



# Interpretable deep learning framework based on contrast-enhanced MRI for predicting histological grade of hepatocellular carcinoma

Wenjun Hu<sup>1#</sup>, Xiuding Cai<sup>2,3#</sup>, Ying Zhao<sup>1#</sup>, Qihao Xu<sup>1</sup>, Xin Wang<sup>4</sup>, Qingwei Song<sup>1</sup>, Yu Yao<sup>2,3</sup>, Ailian Liu<sup>1,5</sup>

<sup>1</sup>Department of Radiology, First Affiliated Hospital of Dalian Medical University, Dalian, China; <sup>2</sup>Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, China; <sup>3</sup>School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China; <sup>4</sup>College of Medical Imaging, Dalian Medical University, Dalian, China; <sup>5</sup>Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging, Dalian, China

**Contributions:** (I) Conception and design: All authors; (II) Administrative support: A Liu; (III) Provision of study materials or patients: Q Song; (IV) Collection and assembly of data: W Hu, Y Zhao, Q Xu, X Wang; (V) Data analysis and interpretation: W Hu, X Cai, Y Yao; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

<sup>#</sup>These authors contributed equally to this work.

**Correspondence to:** Ailian Liu, MD. Department of Radiology, First Affiliated Hospital of Dalian Medical University, No. 222, Changchun Road, Xigang District, Dalian 116011, China; Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging, Dalian, China. Email: liuailian@dmu.edu.cn.

**Background:** Histopathological grading is a key prognostic marker for hepatocellular carcinoma (HCC). However, the clinical application of deep learning models (DLMs) for predicting HCC grading from medical imaging is limited by their black-box nature. We aimed to develop an interpretable DLM, interpretable HCC grading network (iHCG-Net), to predict HCC grading preoperatively using multi-phase contrast-enhanced magnetic resonance imaging (CEMRI).

**Methods:** This study retrospectively enrolled 370 HCC patients who underwent preoperative CEMRI before curative resection. Based on postoperative pathology, the patients were categorized into high-grade (n=136) and low-grade (n=234) HCC groups. They were then stratified into a training cohort (n=259) and a time-independent validation cohort (n=111). Twenty-three clinical-radiological features were collected for all patients. The iHCG-Net, based on the Concept Bottleneck Model (CBM) framework, first encodes CEMRI images using a DenseNet-121 backbone and then leverages a concept regressor to predict the twenty-three clinical-radiological features for final prediction of HCC histological grade. A feature importance score plot was generated to assess the contribution of each feature to the differential diagnosis. Nine baseline predictive models were developed for comparison. The models were evaluated using receiver operating characteristic (ROC) curve analysis and DeLong's test.

**Results:** iHCG-Net demonstrated strong predictive performance for HCC grading, achieving areas under the receiver operating characteristic curve (AUCs) of 0.893 in the training cohort and 0.802 in the validation cohort. The model significantly outperformed conventional models, including the clinical-radiological model (CM), radiomics models (RMs), and a clinical-radiomic combined model (CRM) (AUCs: 0.675–0.778, 0.617–0.723;  $P < 0.05$ ). Furthermore, iHCG-Net exhibited performance comparable to that of the DLM (AUCs: 0.920, 0.774;  $P > 0.05$ ), while providing inherent interpretability and mitigating the risk of overfitting. Feature importance analysis identified intratumoral arteries as the most influential feature for predicting HCC grading, with an importance score of 0.213.

**Conclusions:** The iHCG-Net can be a promising interpretable artificial intelligence tool for the preoperative prediction of HCC grading.

**Keywords:** Hepatocellular carcinoma (HCC); histological grade; magnetic resonance imaging (MRI); deep learning; interpretability

Submitted Feb 05, 2025. Accepted for publication Oct 30, 2025. Published online Dec 31, 2025.

doi: 10.21037/qims-2025-269

View this article at: <https://dx.doi.org/10.21037/qims-2025-269>

## Introduction

Hepatocellular carcinoma (HCC) ranks as the fifth most common malignancy and the second leading cause of cancer-related mortality worldwide (1). Histopathological grading is an important biomarker for predicting HCC recurrence rates and overall survival. Preoperative prediction of HCC grading can guide surgical resection, assess liver transplantation eligibility, and prevent unnecessary chemotherapy (2,3). Currently, biopsy is the main method for assessing HCC grading before surgery, but it has not been widely used in clinical practice due to its invasiveness. Therefore, developing a non-invasive imaging tool for preoperative HCC grading prediction is crucial.

In recent years, deep learning algorithms based on medical imaging have emerged as a promising non-invasive approach for oncology assessment (4,5). Several deep learning architectures have been developed for HCC grading prediction, including 3D SE-DenseNet, multi-scale patches convolutional neural network, and Multimodality-Contribution-Aware TripNet (6-8). However, all these studies aimed to improve the efficacy of deep learning models (DLMs) without considering the interpretability of the models. The black-box nature of DLMs is a major obstacle to their clinical application. Previous studies have attempted to interpret liver cancer prediction models using post-hoc explanation methods such as saliency maps (9-11). Nevertheless, post-hoc methods may lead to misleading interpretations, and the reliability of these interpretations is influenced by model structure and data distribution (12).

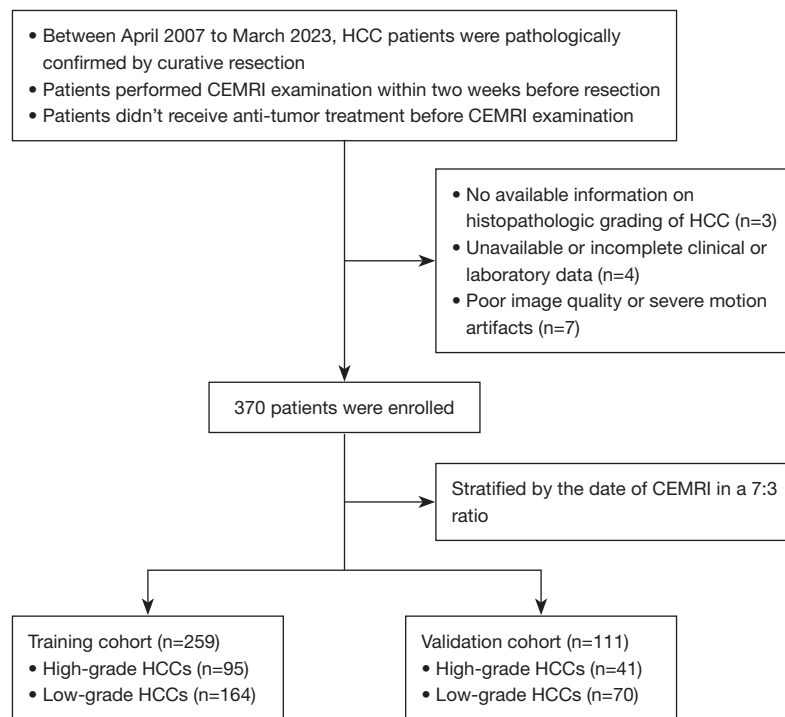
Recently, Koh *et al.* (13) proposed an intrinsically interpretable technique called the Concept Bottleneck Model (CBM). This technique simulates human cognition by incorporating an intermediate concept layer, which produces final classification results based on a comprehensive judgment of high-level concepts. This approach overcomes the limitations of post-hoc interpretation methods. However, its application in medical imaging remains to be explored. Moreover, enhancements in model interpretability often come at the cost of predictive

performance. Therefore, this study aimed to integrate the clinical radiological features based on multi-phase contrast-enhanced magnetic resonance imaging (CEMRI) and the CBM architecture to establish an interpretable HCC grading network (iHCG-Net). In addition, nine baseline prediction models, including various traditional machine learning (ML) models, black-box DLMs and multi-label combination models, were constructed for further comparison to investigate whether the iHCG-Net can achieve a balance among interpretability, efficacy and generalization ability. We present this article in accordance with the TRIPOD + AI reporting checklist (available at <https://qims.amegroups.com/article/view/10.21037/qims-2025-269/rc>).

