e1002711 (2018).
- Starke, S. *et al.* 2D and 3D convolutional neural networks for outcome modelling of locally advanced head and neck squamous cell carcinoma. *Sci. Rep.* **10**, 15625 (2020).
- Nie, D. *et al.* Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Sci. Rep.* **9**, 1103 (2019).
- Katzman, J. L. *et al.* DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).
- Efron, B. 10 Nonparametric Confidence Intervals. In *The Jackknife, the Bootstrap and Other Resampling Plans* (ed. Efron, B.) 75–90 (Society for Industrial and Applied Mathematics, 1982). <https://doi.org/10.1137/1.9781611970319.ch10>.
- Wei, L. *et al.* A deep survival interpretable radiomics model of hepatocellular carcinoma patients. *Phys. Med.* **82**, 295–305 (2021).
- Liu, Q.-P., Xu, X., Zhu, F.-P., Zhang, Y.-D. & Liu, X.-S. Prediction of prognostic risk factors in hepatocellular carcinoma with transarterial chemoembolization using multi-modal multi-task deep learning. *eClinicalMedicine* **23**, 100379 (2020).

29. Wang, H. *et al.* Development and validation of a deep learning model for survival prognosis of transcatheter arterial chemoembolization in patients with intermediate-stage hepatocellular carcinoma. *Eur. J. Radiol.* **156**, 110527 (2022).
30. Zhang, L. *et al.* Deep learning predicts overall survival of patients with unresectable hepatocellular carcinoma treated by transarterial chemoembolization plus sorafenib. *Front. Oncol.* <https://doi.org/10.3389/fonc.2020.593292> (2020).
31. Liu, Y. *et al.* CT radiomics combined with clinical variables for predicting the overall survival of hepatocellular carcinoma patients after hepatectomy. *Transl. Oncol.* **26**, 101536 (2022).
32. Liu, Q. *et al.* A radiomics nomogram for the prediction of overall survival in patients with hepatocellular carcinoma after hepatectomy. *Cancer Imaging* **20**, 82 (2020).
33. Zheng, B.-H. *et al.* Radiomics score: A potential prognostic imaging feature for postoperative survival of solitary HCC patients. *BMC Cancer* **18**, 1148 (2018).
34. Meng, X.-P. *et al.* Radiomics analysis on multiphase contrast-enhanced CT: A survival prediction tool in patients with hepatocellular carcinoma undergoing transarterial chemoembolization. *Front. Oncol.* <https://doi.org/10.3389/fonc.2020.01196> (2020).
35. Bernatz, S. *et al.* CT-radiomics and clinical risk scores for response and overall survival prognostication in TACE HCC patients. *Sci. Rep.* **13**, 533 (2023).
36. Mackin, D. *et al.* Measuring computed tomography scanner variability of radiomics features. *Investig. Radiol.* **50**, 757–765 (2015).
37. Berenguer, R. *et al.* Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters. *Radiology* **288**, 407–415 (2018).
38. Lu, L., Ehmke, R. C., Schwartz, L. H. & Zhao, B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PLoS One* **11**, e0166550 (2016).
39. Jin, H. & Kim, J. H. Evaluation of feature robustness against technical parameters in CT radiomics: Verification of phantom study with patient dataset. *J. Signal Process. Syst.* **92**, 277–287 (2020).
40. Ziegelmayer, S. *et al.* Feature robustness and diagnostic capabilities of convolutional neural networks against radiomics features in computed tomography imaging. *Investig. Radiol.* **57**, 171–177 (2022).
41. Huynh, B. N. *et al.* Head and neck cancer treatment outcome prediction: A comparison between machine learning with conventional radiomics features and deep learning radiomics. *Front. Med.* <https://doi.org/10.3389/fmed.2023.1217037> (2023).
42. Whitfield, J. B. Gamma glutamyl transferase. *Crit. Rev. Clin. Lab. Sci.* **38**, 263–355 (2001).
43. Xu, X.-S. *et al.* Model based on  $\gamma$ -glutamyltransferase and alkaline phosphatase for hepatocellular carcinoma prognosis. *World J. Gastroenterol.* *WJG* **20**, 10944–10952 (2014).
44. Yang, Z. *et al.* Elevation of serum GGT and LDH levels, together with higher BCLC staging are associated with poor overall survival from hepatocellular carcinoma: A retrospective analysis. *Discov. Med.* **19**, 409–418 (2015).
45. Ma, W., Wang, H. & Teng, L. Correlation analysis of preoperative serum alpha-fetoprotein (AFP) level and prognosis of hepatocellular carcinoma (HCC) after hepatectomy. *World J. Surg. Oncol.* **11**, 212 (2013).
46. Tandon, P. & Garcia-Tsao, G. Prognostic indicators in hepatocellular carcinoma: A systematic review of 72 studies. *Liver Int. Off. J. Int. Assoc. Study Liver* **29**, 502–510 (2009).
47. Lu, X.-Y. *et al.* Pathobiological features of small hepatocellular carcinoma: Correlation between tumor size and biological behavior. *J. Cancer Res. Clin. Oncol.* **137**, 567–575 (2011).
48. Wu, G. *et al.* Importance of tumor size at diagnosis as a prognostic factor for hepatocellular carcinoma survival: A population-based study. *Cancer Manag. Res.* **10**, 4401–4410 (2018).
49. Lau, T. & Ahmad, J. Clinical applications of the model for end-stage liver disease (MELD) in hepatic medicine. *Hepatic Med. Evid. Res.* **5**, 1–10 (2013).
50. Hagan, M. T. *et al.* Liver volume in the cirrhotic patient: Does size matter?. *Dig. Dis. Sci.* **59**, 886–891 (2014).
51. Kim, S. *et al.* Radiomics on gadoxetic acid-enhanced magnetic resonance imaging for prediction of postoperative early and late recurrence of single hepatocellular carcinoma. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **25**, 3847–3855 (2019).
52. Teixeira, L. O. *et al.* Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images. *Sensors* **21**, 7116 (2021).

### Author contributions

F.S., A.K., S.L. and J.P.K. conceived the study; F.S., A.K., A.Z. and S.L. designed the experiments; F.S., A.K. and H.N. collected the data; F.S., A.K., C.R., A.Z., S.L. and J.P.K. analyzed the results; A.K. visualized the results; R.T.H., S.L. and J.P.K. provided resources and supervision; F.S., A.K., S.L. and J.P.K. prepared the manuscript; F.S., A.K., H.N., C.R., R.T.H., A.Z., S.L., J.P.K. reviewed the manuscript draft. All authors read and agreed to the final version of the manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-50451-3>.

**Correspondence** and requests for materials should be addressed to F.S. or A.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024