Prediction of hepatocellular carcinoma and Edmondson-Steiner grade using an integrated workflow of multiple machine learning algorithms

Likai Han  
Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College

Xiaojun Peng  
Cosmos Wisdom Biotech Co., Ltd

Xingen Hu  
Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College

Tianshi Ma  
Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College

Zhenyu Shu  
Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College

Guoqing Ru  
Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College

Junshun Gao  
Xiao Shan Economic & Technological Development Zone

Lili Yu (✉ yulili@hmc.edu.cn)  
Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College

Research Article

Keywords: machine learning, radiomics, hepatocellular carcinoma, Edmondson-Steiner grade, diagnostic

Posted Date: May 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2905568/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Early diagnosis of hepatocellular carcinoma (HCC) is indeed a great challenge. Based on traditional methods, the specificity and sensitivity of US/AFP are insufficient to detect the early onset of HCC. In this study, we constructed a prediction model for HCC diagnosis and Edmondson-Steiner (ES) grade using machine learning algorithms. The prediction model was constructed based on CT/MRI images, blood AFP, and pathological diagnosis datasets of 171 patients from Zhejiang Provincial People's Hospital. First, the automatic liver segmentation method of deep learning algorithm is used to locate the region of interest, and then PyRadiomics (engineering hard-coded feature algorithm) and Boruta (random forest algorithm) are used to extract and screen disease-related image features. By comparing the performance of various algorithms, we choose "plr" as the optimal algorithm for the HCC diagnosis model with AUC of 0.990, Kappa of 0.893 and accuracy of 0.952. "gbm" is the optimal algorithm for the ES grade prediction model with AUC 0.941, Kappa 0.777, and accuracy rate 0.902 in the TCGA-LIHC dataset. Compared with traditional diagnostic models based on clinical features, our model significantly improves the predictive performance. AUC increased from 0.733 to 0.933. This study shows that processing image data using deep learning methods can yield important features compared to conventional methods. Choosing an appropriate machine learning algorithm to build a predictive model can significantly improve the performance of disease diagnosis.

1. Introduction

Hepatocellular carcinoma (HCC) is the most common type of liver cancer and is a leading cause of cancer-related deaths globally(1). Early detection and prompt treatment are crucial in improving the prognosis and survival of HCC patients. If detected at an early stage, the five-year survival rate can be as high as 80%(2). Currently, the most commonly used diagnostic method is imaging, such as ultrasound, computed tomography (CT), or magnetic resonance imaging (MRI). Blood tests, such as alpha-fetoprotein (AFP) levels, can also be used to help diagnose HCC. However, these tests have limitations and can produce false results(3). In some cases, biopsy may be necessary for a definite diagnosis. Advances in technology, such as the use of contrast-enhanced ultrasound and laparoscopic biopsy, are helping to improve the accuracy of HCC diagnosis(4). Nevertheless, early diagnosis can still be challenging due to the lack of specific symptoms and the slow progression of the disease in its early stages.

Machine learning methods can be applied in the diagnosis of hepatocellular carcinoma (HCC), a type of liver cancer, by analyzing various medical imaging modalities such as ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI). These methods can help to automatically detect and classify liver lesions, reducing the need for human interpretation and increasing the accuracy and speed of diagnosis. In addition, machine learning algorithms can be trained on large datasets of annotated medical images, allowing for the identification of complex patterns and relationships that may be difficult for humans to detect. Some examples of machine learning methods applied to HCC diagnosis include convolutional neural networks, decision trees, and random forests.
2. Materials and Methods

2.1 Ethics Statement

This study was approved by the Ethical Committee and was in accordance with the World Medical Association (Declaration of Helsinki). Each participant was informed of the experiment procedure and signed a written informed consent before the study began.

2.2 Study design

This study was conducted through five sections shown in Fig. 1A. First, we divided the collected patients into training cohort and validation cohort. The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) database was used for independent testing cohorts. Then, the imaging data and clinical data were preprocessed separately. Imaging features distinctive of HCC could be observed by multiphasic CT and LAVA-Flex MRI. Referring to previous reports, we selected an optimal imaging protocol for modeling. In many previous studies, researchers had often used manual contouring of regions of interest (ROI), a strategy that was not only labor-intensive, but also subjective and unrepeatable. In this study, a deep learning algorithm was used to automatically segment the liver mask to construct the ROI. Then radiomics feature extraction and optimal feature selection were performed on the ROI. Clinical data includes pathological diagnosis information and the concentration of biomarkers in blood. Through statistical analysis, we selected three clinical data (gender, age and AFP) most associated with HCC for modeling. Finally, we made a comparison among all models (Radiomics model, AFP model and Radiomics-AFP model).