## Methods

### Study participants

This retrospective study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments, and was approved by the Medical Ethics Committee of the First Affiliated Hospital of Dalian Medical University (No. PJ-KS-KY-2022-180). Written informed consent was waived due to the retrospective nature of the research. We initially identified patients diagnosed with HCC between April 2007 and March 2023. The inclusion criteria were: (I) pathologically confirmed HCC after curative (R0) resection; (II) preoperative CEMRI performed within two weeks prior to resection; (III) no anti-tumor treatment before CEMRI examination. Exclusion criteria were: (I) unavailable histopathologic grading information (n=3); (II) unavailable or incomplete clinical or laboratory data (n=4); (III) poor CEMRI image quality due to severe artifacts (n=7). A final cohort of 370 eligible patients was enrolled. All participants were non-randomly allocated to a training cohort (n=259; examinations from April 2007 to April 2016) and a time independent validation cohort (n=111, examinations from May 2016 to March 2023) in a 7:3 ratio. This allocation strategy conforms to a Type 2b



**Figure 1** The flow chart of this study population. CEMRI, contrast-enhanced magnetic resonance imaging; HCC, hepatocellular carcinoma.

study under the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement, a design that provides a more robust evaluation of model performance compared to random split or internal validation alone. The patient selection process is summarized in *Figure 1*.

Baseline patient characteristics were retrospectively collected, including demographics (age and gender) and laboratory markers such as alpha-fetoprotein (AFP), albumin (ALB), alanine aminotransferase (ALT), aspartate aminotransferase (AST),  $\gamma$ -glutamyl transferase (GGT), and total bilirubin (TBIL). The Child-Pugh classification was also recorded.

### **HCC histological grading**

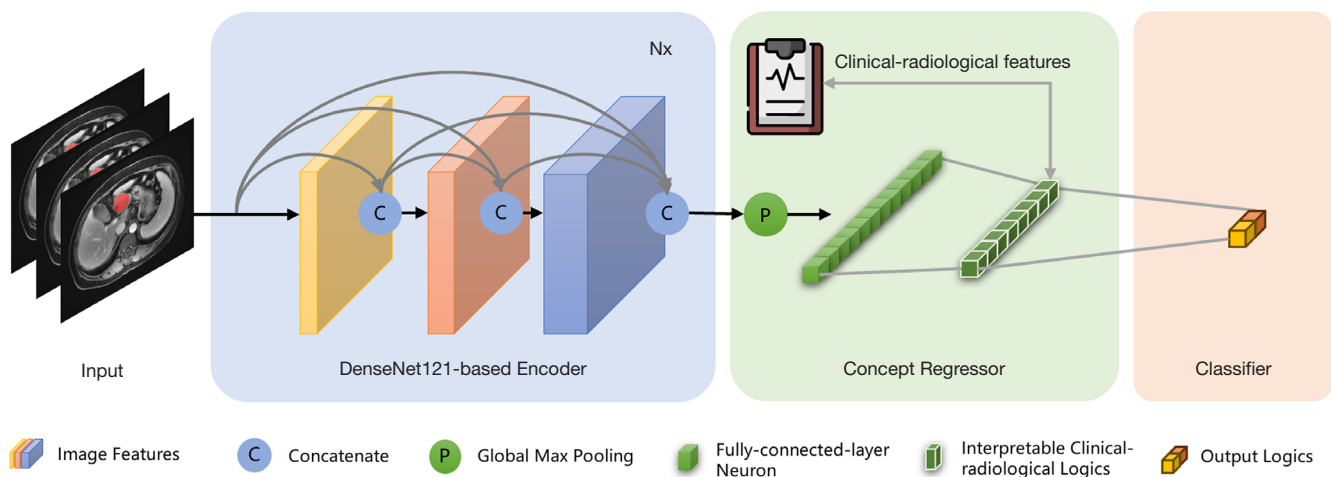
Two pathologists, blinded to all clinical data, independently assessed the pathologic grade of HCCs. Discrepancies were resolved through discussion. HCCs were classified into high-grade (poorly-differentiated) and low-grade (moderately/well-differentiated) tumors according to the fifth edition of the World Health Organization classification of digestive system tumors (14).

### **MRI acquisition**

All patients underwent magnetic resonance imaging (MRI) examinations using 1.5 T or 3.0 T scanners. The arterial phase (AP) scans, portal venous phase (PVP), and delayed phase (DP) images were collected, with scanning parameter details provided in [Appendix 1](#).

### **Image interpretation**

Two radiologists [observer 1 (W.H.) and observer 2 (Y.Z.)], with 5 and 10 years of experience in abdominal MRI interpretation respectively and blinded to all clinical and pathological groupings, independently analyzed all MRI images. They observed and recorded 13 MRI features of HCC, including tumor size, position, shape, margin, number, intratumor necrosis, hemorrhage, encapsulation, intratumoral arteries, arterial peritumoral enhancement, AP hyperenhancement, washout appearance, and liver cirrhosis. Consensus on the feature assessment was reached through discussion in cases of disagreement. The detailed assessment criteria for each feature are presented in [Appendix 2](#).



**Figure 2** The architecture of iHCG-Net. C, concatenate; iHCG-Net, interpretable hepatocellular carcinoma grading network; P, global max pooling.

### Image segmentation and preprocessing

All preoperative MR images were preprocessed to ensure data uniformity. First, the images were resampled to an isotropic voxel size of  $1 \times 1 \times 1 \text{ mm}^3$  using a linear interpolation algorithm to standardize spatial resolution. Subsequently, intensity normalization was applied across all images to mitigate scanner-related variations. A radiologist (W.H.) manually delineated the volumetric region of interest (VOI) for each tumor on all slices using ITK-SNAP software. For multifocal HCCs, the largest nodule was selected. To evaluate the inter-observer reproducibility of the segmentation, a subset of thirty tumors was randomly selected and independently segmented by a second radiologist (Y.Z.).

### iHCG-Net architecture

The proposed iHCG-Net architecture, illustrated in Figure 2, comprises three principal components. First, a DenseNet-121 backbone extracts image features. Subsequently, a concept regressor then maps these image features into a 23-dimensional clinical-radiological space, which comprises 13 MRI, 3 demographic, and 7 laboratory features. These features, routinely assessed by clinicians and radiologists in daily practice, enable the model to base its decisions on human-understandable concepts, thereby providing inherent interpretability. Finally, a classifier links these concepts to the label space for grade prediction. For each contrast phase, the three largest tumor images were

selected, with backgrounds removed. Nine such images from three phases were concatenated, resized to  $224 \times 224 \times 9$ , and fed into DenseNet-121 network  $F$ . In DenseNet-121, the input tensor  $x$  first passes through a  $7 \times 7$  convolutional layer, batch normalization, ReLU activation, and  $3 \times 3$  max-pooling for initial extraction and downsampling. Then it goes through multiple dense blocks with dense layers (where each convolutional layer's input combines previous layers' outputs for feature reuse). After each dense block, a transition layer reduces feature-map dimensionality. Finally, the features are projected into the clinical-radiological space via global average pooling and a fully-connected layer  $g(\cdot)$ . During the training phase, we utilized L1 loss as the clinical-radiological concept loss, defined as:

$$L_{CBM} = \frac{1}{m} \sum_{j=1}^m \|g[F(x)]_j - c_j\| \quad [1]$$

where  $g[F(x)]_j$  represents the predicted  $j$ -th feature of the patient, and  $c_j$  denotes the actual clinical-radiological feature of the patient. Finally, a simple linear classification layer was added after the clinical-radiological feature space to predict the outcome. We used cross-entropy loss function to optimize the classification task, minimizing the difference between the model's output and the true labels. The cross-entropy loss function can be expressed as:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{i,k} \log(\hat{y}_{i,k}) \quad [2]$$

where  $N$  represents the number of samples,  $K$  represents

the number of classes,  $y_{i,k}$  is the true label indicating whether sample  $i$  belongs to class  $k$ , and  $\hat{y}_{i,k}$  is the predicted probability by the model that sample  $i$  belongs to class  $k$ . Our ultimate loss can be represented as:

$$L_{Total} = L_{CE} + \alpha L_{CBM} \quad [3]$$

where  $\alpha$  represents the weighting factor that balances the contribution of the cross-entropy loss and the L1 loss for clinical-radiological concept learning. The predictive performance of the concept regressor for various clinical-radiological features was evaluated by calculating the mean absolute error (MAE).

A feature importance score plot was generated to represent the relative contribution of each feature to predicting HCC grading. Given the nature of iHCG-Net, all features can be transformed into a linear combination as follows:

$$\hat{y} = \alpha_0 \text{concept}_0 + \alpha_1 \text{concept}_1 + \dots + \alpha_{22} \text{concept}_{22} \quad [4]$$

where,  $\hat{y}$  represents the predicted logic before softmax, while  $\text{concept}_i$  and  $\alpha_i$  denote the  $i$ -th predicted feature and its corresponding linear combination weight, respectively. We defined magnitude of  $\alpha_i$ , i.e.,  $|\alpha_i|$  as the importance score of the corresponding feature. The SHapley Additive exPlanations (SHAP) analysis was applied to provide post-hoc interpretation of feature importance and contribution direction for individual predictions across all samples.

A batch size of 32 was used, with AdamW optimizer (beta1 = 0.9, beta2 = 0.999), a learning rate of 0.001, and a 20-epoch cosine annealing strategy over 200 epochs. Early stopping and data augmentation (random cropping and flipping) were applied. Experimenting with hyperparameter values, 0.1 gave the best performance. All work was done on an Intel(R) Core(TM) i7-12700K and NVIDIA GeForce RTX 3090Ti GPU-equipped machine.

### Development of comparative models

Nine baseline predictive models were constructed for comparison with iHCG-Net, including a clinical-radiological model (CM), five radiomics models (RMs) using various classical ML algorithms [AdaBoost (AB), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and Gradient Boosting (GB)], a clinical-radiomic combined model (CRM), a DLM using image data; and a deep learning combined model (DLCM) integrating both image and clinical-radiological data. The

development pipeline for these baseline models is illustrated in Figure 3. Detailed information on model construction and feature selection is presented in Appendix 3. In addition, we introduced a baseline model termed the clinical-radiological linear model (CRLM), which directly input ground-truth clinical-radiological features into the linear classifier of iHCG-Net for validation, to determine whether the concept regressor adds more value than directly using existing data.

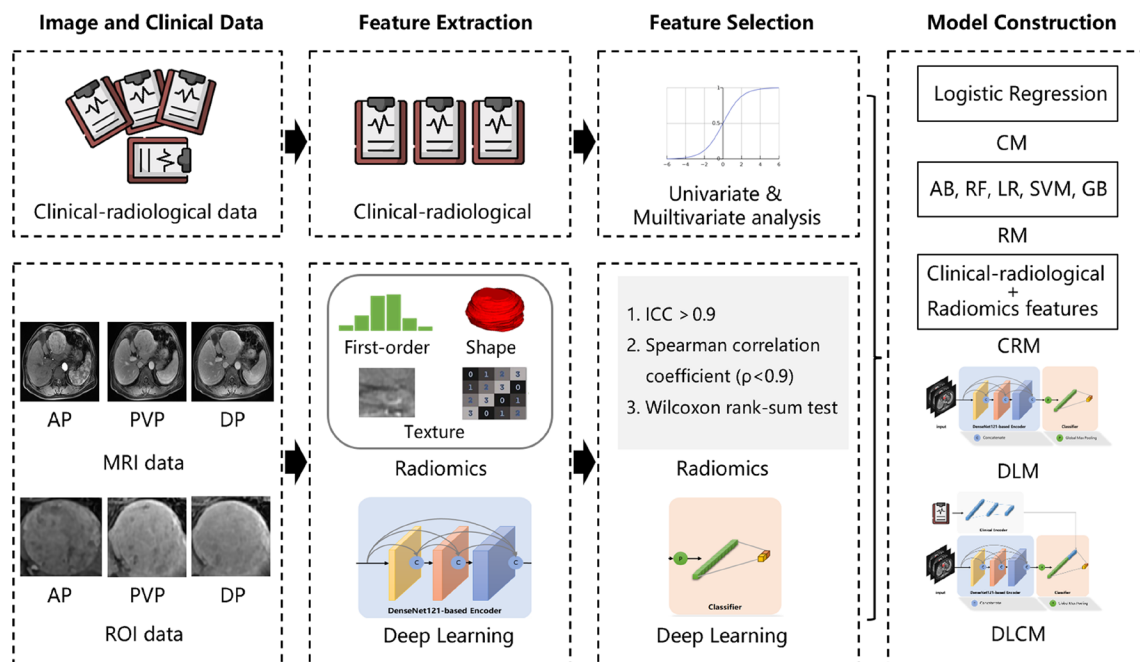
### Statistical analysis

All statistical analyses were performed using Python (version 3.9.18). The Kolmogorov-Smirnov tests were applied for normality testing. Normality distributed continuous data were compared with the independent samples  $t$ -test, while non-normally distributed data were analyzed with the Mann-Whitney  $U$  test. Categorical variables were compared using the Chi-squared test or Fisher's exact test. Associations between variables were assessed with Pearson or Spearman correlation analysis. The model's performance was rigorously evaluated through a five-fold cross-validation procedure on the training set for internal validation. Its generalization ability was subsequently assessed on the independent temporal validation cohort. Predictive performance was quantified using receiver operating characteristic (ROC) analysis, with the area under the curve (AUC), accuracy, sensitivity, and specificity reported. The DeLong's test was employed to compare the AUC values. A two-sided  $P$  value of less than 0.05 was considered statistically significant.

## Results

### General data

This study ultimately enrolled a total of 370 patients. The study population had a median age of 62 years and was predominantly male (79.2%). There were 259 and 111 patients finally enrolled in the training (164 low-grade and 95 high-grade HCCs) and validation cohorts (70 low-grade and 41 high-grade HCCs), respectively. Despite the temporal interval between the cohorts, no significant differences were observed in the prevalence of high-grade HCC (36.7% *vs.* 36.9%,  $P > 0.99$ ) or other clinical characteristics (all  $P > 0.05$ ; Table 1), indicating that the two cohorts were well-balanced and justifying their use for model training and validation.



**Figure 3** The flow-chart of the baseline models development. The detailed structures of each model are provided in Appendix 3. AB, AdaBoost; AP, arterial phase; CM, clinical-radiological model; CRM, clinical-radiomic combined model with SVM classifier; DLCM, deep learning combined model; DLM, deep learning model; DP, delayed phase; GB, Gradient Boosting; ICC, intraclass correlation coefficients; LR, Logistic Regression; MRI, magnetic resonance imaging; PVP, portal venous phase; RF, Random Forest; RM, radiomics model with SVM classifier; ROI, region of interest; SVM, Support Vector Machine.

### Clinical-radiological and radiomics feature selection

Multivariate logistic regression analysis identified AFP [odds ratio (OR) =2.313; 95% confidence interval (CI): 1.282–4.174;  $P=0.005$ ], tumor shape (OR =2.013; 95% CI: 1.233–3.287;  $P=0.005$ ), and liver cirrhosis (OR =1.669; 95% CI: 1.039–2.680;  $P=0.034$ ) as independent predictors for HCC grading, as summarized in *Table 2*.

Among the 3,111 extracted radiomics features, 2,703 features showed high reproducibility. Next, redundant radiomics features were removed using the Spearman correlation method, 758 radiomics features were retained for further analysis. Finally, the top 10 radiomics features significantly associated with HCC grade, as identified by the Wilcoxon rank-sum test, were used to construct the RM for comparison with iHCG-Net (*Appendix 4*). These high-dimensional features often capture sub-visual patterns beyond human perception and, in contrast to intuitive clinical-radiological concepts, represent relatively abstract statistical descriptors that remain challenging for clinicians to interpret.