2.3 Study Population

We obtained data on 171 high-risk adult patients with HCC in Zhejiang Provincial People's Hospital from 2018 to 2021 who underwent MRI and CT for suspected liver tumors at 3-month intervals. The inclusion and exclusion details of patients were as follows. Inclusion criteria: patients who preoperatively underwent CT or MRI and received radical resection at Zhejiang Provincial People Hospital. Exclusion criteria: 1. Patients who had initially borderline resectable/unresectable cancers according to the NCCN guideline; 2. Patients who received neoadjuvant therapy; 3. Patients who did not receive a contrast-enhanced CT or MRI scan within two weeks before surgery; 4. Patients who did not receive AFP testing within two weeks before surgery; 5. Patients who died from surgical complications within 30 days after surgery; 6. Patients who received consecutive surgical operations.

2.4 Data splitting

The dataset was divided into training and validation cohorts (training: validation = 7:3), and random sampling was repeated until there was no significant difference (P-value > 0.05) between the two groups for all variables (Table 1). The P value was calculated using Welch t test for continuous variables and Fisher exact test for categorical variables. This resulted in allocation of 39 patients with bile duct dilation and 90 patients with hepatocellular carcinoma (HCC) to the training cohort, and 14 patients with bile duct...
dilation and 28 patients with HCC to the validation cohort. To validate the performance of the Edmondson-Steiner rank prediction model on external data, the public database TCGA-LIHC containing 41 HCC (Edmondson-Steiner grade I/II: 26, Edmondson-Steiner grade III/IV: 15) patients was used for an independent testing cohort. The details of profile and information of patients were presented in Fig. 1B and Table 1.
<table>
<thead>
<tr>
<th></th>
<th>Training (N = 129)</th>
<th>Validation (N = 42)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Envelope</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete</td>
<td>23 (17.8%)</td>
<td>7 (16.7%)</td>
<td>0.544</td>
</tr>
<tr>
<td>Incomplete</td>
<td>33 (25.6%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>34 (26.4%)</td>
<td>7 (16.7%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>39 (30.2%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td><strong>Bile duct invasion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>85 (65.9%)</td>
<td>27 (64.3%)</td>
<td>1</td>
</tr>
<tr>
<td>Unknown</td>
<td>41 (31.8%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 (2.3%)</td>
<td>1 (2.4%)</td>
<td></td>
</tr>
<tr>
<td><strong>Microvascular invasion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>53 (41.1%)</td>
<td>16 (38.1%)</td>
<td>0.971</td>
</tr>
<tr>
<td>Unknown</td>
<td>39 (30.2%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>37 (28.7%)</td>
<td>12 (28.6%)</td>
<td></td>
</tr>
<tr>
<td><strong>Lymph node metastasis</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not transferred</td>
<td>6 (4.7%)</td>
<td>0 (0%)</td>
<td>0.319</td>
</tr>
<tr>
<td>Unknown</td>
<td>123 (95.3%)</td>
<td>42 (100%)</td>
<td></td>
</tr>
<tr>
<td><strong>Adjacent organ invasion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>24 (18.6%)</td>
<td>6 (14.3%)</td>
<td>0.527</td>
</tr>
<tr>
<td>Unknown</td>
<td>102 (79.1%)</td>
<td>36 (85.7%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 (2.3%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td><strong>Vascular tumor thrombus</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>88 (68.2%)</td>
<td>27 (64.3%)</td>
<td>0.877</td>
</tr>
<tr>
<td>Unknown</td>
<td>39 (30.2%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>2 (1.6%)</td>
<td>1 (2.4%)</td>
<td></td>
</tr>
<tr>
<td>Perineural invasion</td>
<td>No</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
<tr>
<td>---------------------</td>
<td>----</td>
<td>---------</td>
<td>-----</td>
</tr>
<tr>
<td>No</td>
<td>87 (67.4%)</td>
<td>27 (64.3%)</td>
<td>0.93</td>
</tr>
<tr>
<td>Unknown</td>
<td>39 (30.2%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 (2.3%)</td>
<td>1 (2.4%)</td>
<td></td>
</tr>
<tr>
<td>Edmondson-Steiner grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>10 (7.8%)</td>
<td>2 (4.8%)</td>
<td>0.955</td>
</tr>
<tr>
<td>II</td>
<td>41 (31.8%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>34 (26.4%)</td>
<td>11 (26.2%)</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>5 (3.9%)</td>
<td>1 (2.4%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>39 (30.2%)</td>
<td>14 (33.3%)</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>44 (34.1%)</td>
<td>14 (33.3%)</td>
<td>1</td>
</tr>
<tr>
<td>Male</td>
<td>85 (65.9%)</td>
<td>28 (66.7%)</td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>60.0 (11.9)</td>
<td>62.2 (10.7)</td>
<td>0.214</td>
</tr>
<tr>
<td>Median [Min, Max]</td>
<td>60.0 [25.0, 86.0]</td>
<td>63.0 [32.0, 85.0]</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>39 (30.2%)</td>
<td>14 (33.3%)</td>
<td>0.945</td>
</tr>
<tr>
<td>ES.I.II</td>
<td>51 (39.5%)</td>
<td>16 (38.1%)</td>
<td></td>
</tr>
<tr>
<td>ES.III.IV</td>
<td>39 (30.2%)</td>
<td>12 (28.6%)</td>
<td></td>
</tr>
<tr>
<td>AFP (µg/L)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
<td>1380 (5120)</td>
<td>2810 (16200)</td>
<td>0.425</td>
</tr>
<tr>
<td>Median [Min, Max]</td>
<td>5.10 [0.600, 30300]</td>
<td>3.95 [1.10, 105000]</td>
<td></td>
</tr>
</tbody>
</table>