### Model development and evaluation

The CM, developed based on AFP, tumor shape, and liver cirrhosis, achieved AUCs of 0.675 in the training cohort and 0.617 in the validation cohort. Ten selected radiomics features were used to construct RMs with five distinct ML classifiers (*Table 3* and *Figure 4*). For all classifiers except LR and SVM, significant differences in AUC were observed between the training and validation cohorts ( $P<0.05$ ). The SVM-based model significantly outperformed the LR-based model ( $P<0.05$ ), with corresponding AUCs of 0.764 and 0.682 in the training and validation cohorts, respectively. CRM was built by integrating radiomics features, AFP, tumor shape, and liver cirrhosis using an SVM classifier, achieving AUCs of 0.778 in the training cohort and 0.723 in the validation cohort. The deep-learning models DLM and DLCM had the best performance in the training cohort, with AUCs of 0.920 and 0.982 respectively. However, in the validation cohort, their AUCs were only 0.774 and 0.787, showing severe overfitting ( $P<0.05$ ). The iHCG-Net achieved an AUC of 0.893, accuracy of 0.781,

**Table 1** Clinical-radiological characteristics in the training and validation cohorts

| Characteristics             | Training cohort (n=259) |                   |                | Validation cohort (n=111) |                   |                | P <sup>‡</sup> |
|-----------------------------|-------------------------|-------------------|----------------|---------------------------|-------------------|----------------|----------------|
|                             | High-grade (n=95)       | Low-grade (n=164) | P <sup>†</sup> | High-grade (n=41)         | Low-grade (n=70)  | P <sup>†</sup> |                |
| Gender                      |                         |                   |                |                           |                   |                |                |
| Male                        | 72 (75.8)               | 132 (80.5)        | 0.463          | 34 (82.9)                 | 55 (78.6)         | 0.757          | 0.867          |
| Female                      | 23 (24.2)               | 32 (19.5)         |                | 7 (17.1)                  | 15 (21.4)         |                |                |
| Age (years)                 | 60.0 (51.0, 68.0)       | 62.5 (55.8, 68.0) | 0.078          | 57.0 (53.0, 63.0)         | 63.0 (56.0, 67.0) | 0.072          | 0.577          |
| History of hepatitis B or C |                         |                   | 0.845          |                           |                   | >0.99          | 0.425          |
| Negative                    | 25 (26.3)               | 40 (24.4)         |                | 12 (29.3)                 | 21 (30.0)         |                |                |
| Positive                    | 70 (73.7)               | 124 (75.6)        |                | 29 (70.7)                 | 49 (70.0)         |                |                |
| AFP (IU/mL)                 |                         |                   | 0.015*         |                           |                   | 0.052          | 0.664          |
| ≤400                        | 70 (73.7)               | 142 (86.6)        |                | 28 (68.3)                 | 60 (85.7)         |                |                |
| >400                        | 25 (26.3)               | 22 (13.4)         |                | 13 (31.7)                 | 10 (14.3)         |                |                |
| ALT (U/L)                   |                         |                   | 0.239          |                           |                   | 0.768          | 0.346          |
| ≤50                         | 77 (81.1)               | 121 (73.8)        |                | 28 (68.3)                 | 51 (72.9)         |                |                |
| >50                         | 18 (18.9)               | 43 (26.2)         |                | 13 (31.7)                 | 19 (27.1)         |                |                |
| AST (U/L)                   |                         |                   | >0.99          |                           |                   | >0.99          | 0.385          |
| ≤40                         | 68 (71.6)               | 118 (72.0)        |                | 27 (65.9)                 | 47 (67.1)         |                |                |
| >40                         | 27 (28.4)               | 46 (28.0)         |                | 14 (34.1)                 | 23 (32.9)         |                |                |
| GGT (U/L)                   |                         |                   | >0.99          |                           |                   | 0.091          | 0.116          |
| ≤60                         | 59 (62.1)               | 103 (62.8)        |                | 17 (41.5)                 | 42 (60.0)         |                |                |
| >60                         | 36 (37.9)               | 61 (37.2)         |                | 24 (58.5)                 | 28 (40.0)         |                |                |
| TBIL (μmol/L)               |                         |                   | 0.998          |                           |                   | 0.530          | 0.801          |
| ≤19                         | 68 (71.6)               | 116 (70.7)        |                | 28 (68.3)                 | 53 (75.7)         |                |                |
| >19                         | 27 (28.4)               | 48 (29.3)         |                | 13 (31.7)                 | 17 (24.3)         |                |                |
| ALB (g/L)                   |                         |                   | 0.433          |                           |                   | 0.683          | 0.490          |
| ≤40                         | 42 (44.2)               | 63 (38.4)         |                | 20 (48.8)                 | 30 (42.9)         |                |                |
| >40                         | 53 (55.8)               | 101 (61.6)        |                | 21 (51.2)                 | 40 (57.1)         |                |                |
| Child-Pugh class            |                         |                   | >0.99          |                           |                   | 0.381          | 0.442          |
| A                           | 81 (85.3)               | 139 (84.8)        |                | 31 (75.6)                 | 59 (84.3)         |                |                |
| B                           | 14 (14.7)               | 25 (15.2)         | 0.081          | 10 (24.4)                 | 11 (15.7)         | 0.012*         | 0.643          |
| Tumor size (cm)             | 3.7 (2.7, 5.8)          | 3.5 (2.0, 5.2)    |                | 4.2 (2.6, 7.0)            | 3.0 (2.1, 4.5)    |                |                |
| Tumor position              |                         |                   | 0.464          |                           |                   | 0.859          | 0.294          |
| Left lobe                   | 26 (27.4)               | 46 (28.0)         |                | 8 (19.5)                  | 14 (20.0)         |                |                |
| Junction lobe               | 2 (2.1)                 | 8 (4.9)           |                | 2 (4.9)                   | 2 (2.9)           |                |                |
| Right lobe                  | 67 (70.5)               | 108 (65.9)        |                | 31 (75.6)                 | 54 (77.1)         |                |                |
| Caudate lobe                | 0 (0.0)                 | 2 (1.2)           |                | 0 (0.0)                   | 0 (0.0)           |                |                |

**Table 1** (continued)

Table 1 (continued)

| Characteristics                  | Training cohort (n=259) |                   |                | Validation cohort (n=111) |                  |                | P <sup>‡</sup> |
|----------------------------------|-------------------------|-------------------|----------------|---------------------------|------------------|----------------|----------------|
|                                  | High-grade (n=95)       | Low-grade (n=164) | P <sup>†</sup> | High-grade (n=41)         | Low-grade (n=70) | P <sup>†</sup> |                |
| Tumor shape                      |                         |                   | 0.018*         |                           |                  | 0.076          | 0.887          |
| Circular                         | 52 (54.7)               | 115 (70.1)        |                | 21 (51.2)                 | 49 (70.0)        |                |                |
| Irregular                        | 43 (45.3)               | 49 (29.9)         |                | 20 (48.8)                 | 21 (30.0)        |                |                |
| Tumor margin                     |                         |                   | 0.130          |                           |                  | 0.885          | >0.99          |
| Smooth                           | 54 (56.8)               | 110 (67.1)        |                | 25 (61.0)                 | 45 (64.3)        |                |                |
| Non-smooth                       | 41 (43.2)               | 54 (32.9)         |                | 16 (39.0)                 | 25 (35.7)        |                |                |
| Tumor number                     |                         |                   | 0.017*         |                           |                  | 0.318          | 0.658          |
| Unifocal                         | 78 (82.1)               | 152 (92.7)        |                | 39 (95.1)                 | 62 (88.6)        |                |                |
| Multifocal                       | 17 (17.9)               | 12 (7.3)          |                | 2 (4.9)                   | 8 (11.4)         |                |                |
| Intratumor necrosis              |                         |                   | 0.817          |                           |                  | 0.271          | 0.872          |
| Absent                           | 54 (56.8)               | 97 (59.1)         |                | 20 (48.8)                 | 43 (61.4)        |                |                |
| Present                          | 41 (43.2)               | 67 (40.9)         |                | 21 (51.2)                 | 27 (38.6)        |                |                |
| Intratumor hemorrhage            |                         |                   | 0.510          |                           |                  | 0.544          | >0.99          |
| Absent                           | 66 (69.5)               | 106 (64.6)        |                | 25 (61.0)                 | 48 (68.6)        |                |                |
| Present                          | 29 (30.5)               | 58 (35.4)         |                | 16 (39.0)                 | 22 (31.4)        |                |                |
| Tumor encapsulation              |                         |                   | 0.524          |                           |                  | >0.99          | >0.99          |
| Absent                           | 34 (35.8)               | 51 (31.1)         |                | 14 (34.1)                 | 23 (32.9)        |                |                |
| Present                          | 61 (64.2)               | 113 (68.9)        |                | 27 (65.9)                 | 47 (67.1)        |                |                |
| Intratumoral arteries            |                         |                   | 0.400          |                           |                  | 0.480          | 0.726          |
| Absent                           | 62 (65.3)               | 97 (59.1)         |                | 24 (58.5)                 | 47 (67.1)        |                |                |
| Present                          | 33 (34.7)               | 67 (40.9)         |                | 17 (41.5)                 | 23 (32.9)        |                |                |
| Arterial peritumoral enhancement |                         |                   | 0.132          |                           |                  | 0.226          | 0.433          |
| Absent                           | 48 (50.5)               | 100 (61.0)        |                | 22 (53.7)                 | 47 (67.1)        |                |                |
| Present                          | 47 (49.5)               | 64 (39.0)         |                | 19 (46.3)                 | 23 (32.9)        |                |                |
| Arterial phase hyperenhancement  |                         |                   | 0.229          |                           |                  | 0.145          | 0.487          |
| Absent                           | 28 (29.5)               | 36 (22.0)         |                | 12 (29.3)                 | 11 (15.7)        |                |                |
| Present                          | 67 (70.5)               | 128 (78.0)        |                | 29 (70.7)                 | 59 (84.3)        |                |                |
| Washout appearance               |                         |                   | 0.295          |                           |                  | 0.466          | 0.358          |
| Absent                           | 37 (38.9)               | 52 (31.7)         |                | 14 (34.1)                 | 18 (25.7)        |                |                |
| Present                          | 58 (61.1)               | 112 (68.3)        |                | 27 (65.9)                 | 52 (74.3)        |                |                |
| Liver cirrhosis                  |                         |                   | 0.013*         |                           |                  | 0.641          | 0.316          |
| Absent                           | 31 (32.6)               | 81 (49.4)         |                | 22 (53.7)                 | 33 (47.1)        |                |                |
| Present                          | 64 (67.4)               | 83 (50.6)         |                | 19 (46.3)                 | 37 (52.9)        |                |                |

Quantitative data were presented as median (25th, 75th percentiles), using the Mann-Whitney *U* test. Categorical data were presented as n (%), using the Chi-squared test or Fisher exact test. <sup>†</sup>, comparison between low-grade and high-grade groups; <sup>‡</sup>, comparison between training and validation cohorts; \*, P<0.05. AFP, alpha-fetoprotein; ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT,  $\gamma$ -glutamyl transpeptidase; TBIL, total bilirubin.

**Table 2** Univariate and multivariate analyses of clinical-radiological characteristics for predicting HCC grading in the training cohort

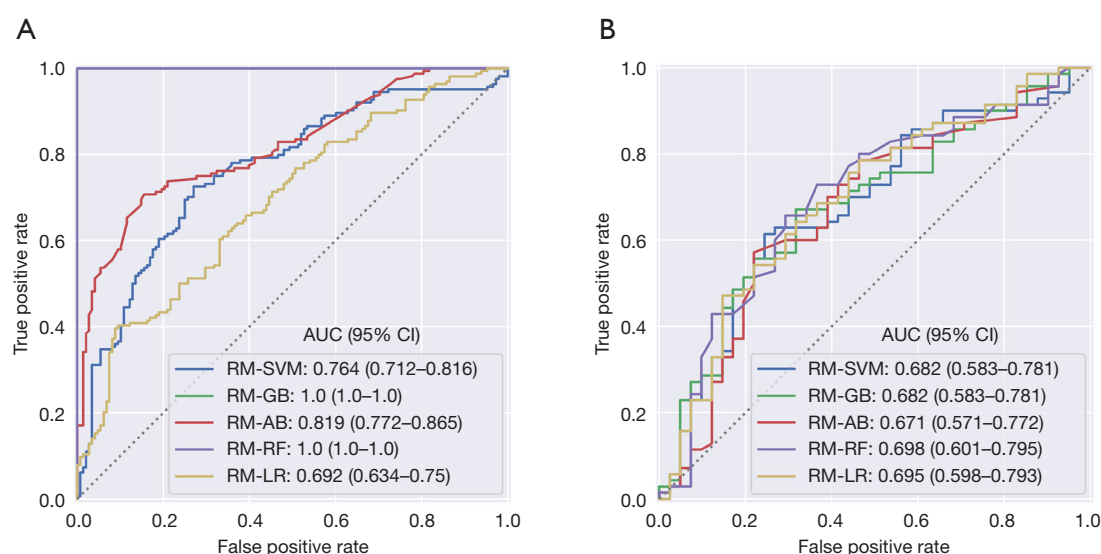
| Variables                        | Univariate analysis |         | Multivariate analysis |         |
|----------------------------------|---------------------|---------|-----------------------|---------|
|                                  | Odd ratio (95% CI)  | P value | OR (95% CI)           | P value |
| Gender                           | 0.961 (0.558–1.656) | 0.886   |                       |         |
| Age                              | 1.335 (0.962–1.854) | 0.084   |                       |         |
| History of hepatitis B or C      | 0.791 (0.473–1.324) | 0.373   |                       |         |
| AFP                              | 2.275 (1.286–4.023) | 0.005   | 2.313 (1.282–4.174)   | 0.005   |
| ALT                              | 1.137 (0.672–1.924) | 0.632   |                       |         |
| AST                              | 1.058 (0.646–1.734) | 0.823   |                       |         |
| GGT                              | 0.911 (0.581–1.428) | 0.683   |                       |         |
| TBIL                             | 1.339 (0.803–2.233) | 0.263   |                       |         |
| ALB                              | 1.285 (0.817–2.022) | 0.278   |                       |         |
| Child-Pugh class                 | 0.786 (0.417–1.483) | 0.458   |                       |         |
| Tumor size                       | 1.476 (0.979–2.224) | 0.063   |                       |         |
| Tumor position                   | 1.202 (0.792–1.825) | 0.388   |                       |         |
| Tumor shape                      | 2.278 (1.419–3.655) | 0.001   | 2.013 (1.233–3.287)   | 0.005   |
| Tumor margin                     | 0.735 (0.465–1.163) | 0.189   |                       |         |
| Tumor number                     | 0.743 (0.370–1.491) | 0.403   |                       |         |
| Intratumor necrosis              | 0.870 (0.556–1.362) | 0.543   |                       |         |
| Intratumor hemorrhage            | 0.852 (0.536–1.354) | 0.497   |                       |         |
| Tumor encapsulation              | 1.026 (0.630–1.671) | 0.918   |                       |         |
| Intratumoral arteries            | 1.007 (0.642–1.578) | 0.976   |                       |         |
| Arterial peritumoral enhancement | 0.811 (0.518–1.271) | 0.361   |                       |         |
| Arterial phase hyperenhancement  | 1.400 (0.820–2.389) | 0.218   |                       |         |
| Washout appearance               | 1.522 (0.941–2.459) | 0.087   |                       |         |
| Liver cirrhosis                  | 1.597 (1.019–2.504) | 0.041   | 1.669 (1.039–2.680)   | 0.034   |

AFP, alpha-fetoprotein; ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT,  $\gamma$ -glutamyl transpeptidase; HCC, hepatocellular carcinoma; TBIL, total bilirubin.