2.5 CT Technique

CT was performed with a dedicated scanner (Discovery; GE Healthcare, Milwaukee, WI, USA). Two-phase (arterial, 3-min delayed phase) CT was obtained using iodine contrast media (1.5 mL/kg, Iobrix 350, Acuzen, Seoul, Korea, or Optiray 350, Reyon Pharma, Seoul, Korea). Scan coverage was from the base of lung to the iliac crest. Depending on patients' weight, 80 kVp (body weight ≤ 70 kg) or 100 kVp (body weight > 70 kg) was applied. The dose-length product was automatically calculated, and the effective dose was estimated using a coefficient of 0.015.
2.6 MRI Technique

All examinations were conducted on a 3.0 T MR scanner with a 50 mT/m maximum gradient length and 200 T/m/s maximum slew rate (Discovery MR 750; GE Medical Systems, Milwaukee, WI) using a 16-channel body array coil with eight anterior and eight posterior elements, each arranged in a 4 × 4 configuration. All examinations were in the supine position. The sequences included axial LAVA-Flex MRI as well as other conventional sequences. LAVA-Flex MRI was performed during one breath-hold with the following parameters: TR/TE = 4.2/2.6 msec, 1.3 msec; flip angle = 12; matrix = 320–384 × 224; section thickness = 5 mm; intersection gap = 0 mm; one acquired signal; FOV = 26–33 cm; and bandwidth = 166.7 kHz while NEX = 0.69. The acquisition time for LAVA-Flex was 17 seconds. This acquisition was performed with an array spatial sensitivity encoding technique (ASSET; GE Healthcare) by using a recommended acceleration factor of 2.0. Dynamically enhanced imaging was performed with axial 3D LAVA sequencing. Scanning parameters included TR/TE = 4.0/1.9 msec; flip angle = 15°; matrix = 320 × 224; FOV = 26–33 cm; section thickness = 4–6 mm; and NEX = 0.75. Gadolinium chelate (Magnevist, Schering Guangzhou, China) was administered intravenously (0.2 mmol/L per kg of body weight) at approximately 3.5 mL/s using a double tube high-pressure injector (Spectris MR Injection System, Medrad, Pittsburgh, PA) and was followed by a 20 mL saline solution flush at the same speed. With the patient holding his/her breathe, two arterial phase images were achieved in 19 seconds. From injection, two 60 seconds portal phase images and one 180 seconds equilibrium phase image were obtained.

2.7 Liver segmentation

For each patient, multi-phase CT and multi-sequence MRI methods were used for liver image acquisition. For each method, we tried to extract liver region masks using an automated liver segmentation algorithm. Arterial phase images were used for liver segmentation because most HCC lesions display arterial phase hyperenhancement (APHE), which is reflected in the current LI-RADS criteria. For CT imaging, ‘Arterial Phase 5.0’ protocol was selected. And for MRI imaging, ‘Ax LAVA-Flex + C’ protocol was selected. In this study the state-of-the-art deep learning 3D liver segmentation algorithm(8) was used to automatically segment CT and MRI imaging data. We used the automatic segmentation tool liver_segm (https://github.com/OnofreyLab/liver-segm/) to obtain the segmented mask of the liver. Briefly, first convert the image data in dcm format into nii format file. Then, the input images were standardized to have isotropic voxel spacing of 2mm3 and intensities were scaled so that the 25th and 75th percentile ranged between −0.5 and +0.5. A deep learning neural network model built from a previous study of early, intermediate, and advanced HCC cancer subjects was used for prediction.

2.8 Radiomics feature extraction

To reduce the variation in different scanners, a two-step method of the image preprocessing was conducted before radiomics features extraction. Firstly, due to different pixel sizes and slice thicknesses of various scanners, all the slices were resampled to 1 × 1 × 1 mm3 using the bicubic interpolation(9). Secondly, the images were normalized to 64 grey levels to compensate for the variation of scanners. For each liver segmented mask, various features that can be extracted using PyRadiomics(10). Radiomics
features were extracted covering 19 first order features, 16 3D shape features, 10 2D shape features, 24 gray level co-occurrence matrix (GLCM) features, 16 gray level size zone matrix (GLSZM) features, 16 gray level run length matrix (GLRLM) features, 5 neighbouring gray tone difference matrix (NGTDM) features, 14 gray level dependence matrix (GLDM) features. Therefore, we could obtain 120 radiomics features for each patient(11). Radiomics features extraction process was conducted using the PyRadiomics tools available in github git://github.com/Radiomics/pyradiomics. More details about the feature extraction methods could be found in the study by URL https://pyradiomics.readthedocs.io/en/latest/index.html.

2.9 Radiomics feature selection

Before the process of feature selection, the extracted features were normalized using the Z-score, with the mean and standard deviation of the features in the training set utilized to normalize the corresponding features in the validation set. To reduce the dimension of radiomics features, we exported Boruta algorithms for feature selection in the training set(12). Boruta iteratively compares importances of attributes with importance of shadow attributes, created by shuffling original ones. Attributes that have significantly least importance than shadow ones are being consecutively dropped. On the other hand, properties that are significantly better than shadows are acknowledged and confirmed. Shadows are re-created in each iteration. Algorithm stops when only confirmed attributes are left, or when it reaches 500 importance source runs. If the second scenario occurs, some attributes may be left without a decision. They are claimed tentative. The importance scores of features were calculated according to the method described above, and the top 10 features most associated with HCC and Edmondson-Steiner grade were selected.

2.10 Model construction

In this study, 7 types of supervised machine learning classifiers, including Gradient Boosting Machine (GBM), Support Vector Machine (SVM), Random Forest (RF), Neural Network (NN), Penalized Logistic Regression (PLR), Elastic Net (EN), and Decision Tree (DT) were assessed. All classifiers were implemented using R package caret(13) (method "gbm" for GBM, "svmRadial" for SVM, "ranger" for RF, "avNNet" for NN, "glmnet" for EN and "C5.0" for DT). Classifiers were trained using repeated 10-fold cross-validation of training dataset, and their predictive performance was evaluated in the test dataset. We defined maximum value of area under the receiver operating characteristic (ROC) curve (AUC) as best performance. To calculate the predicted variable importance, 100 sets of independent training were performed using different random seed. The median of variable importance obtained in each training was used as a representative value. Each variable importance is calculated with varImp function of the caret package.