**Table 3** Discrimination performance of different traditional ML classifiers in the training and validation cohorts

| Model | Training cohort     |       |       |       | Validation cohort   |       |       |       | P <sup>†</sup> |
|-------|---------------------|-------|-------|-------|---------------------|-------|-------|-------|----------------|
|       | AUC (95% CI)        | SEN   | SPE   | ACC   | AUC (95% CI)        | SEN   | SPE   | ACC   |                |
| SVM   | 0.764 (0.712–0.816) | 0.774 | 0.642 | 0.712 | 0.682 (0.583–0.781) | 0.729 | 0.512 | 0.649 | 0.169          |
| GB    | 1.000 (1.000–1.000) | 1.000 | 1.000 | 1.000 | 0.682 (0.583–0.781) | 0.714 | 0.537 | 0.649 | <0.0001*       |
| AB    | 0.819 (0.772–0.865) | 0.750 | 0.723 | 0.737 | 0.671 (0.571–0.772) | 0.671 | 0.61  | 0.649 | 0.015          |
| RF    | 1.000 (1.000–1.000) | 1.000 | 1.000 | 1.000 | 0.698 (0.601–0.795) | 0.800 | 0.537 | 0.703 | <0.0001*       |
| LR    | 0.692 (0.634–0.750) | 0.677 | 0.561 | 0.622 | 0.695 (0.598–0.793) | 0.700 | 0.585 | 0.658 | 0.946          |

<sup>†</sup>, comparison of AUCs between training and validation cohorts; \*, P<0.05. AB, AdaBoost; ACC, accuracy; AUC, area under the curve; CI, confidence interval; GB, Gradient Boosting; LR, Logistic Regression; ML, machine learning; RF, Random Forest; SEN, sensitivity; SPE, specificity; SVM, Support Vector Machine.



**Figure 4** ROC curves for different traditional ML classifiers in the training cohort (A) and validation cohort (B). AB, AdaBoost; AUC, area under the curve; CI, confidence interval; GB, Gradient Boosting; ML, machine learning; RF, Random Forest; RM, radiomics model with SVM classifier; ROC, receiver operating characteristic; SVM, Support Vector Machine; LR, Logistic Regression.

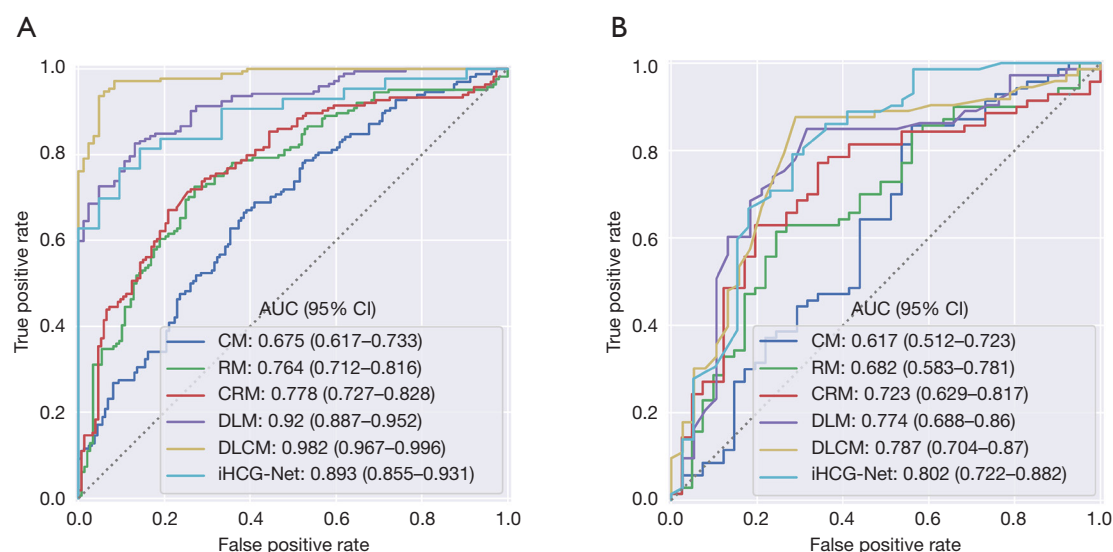
**Table 4** Discrimination performance of different models in the training and validation cohorts

| Model    | Training cohort     |       |       |       | Validation cohort   |       |       |       | P <sup>†</sup> |
|----------|---------------------|-------|-------|-------|---------------------|-------|-------|-------|----------------|
|          | AUC (95% CI)        | SEN   | SPE   | ACC   | AUC (95% CI)        | SEN   | SPE   | ACC   |                |
| CM       | 0.675 (0.617–0.733) | 0.659 | 0.615 | 0.637 | 0.617 (0.512–0.723) | 0.629 | 0.561 | 0.604 | 0.409          |
| RM       | 0.764 (0.712–0.816) | 0.774 | 0.642 | 0.712 | 0.682 (0.583–0.781) | 0.729 | 0.512 | 0.649 | 0.169          |
| CRM      | 0.778 (0.727–0.828) | 0.713 | 0.745 | 0.729 | 0.723 (0.629–0.817) | 0.743 | 0.659 | 0.712 | 0.338          |
| DLM      | 0.920 (0.887–0.952) | 0.843 | 0.833 | 0.84  | 0.774 (0.688–0.860) | 0.849 | 0.632 | 0.775 | 0.005*         |
| DLCM     | 0.982 (0.967–0.996) | 0.802 | 0.976 | 0.859 | 0.787 (0.704–0.870) | 0.877 | 0.684 | 0.811 | <0.0001*       |
| iHCG-Net | 0.893 (0.855–0.931) | 0.698 | 0.952 | 0.781 | 0.802 (0.722–0.882) | 0.861 | 0.641 | 0.784 | 0.079          |

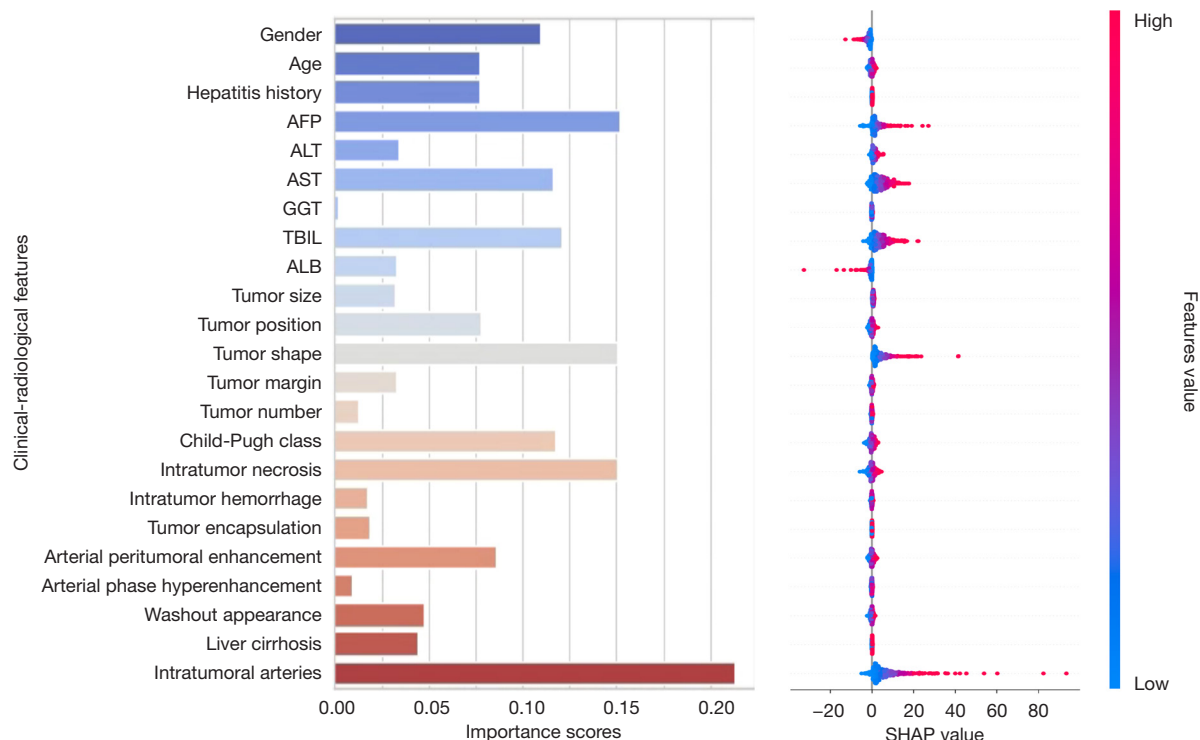
<sup>†</sup>, comparison of AUCs between training and validation cohorts; \*,  $P < 0.05$ . ACC, accuracy; AUC, area under the curve; CI, confidence interval; CM, clinical-radiological model; CRM, clinical-radiomic machine learning model with Support Vector Machine classifier; DLCM, deep learning combined model; DLM, deep learning model; iHCG-Net, interpretable hepatocellular carcinoma grading network; RM, radiomics model with Support Vector Machine classifier; SEN, sensitivity; SPE, specificity.