2.11 Statistical Analysis

Statistical analysis was conducted in R v4.1.3 (www.Rproject.org). Feature selection, Model construction, Delong’s test, ROC and AUC were performed on the packages "Boruta", "caret", "pROC", respectively. The
reported significance levels were two-sided and set at 0.05.

3. Result

3.1 Basic Characteristics

AFP is the most widely used serum biomarker for HCC detection(3). However, no AFP elevation have been found in many HCC and AFP analysis can't be used to screen HCC in these cases(14–16). Here, we systematically assessed preoperative blood AFP levels in 171 patients with suspected HCC in a retrospective study (Fig. 2). This study included 118 patients with HCC (67 patients with ES I/II, and 51 patients with ES III/IV) and 53 patients with bile duct dilatation. Wilcoxon test was used to estimate significant differences in AFP levels in blood between different HCC ES grades and bile duct dilatation. And receiver operator characteristic analysis was performed to evaluate sensitivity and specificity. As shown in Fig. 2A, there was a significant difference in AFP levels between bile duct dilation (control) and HCC (p-value < 2.22e-16). However, the difference in AFP levels between ES I/II and ES III/IV was less significant (p value = 0.0013). Receiver operating characteristic analysis showed that AFP as a diagnostic marker for HCC has good specificity but poor sensitivity. Area under curve (AUC): 0.851, 95% confidence interval: 0.796, 0.907; with cut-off value: 11 ng/mL, sensitivity: 58.5%, specificity: 98.1%. When predicting ES grade by AFP level alone, its performance falls far short of clinical expectations. AUC: 0.68, 95% confidence interval: 0.582, 0.777; with cut-off value: 9 ng/mL, sensitivity: 82.4%, specificity: 53.7%. Given the poor sensitivity of AFP in HCC diagnosis, we need to combine more clinical and imaging data to improve the sensitivity of HCC diagnosis. Given the poor specificity of AFP in ES grade prediction, we can also try to combine this information with machine learning to predict HCC ES grade to guide the treatment of patients. We attempted to combine AFP with patient age, gender, and CT/MRI imaging data for HCC diagnosis. Through the analysis of pathological diagnosis data of HCC patients, the results showed that 'Microvascular invasion' was a pathological factor significantly (p = 0.014) associated with ES grade (Table S1). Therefore, we included the pathological diagnostic information "microvascular invasion" as an additional clinical factor for ES grade prediction.

3.2 Prediction of HCC with AFP, Clinical and Radiomics Features

This study developed and compared three types of predictive models, including AFP + clinical models, radiomics + clinical models, and AFP + radiomics + clinical models. By Radiomics feature selection analysis, top 10 features most associated with HCC were selected by the Boruta algorithm(12) to build models. The final selected features for modeling were shown in Fig. 3. In order to construct a model with the best predictive performance, 7 types of supervised machine learning algorithms were explored. First, models based on different machine learning algorithms are built in the training set. Then AUC, accuracy, Kappa, sensitivity and specificity were calculated in the validation set to evaluate the predictive performance of the model (Figure S1). The model with the maximum AUC value was selected as the optimal model. The results showed that "plr" was the optimal algorithm for building three model. The predictive performance of the AFP-clinical model in the validation set is: AUC of 0.941, Kappa of 0.786,
and accuracy of 0.905. The predictive performance of the radiomics + clinical model was the validation set is: AUC of 0.967, Kappa of 0.945, and accuracy of 0.976. The predictive performance of the AFP + radiomics + clinical model in the validation set is: AUC of 0.990, Kappa of 0.893, and accuracy of 0.952. To evaluate the role of imaging features in the prediction model, we screened out the cases with blood AFP levels below 11ug/ml from the training set and test set, and defined them as the AFP negative dataset. Based on the optimal model selected above, we compared the predictive performance between three models in validation dataset (Fig. 4A) and AFP negative dataset (Fig. 4B). Compared with the other two models in the validation set, the AFP + radiomics + clinical model has slightly improved performance in AUC. However, in the AFP negative dataset, the AFP + radiomics + clinical model has a significant improvement in the predictive performance compared to the AFP + clinical model. AUC increased from 0.733 to 0.933, Sensitivity increased from 0.513 to 0.923. By the Delong test, the AFP + radiomics + clinical model was preferred, there was significant (p value = 0.001) difference in the predictive performance compared with AFP + clinical model in the AFP negative set (Table S2). The result indicated that imaging features can significantly improve the detection rate of patients with AFP-negative HCC.