sensitivity of 0.698, and specificity of 0.952 in the training cohort, with corresponding values of 0.802, 0.784, 0.861, and 0.640 in the validation cohort. The performance of iHCG-Net was superior to that of CM, RM, and CRM ( $P < 0.05$ ), and was comparable to that of DLM ( $P > 0.05$ ). No significant difference in AUC was observed between the training and validation cohorts for iHCG-Net ( $P > 0.05$ ). The performance and ROC curves of all predictive models are exhibited in Table 4 and Figure 5, respectively. Statistical comparisons between the different predictive models based

on DeLong's test are available in Appendix 4. Both the feature importance score plot and SHAP analysis identified intratumoral arteries as the most influential feature for HCC grade prediction (importance score = 0.213) (Figure 6). All clinical-radiological features exhibited a negative correlation between importance scores and MAE (Pearson's  $r = -0.103$ ; Spearman's  $r = -0.086$ ; both  $P < 0.0001$ ) (Appendix 4), indicating that the concept regressor could predict more important features with lower error. The CRLM, based solely on ground truth measurements, achieved an AUC of



**Figure 5** ROC curves for different predictive models in the training cohort (A) and validation cohort (B). CM, clinical-radiological model; CRM, clinical-radiomic combined model with SVM classifier; DLM, deep learning model; DLCM, deep learning combined model; iHCG-Net, interpretable hepatocellular carcinoma grading network; RM, radiomics model with SVM classifier; ROC, receiver operating characteristic.



**Figure 6** Importance score plot and SHAP plot. Left panel: X-axis denotes the importance scores of each feature. Right panel: X-axis represents the SHAP values of each sample; Y-axis indicates the magnitudes of feature values, with red denoting high values and blue denoting low values. AFP, alpha-fetoprotein; ALB, albumin; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT,  $\gamma$ -glutamyl transpeptidase; SHAP, SHapley Additive exPlanation; TBIL, total bilirubin.

0.576 (95% CI: 0.466–0.687), with corresponding accuracy, sensitivity, and specificity of 0.577, 0.699, and 0.342, respectively. This performance was significantly inferior to that of iHCG-Net ( $P=0.001$ ).

## Discussion

To our knowledge, this study represents the first effort to develop and validate an interpretable deep learning framework for preoperative prediction of HCC grading. Our study demonstrated that iHCG-Net can effectively predict HCC grading with intrinsic interpretability. Its performance exceeds that of traditional ML models and is comparable to DLMs. Notably, iHCG-Net mitigated the overfitting observed in the conventional DLMs. The feature importance score plot showed that the intra-tumoral artery was the most influential feature for predicting the HCC grading.

In our study, AFP, tumor shape, and liver cirrhosis were included in the CM, with their correlations to HCC grading confirmed in previous studies (15,16). High-throughput radiomics features provide a comprehensive description of tumor heterogeneity. Previous research applied various ML algorithms to create RMs for HCC grading (17,18), but the best approach remains unclear. Here, five common ML classifiers were used to build the RM. All classifiers, except SVM and LR, exhibited overfitting. A meta-analysis by Wang *et al.* (19) showed SVM and LR are the most common classifiers for this task, validating our finding. In our analysis, the SVM-based model demonstrated superior performance over the LR-based model, likely attributable to its enhanced adaptability to high-dimensional data, superior capability in modeling complex nonlinear relationships, and greater robustness to noise and outliers (20,21). Although the CM and RM models offer interpretability, their predictive performance was limited. This may be due to the lower dimensionality of the data and the limitations of traditional modeling in representing tumor biological behaviors.

Deep learning algorithms can autonomously extract and learn high-dimensional features from input data using multi-layer artificial neural networks, making them highly effective in modeling complex patterns and decision-making processes (22). Consistent with previous research (23,24), the DLM demonstrated higher prediction efficiency compared to the CM and RM models. The performance of DLCM improved compared to DLM, indicating that the integration of multi-dimensional risk factors enhances

their interactions and optimizes the model's effectiveness. However, the black-box nature of DLMs makes it difficult for doctors to fully trust their predictions of HCC grading. In addition, despite the use of data augmentation, DLM and DLCM still had serious overfitting problems. On one hand, the limited sample size restricted the generalization ability of the models. On the other hand, the complex structures and large number of parameters of DLMs made them overly sensitive to interference factors, such as data deviation and noise, leading to overfitting (25).

By introducing an intermediate concept layer, iHCG-Net enables clinicians to clearly visualize the key concepts extracted and used by the model during the learning process, thus eliminating the black-box nature of DLMs. The study results showed that the iHCG-Net enhanced interpretability without compromising model performance. Its efficacy was comparable to that of the DLM and effectively avoided overfitting. This can be attributed to the following factors (13): (I) the architecture of CBM enables the DLM to first learn the concepts with clear semantics related to the task instead of directly outputting based on the original data. This reduces the model's need for large sample sizes and prevents it from fitting noise, data deviation, or other irrelevant information. (II) The clinical-radiological features in the study are more stable than the original data, reducing the impact of varying datasets and image acquisition parameters on model efficacy. Additionally, the significantly lower AUC of CRLM compared to iHCG-Net confirms that the concept regressor provides additional value by integrating MRI features and clinical-radiological features rather than relying solely on raw data. This may be attributed to its ability to predict “soft” clinical-radiological features instead of hard diagnostic labels, which better adapts to the complexity of medical data.

The feature importance score plot generated by iHCG-Net can visually display the key features for predicting HCC grading. The results showed that intratumoral arteries had the highest importance score. The SHAP plot provided additional directional information, indicating that both low and high intratumoral artery levels increased the probability of high-grade HCC, suggesting a non-linear relationship between them. This may be attributed to the trend where the number of intratumoral arteries first increases and then decreases during the progression from well-differentiated to moderately differentiated and then to poorly differentiated HCC (26). Reduced vascular endothelial growth factor expression and increased anaerobic metabolism in poorly

differentiated HCC may underlie this phenomenon (27,28). Furthermore, the enhanced heterogeneity of high-grade HCC, characterized by a mixture of hypervascular and hypovascular regions, may also contribute to this non-linear relationship.

There are several limitations in this study. First, its retrospective, single-center design and the limited sample size may affect the generalizability of the findings. Future validation using larger, multi-center datasets is necessary to confirm the robustness of the proposed model. Second, iHCG-Net incorporated 23 commonly used manually-defined concepts. Whether additional unsupervised concepts can further enhance the model's performance is a direction for our further research. Third, in this study, iHCG-Net was only used to predict HCC grading. In the future, we will extend iHCG-Net to predict various biological behaviors and prognosis of HCC. Lastly, this study focused on enhancing the interpretability of DLMS and did not include models such as 3D SE-DenseNet and Multimodality-Contribution-Aware TripNet, which aim to improve model efficacy, in the experimental comparisons. This may, to some extent, limit the comprehensive performance reference of different HCC grading models under the same task scenario. Future studies will expand the sample size and further supplement comparative analyses of such cross-objective designed models.

## Conclusions

We have developed an interpretable model, iHCG-Net. It was comparable in performance to black-box DLMS and effectively avoided overfitting. iHCG-Net has the potential to serve as an imaging biomarker for predicting HCC grading. This could help provide a basis for personalized treatment of HCC patients and promote the integration of artificial intelligence and radiology in clinical practice.

## Acknowledgments

None.

## Footnote

**Reporting Checklist:** The authors have completed the TRIPOD+AI reporting checklist. Available at <https://qims.amegroups.com/article/view/10.21037/qims-2025-269/rc>

**Data Sharing Statement:** Available at <https://qims.amegroups.com/article/view/10.21037/qims-2025-269/dss>

[amegroups.com/article/view/10.21037/qims-2025-269/dss](https://qims.amegroups.com/article/view/10.21037/qims-2025-269/dss)

**Funding:** This work was supported by the Dalian Young Stars of Science and Technology Project Support Program (No. 2022RQ074) and Dalian Medical Science Research Program (No. 2212011).