3.3 Prediction of Edmondson-Steiner grade with AFP, Clinical and Radiomics Features

Based on AFP and radiomics feature, we also tried to predict Edmondson-Steiner grade of HCC. Edmondson-Steiner grade, a histological classification, carries robust prognostic implications for all the endpoints for prognosis(17). The tumors were classified according to the Edmondson–Steiner grade and separated as well-differentiated and non-well-differentiated (moderately and poorly differentiated)(18). In this study, we defined HCC patient with ES I/II as well-differentiated, and ES III/IV as non-well-differentiated. By Radiomics feature selection analysis, top 10 features most associated with ES grade were selected by the Boruta algorithm(12) to build models. The final selected features for modeling were shown in Fig. 3. As previously described, three models (AFP + clinical model, radiomic + clinical model, and AFP + radiomic + clinical model) were constructed to predict Edmondson-Steiner grade of HCC. First, models based on different machine learning algorithms are built in the training set. Then AUC, Accuracy, Kappa, Sensitivity and Specificity was calculated in the validation set to evaluate the predictive performance of the model (Figure S2). The model with the maximum AUC value was selected as the optimal model. The results showed that “gbm” was the optimal algorithm for building the three models. Based on the optimal model, we compared the predictive performance between the AFP model, radiomics model and radiomics-AFP in validation set Fig. 5A. The predictive performance of the AFP-clinical model in the validation set is: AUC of 0.768, Kappa of 0.563, and accuracy of 0.786. The predictive performance of the radiomics + clinical model was the validation set is: AUC of 0.854, Kappa of 0.774, and accuracy of 0.893. The predictive performance of the AFP + radiomics + clinical model in the validation set is: AUC of 0.969, Kappa of 0.851, and accuracy of 0.929. By Delong's test, the difference in prediction performance was significant (p-value = 0.04), when AFP + radiomic + clinical model compared to AFP + clinical model in the validation set (Table S3).

In order to explore the abnormal performance of the model in the validation set, we obtained 41 HCC patients data from the TCGA-LIHC database for an independent testing cohort. The results was satisfying
(Fig. 5B), the prediction performance of the AFP + radiomic + clinical model in the testing cohort (AUC 0.941, Kappa 0.777 and accuracy 0.902) was better than AFP + clinical model. By Delong's test, the difference was significant compared to AFP + clinical model (p-value = 0.0244) in the testing set (Table S3). The model showed equally good predictive performance on the independent testing set as the validation set.

4. Discussion

A powerful risk group HCC prediction tool with satisfactory specificity and sensitivity could help physicians in the clinical diagnosis of HCC, thereby minimizing unnecessary treatment of patients and improving the survival rate of HCC patients. In this study, we constructed and validated a model integrating CT/MRI radiomics and AFP to predict potential HCC patients in a high-risk population for HCC. The results showed that the model significantly outperformed compared to the traditional blood AFP concentrations or independent imaging information in predictive performance (especially sensitivity).

AFP testing remains far from satisfactory for early HCC, owing to insufficient sensitivity. Using 20 ng/mL as a cut-off, the reported sensitivities of AFP for HCC at any stage in cirrhotic patients ranged from 41–65%, with specificities ranging from 80–94% (19). Whilst during the early stages of HCC progression, detection rates were as low as 1/3 (20). Unlike other cancer, HCC can be diagnosed based on imaging features alone (4). Imaging-based diagnosis is recommended by nearly all HCC guidelines. Early diagnosis provides survival benefits to those with HCC but imaging screening lacks sensitivity with an ~84% (21, 22) accuracy for all HCC stages with an accuracy ranging from 47 to 63% in cirrhotic HCC patients at the early disease stages (22, 23). Confounded liver disease and obesity result in a lack of sensitivity of ultrasonography. Furthermore, the interpretation of imaging data is relied on the experienced physicians, and the challenge of standardization remains. In this study, we developed an automated process for extracting image features of the liver, which greatly reduced the degree of interference to the results caused by the doctor's empirical error. Using a powerful machine learning algorithm, the image features, clinical factors and AFP were combined to build a powerful and robust (AUC 0.990, Kappa 0.893 and accuracy 0.952) HCC prediction model. In particular, for the AFP negative population, the model also showed excellent predictive performance (AUC 0.933, Kappa 0.727 and accuracy 0.865). The results suggest that the poor sensitivity due to the absence of elevated AFP in many HCC patients can be well resolved by imaging features.