**Conflicts of Interest:** All authors have completed the ICMJE uniform disclosure form (available at <https://qims.amegroups.com/article/view/10.21037/qims-2025-269/coif>). The authors have no conflicts of interest to declare.

**Ethical Statement:** The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki and its subsequent amendments, and was approved by the Medical Ethics Committee of the First Affiliated Hospital of Dalian Medical University (No. PJ-KS-KY-2022-180). Written informed consent was waived due to the retrospective nature of the research.

**Open Access Statement:** This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

## References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87-108.
2. Oishi K, Itamoto T, Amano H, Fukuda S, Ohdan H, Tashiro H, Shimamoto F, Asahara T. Clinicopathologic features of poorly differentiated hepatocellular carcinoma. *J Surg Oncol* 2007;95:311-6.
3. Okusaka T, Okada S, Ueno H, Ikeda M, Shimada K, Yamamoto J, Kosuge T, Yamasaki S, Fukushima N, Sakamoto M. Satellite lesions in patients with small hepatocellular carcinoma with reference to clinicopathologic features. *Cancer* 2002;95:1931-7.
4. Sharma P, Nayak DR, Balabantaray BK, Tanveer M,

- Nayak R. A survey on cancer detection via convolutional neural networks: Current challenges and future directions. *Neural Netw* 2024;169:637-59.
5. Chen F, He Y, Liu L, Cheng J, Li Y, Xu Y, Li X, Zhang H, Yuan K, Chen W, Cai P, Li X. The value of using deep learning to predict proliferative hepatocellular carcinoma based on multiphasic magnetic resonance imaging. *Quant Imaging Med Surg* 2025;15:9778-91.
  6. Zhou Q, Zhou Z, Chen C, Fan G, Chen G, Heng H, Ji J, Dai Y. Grading of hepatocellular carcinoma using 3D SE-DenseNet in dynamic enhanced MR images. *Comput Biol Med* 2019;107:47-57.
  7. Gu D, Guo D, Yuan C, Wei J, Wang Z, Zheng H, Tian J. Multi-scale patches convolutional neural network predicting the histological grade of hepatocellular carcinoma. *Annu Int Conf IEEE Eng Med Biol Soc* 2021;2021:2584-7.
  8. Jia X, Sun Z, Mi Q, Yang Z, Yang D. A Multimodality-Contribution-Aware TripNet for Histologic Grading of Hepatocellular Carcinoma. *IEEE/ACM Trans Comput Biol Bioinform* 2022;19:2003-16.
  9. Liu SC, Lai J, Huang JY, Cho CF, Lee PH, Lu MH, Yeh CC, Yu J, Lin WC. Predicting microvascular invasion in hepatocellular carcinoma: a deep learning model validated across hospitals. *Cancer Imaging* 2021;21:56.
  10. Wang SH, Han XJ, Du J, Wang ZC, Yuan C, Chen Y, Zhu Y, Dou X, Xu XW, Xu H, Yang ZH. Saliency-based 3D convolutional neural network for categorising common focal liver lesions on multisequence MRI. *Insights Imaging* 2021;12:173.
  11. Wang CJ, Hamm CA, Savic LJ, Ferrante M, Schobert I, Schlachter T, Lin M, Weinreb JC, Duncan JS, Chapiro J, Letzen B. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. *Eur Radiol* 2019;29:3348-57.
  12. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQH, Nguyen CDT, Ngo V-D, Seekins J, Blankenberg FG, Ng AY, Lungren MP, Rajpurkar P. Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence* 2022;4:867-78.
  13. Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, Liang P. Concept bottleneck models. *ICML'20: Proceedings of the 37th International Conference on Machine Learning*. Article No.: 495, Pages 5338-53.
  14. Nagtegaal ID, Odze RD, Klimstra D, Paradis V, Rugge M, Schirmacher P, Washington KM, Carneiro F, Cree IA; WHO Classification of Tumours Editorial Board. The 2019 WHO classification of tumours of the digestive system. *Histopathology* 2020;76:182-8.
  15. Karahan OI, Yikilmaz A, Artis T, Canoz O, Coskun A, Torun E. Contrast-enhanced dynamic magnetic resonance imaging findings of hepatocellular carcinoma and their correlation with histopathologic findings. *Eur J Radiol* 2006;57:445-52.
  16. Huang K, Dong Z, Cai H, Huang M, Peng Z, Xu L, Jia Y, Song C, Li ZP, Feng ST. Imaging biomarkers for well and moderate hepatocellular carcinoma: preoperative magnetic resonance image and histopathological correlation. *BMC Cancer* 2019;19:364.
  17. Wu C, Du X, Zhang Y, Zhu L, Chen J, Chen Y, Wei Y, Liu Y. Five machine learning-based radiomics models for preoperative prediction of histological grade in hepatocellular carcinoma. *J Cancer Res Clin Oncol* 2023;149:15103-12.
  18. Hu X, Li C, Wang Q, Wu X, Chen Z, Xia F, Cai P, Zhang L, Fan Y, Ma K. Development and External Validation of a Radiomics Model Derived from Preoperative Gadoteric Acid-Enhanced MRI for Predicting Histopathologic Grade of Hepatocellular Carcinoma. *Diagnostics (Basel)* 2023;13:413.
  19. Wang Q, Wang A, Wu X, Hu X, Bai G, Fan Y, Stål P, Brismar TB. Radiomics models for preoperative prediction of the histopathological grade of hepatocellular carcinoma: A systematic review and radiomics quality score assessment. *Eur J Radiol* 2023;166:111015.
  20. Zheng Y, Zhou D, Liu H, Wen M. CT-based radiomics analysis of different machine learning models for differentiating benign and malignant parotid tumors. *Eur Radiol* 2022;32:6953-64.
  21. Nordin NI, Mustafa WA, Lola MS, Madi EN, Kamil AA, Nasution MD, K Abdul Hamid AA, Zainuddin NH, Aruchunan E, Abdullah MT. Enhancing COVID-19 Classification Accuracy with a Hybrid SVM-LR Model. *Bioengineering (Basel)* 2023;10:1318.
  22. Lakshmipriya B, Pottakkat B, Ramkumar G. Deep learning techniques in liver tumour diagnosis using CT and MR imaging - A systematic review. *Artif Intell Med* 2023;141:102557.
  23. Mao Y, Wang J, Zhu Y, Chen J, Mao L, Kong W, Qiu Y, Wu X, Guan Y, He J. Gd-EOB-DTPA-enhanced MRI radiomic features for predicting histological grade of hepatocellular carcinoma. *Hepatobiliary Surg Nutr* 2022;11:13-24.
  24. Cucchetti A, Piscaglia F, Grigioni AD, Ravaioli M, Cescon M, Zanello M, Grazi GL, Golfieri R, Grigioni WF, Pinna AD. Preoperative prediction of hepatocellular

- carcinoma tumour grade and micro-vascular invasion by means of artificial neural network: a pilot study. *J Hepatol* 2010;52:880-8.
25. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging* 2019;49:939-54.
  26. Rhee H, Park YN, Choi JY. Advances in Understanding Hepatocellular Carcinoma Vasculature: Implications for Diagnosis, Prognostication, and Treatment. *Korean J Radiol* 2024;25:887-901.
  27. Asayama Y, Yoshimitsu K, Irie H, Nishihara Y, Aishima S, Tajima T, Hirakawa M, Ishigami K, Kakihara D, Taketomi A, Honda H. Poorly versus moderately differentiated hepatocellular carcinoma: vascularity assessment by computed tomographic hepatic angiography in correlation with histologically counted number of unpaired arteries. *J Comput Assist Tomogr* 2007;31:188-92.
  28. Asayama Y, Yoshimitsu K, Nishihara Y, Irie H, Aishima S, Taketomi A, Honda H. Arterial blood supply of hepatocellular carcinoma and histologic grading: radiologic-pathologic correlation. *AJR Am J Roentgenol* 2008;190:W28-34.

**Cite this article as:** Hu W, Cai X, Zhao Y, Xu Q, Wang X, Song Q, Yao Y, Liu A. Interpretable deep learning framework based on contrast-enhanced MRI for predicting histological grade of hepatocellular carcinoma. *Quant Imaging Med Surg* 2026;16(1):86. doi: 10.21037/qims-2025-269