Edmondson–Steiner (ES) histopathological grading system was published in the far 1954 (24), and it had been used to grade HCCs pathologically ever since. ES grading of tumor has a predictive importance for survival and recurrence, therefore, it is important to know the ES stage of the tumor when planning treatment (35). However, the ES grade of the tumor can only be determined by core needle biopsy or surgical resection. None of the imaging methods can noninvasively determine the histopathological stage of HCC. In this study, we tried to build a model for predicting ES grade based on imaging, clinical factors and serum AFP level. As the results show, the AFP + radiomic + clinical model exhibits excellent predictive performance on the validation set (AUC = 0.969). Compared with the AFP + clinical model, the prediction
performance is significantly improved \((p = 0.04)\). Considering that the image data in the training set and testing set come from the same institution and the sample size is limited, the performance evaluation of the model may not be universal. So we applied the TCGA-LIHC dataset as a testing set to evaluate the predictive performance of the model in an independent dataset. The results were satisfactory. The AFP + Radiomic + clinical model showed the same excellent prediction performance in the testing set as in the validation set (AUC 0.941, Kappa 0.777 and accuracy 0.902). In particular, the radiomic + clinical model showed surprisingly consistent predictive performance in the validation and testing sets. These results indicate that image features extracted by machine methods are more stable and reliable in predicting ES grade than AFP, especially when dealing with multiple institutions.

The importance of the variables of the AFP + radiomics + clinical model was calculated based on the "varImp" method. The results in Fig. 6A shows that AFP is the most important variable in models for HCC diagnosis and ES grade prediction. For radiomic features, ‘original_firstorder_Skewness’ and ‘original_shape_Sphericity’ were most importance variable of the HCC diagnosis model. And ‘original_shape_LeastAxisLength’ was most importance variable of ES-grade prediction model.

Different machine learning algorithms search for different trends and patterns. One algorithm is not the best on all datasets or all use cases. In this study we choose the optimal algorithm by the performance of the final model. In this study, we evaluate the algorithm by comparing the AUC values of the model in the training and validation sets. Evaluation can be divided into three states: underfitting, overfitting and optimum. Underfitting: the model is too simple to capture the patterns within the data. The model performs poorly on both trained and unseen data; Overfitting: The model is too complicated or too specific, capturing trends that don't generalize. The model accurately predicts data that it was trained on but doesn't accurately predict unseen data; Optimum: The model is able to capture trends in the data well. The model accurately predicted both the data it was trained on and the unseen data. From the results shown in Fig. 6B, we can consider that the point that deviates significantly below the diagonal is overfitting, the point that is located at the lower left of the diagonal is underfitting, and the point that is located at the top right of the diagonal is the best. To conclude: models based on the 'ranger' (random forest) algorithm are prone to overfitting in this study; models based on the 'avNNet' (Neural Network) algorithm are prone to underfitting; 'plr' (Penalized Logistic Regression) and 'gbm' (Gradient Boosting Machine) algorithm are located in the top right corner of the graph, and are the best algorithms for the HCC diagnosis model and ES grade prediction model, respectively.

Some limitations exist during the study. First, given the retrospective nature of this study, which may introduce some selection bias, the results reported here need to be confirmed in prospective trials. Second, the sample size may limit the universality of the study. Further studies based on the large-scale datasets from multiple medical institutions are also required to establish the proposed robustness.

5. Conclusions
In summary, we developed and validated a grading prediction model for HCC and ES that combines clinical and radiomic data. The model adopts automatic radiomics feature extraction algorithm and comprehensive evaluation method based on multiple machine learning algorithms. Compared with traditional diagnostic methods, the model has improved significantly in many predictive performance indexes. This study shows that processing image data using deep learning methods can yield important features compared to conventional methods. Choosing an appropriate machine learning algorithm to build a predictive model can significantly improve the performance of disease diagnosis.

**Declarations**

**Ethics declarations**

**Ethics approval and consent to participate**

The studies involving human participants were reviewed and approved by the Institutional Review Board Committee of Zhejiang Provincial People's Hospital, Affiliated People's Hospital of Hangzhou Medical College. Written informed consent for participation was not required for this study in accordance with the institutional requirements.

**Consent for publication**

Not applicable.

**Availability of data and materials**

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

**Competing interests**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Funding**

This work was supported by grants from Zhejiang Provincial Medicine and Health Research Foundation (2021KY072, 2021KY535).

**Authors' contributions**

LH and LY designed the research. LH and XP wrote the manuscript and drafted the figures and tables. XH and TM collected clinical cases. ZS performed radiological analysis. XP and JG contributed to the data analysis. LY and GR revised the manuscript. All authors contributed to the article and approved the submitted version.
Acknowledgements

We would like to thank our clinical research associates Qinqin Lou for their help in data collection.

References


Figures
Figure 1

A) The workflow of this study. The workflow mainly includes 4 steps: 1) data segmentation into training set; validation set and test set; 2) feature extraction and selection of CT and MRI image data; 3) clinical data cleaning and selection of important indicators; 4) based on AFP and images Features build models; B) Population distribution of patients with bile duct dilatation and ES grade in training, validation, and test sets (TCGA-LIHC); C) Preoperative CT, MRI and blood AFP detection status distribution of patients in training, validation, and test sets (TCGA-LIHC).

Figure 2

A) Differential analysis of AFP in the blood of patients between HCC and bile duct dilatation (Control). The results showed that the concentrations of AFP in the blood of HCC patients were significantly higher than those of control (T-test p value: 2.22e-16). B) Receiver operating
characteristics analysis of blood AFP for HCC diagnosis and ES-grade prediction. C) Confusion matrix analysis of blood AFP for HCC diagnosis and ES-grade prediction.

**Figure 3**

Imaging features used to build predictive models. The importance scores of imaging features were calculated using the Boruta method. Filtered out 'rejected' features, the top 10 scoring imaging features were selected for building the predictive model.
A) Performance evaluation of HCC diagnosis model based on the optimal algorithm in the validation set. "plr" as optimal algorithms were applied to AFP+clinical model, radiomic+clinical model and AFP+radiomic+clinical models respectively. Compared with AFP+clinical models, the AFP+radiomics+clinical model has slightly significant improvement in AUC kappa and accuracy; B) Performance evaluation of HCC diagnosis model based on the optimal algorithm in the AFP negative set.
In the AFP negative dataset, the AFP+radiomics+clinical model has a significant (p value=0.001) improvement in the predictive performance compared to the AFP+clinical model. AUC increased from 0.733 to 0.933, Sensitivity increased from 0.513 to 0.923.

Figure 5

A

![ROC Curve for AFP+Clinical, Radiomics+Clinical, and AFP+Radiomics+Clinical models]

B

![ROC Curve for AFP+Clinical, Radiomics+Clinical, and AFP+Radiomics+Clinical models]
A) Performance evaluation of ES grade prediction model based on the optimal algorithm in the validation set. “gbm” as optimal algorithms were applied to AFP+clinical model, radiomic+clinical model and AFP+radiomic+clinical models respectively. By Delong's test, the difference in prediction performance was significant (p-value = 0.04), when AFP+radiomic+clinical model compared to AFP+clinical model in the validation set; B) Performance evaluation of ES grade prediction model based on the optimal algorithm in the TCGA-LIHC dataset. The prediction performance of the AFP+radiomic+clinical model in the testing cohort (AUC 0.941, Kappa 0.777 and accuracy 0.902) was better than AFP+clinical model. By Delong's test, the difference was significant compared to AFP+clinical model (p-value=0.0244) in the testing set.
Figure 6

A) The importance of the variables of the AFP+radiomics+clinical model. The importance score was calculated based on the "varImp" method; B) Performance comparison between models built by different algorithms. 'plr' (Penalized Logistic Regression) and 'gbm' (Gradient Boosting Machine) algorithm are located in the top right corner of the graph, and are the best algorithms for the HCC diagnosis model and ES grade prediction model, respectively.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx