Preoperative prediction of pathological grading of hepatocellular carcinoma using machine learning-based ultrasomics: A multicenter study

Shanshan Ren\textsuperscript{a,b}, Qinghua Qi\textsuperscript{c}, Shunhua Liu\textsuperscript{b}, Shaobo Duan\textsuperscript{a,b,d}, Bing Mao\textsuperscript{a,b,d}, Zhiyang Chang\textsuperscript{a,b,d}, Ye Zhang\textsuperscript{a,b,d}, Shuangyang Wang\textsuperscript{a,b,d}, Lianzhong Zhang\textsuperscript{a,b,*}

\textsuperscript{a} Henan University People’s Hospital, Zhengzhou, Henan, China
\textsuperscript{b} Henan Provincial People’s Hospital, Zhengzhou, Henan, China
\textsuperscript{c} First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
\textsuperscript{d} Zhengzhou University People’s Hospital, Zhengzhou, Henan, China

ARTICLE INFO

Keywords:
- Hepatocellular carcinoma
- Machine learning
- Ultrasomics
- Pathological grading
- Ultrasound

ABSTRACT

Purpose: The present study investigated the value of ultrasomics signatures in the preoperative prediction of the pathological grading of hepatocellular carcinoma (HCC) via machine learning.

Methods: A total of 193 patients were collected from three hospitals. The patients from two hospitals (\(n = 160\)) were randomly divided into training set (\(n = 128\)) and test set (\(n = 32\)) at a 8:2 ratio. The patients from a third hospital were used as an independent validation set (\(n = 33\)). The ultrasomics features were extracted from the tumor lesions on the ultrasound images. Support vector machine (SVM) was used to construct three preoperative pathological grading models for HCC on each dataset. The performance of the three models was evaluated by area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and accuracy.

Results: The ultrasomics signatures extracted from the grayscale ultrasound images could successfully differentiate between high- and low-grade HCC lesions on the training set, test set, and the independent validation set (\(p < 0.05\)). On the test set and the validation set, the combined model’s performance was the highest, followed by the ultrasomics model and the clinical model successively (\(p < 0.05\)). Their AUC (along with 95% CI) of these models was 0.874 (0.709–0.964), 0.789 (0.608–0.912), 0.720 (0.534–0.863) and 0.849 (0.682–0.949), 0.825 (0.654–0.935), 0.770 (0.591–0.898), respectively.

Conclusion: Machine learning-based ultrasomics signatures could be used for noninvasive preoperative prediction of pathological grading of HCC. The combined model displayed a better predictive performance for pathological grading of HCC and had a stronger generalization ability.

1. Introduction

Hepatocellular carcinoma (HCC) is the sixth most common cancer in the world [1] and the second most common cause of cancer-related deaths [2]. Hepatectomy is the preferred treatment for early HCC patients with normal liver function [3,4]. However, HCC is easy to relapse and metastasize, and associated with a poor prognosis [5]. Studies have shown that recurrence was closely related to pathological grading [6,7].

For HCC cases with isolated lesions \(< 2\) cm, the 3- and 5-year recurrence-free survival rates for low-grade HCC were 64 and 50% respectively, while those for high-grade HCC were 39 and 29% respectively [8]. It indicates that high-grade HCC has a higher recurrence rate than low-grade HCC. Moreover, compared with high-grade HCC, low-grade HCC also has a higher surgical cure rate and higher short- and long-term survival rate [9–11]. Therefore, an accurate prediction of the pathological grading of HCC is of high importance for clinical decision-making.

Abbreviations: AUC, area under the curve; ICC, intra-class correlation coefficient; ROC, receiver operating characteristic; AFP, alpha-fetoprotein; ALT, alanine aminotransferase; AST, aspartate aminotransferase; PT, prothrombin time; HCC, hepatocellular carcinoma; MVI, microvascular invasion; MWA, microwave ablation; RFA, radiofrequency ablation; DICOM, digital imaging and communications in medicine; EHR, Electronic health records; ES, edmondson-steiner; GLCM, gray-level co-occurrence matrix; GLDM, gray-level dependence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; NGTDM, neighboring gray tone difference matrix; ROI, region of interest; ML, machine learning; LASSO, least absolute shrinkage and selection operator; SVM, support vector machine; CI, confidence interval.

* Corresponding author at: Henan Provincial People’s Hospital, 7 Weiwu Road, Jinshui District, Zhengzhou, Henan, 450003, China.
E-mail address: llw0624@gs.zzu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.ejrad.2021.109891

Received 1 May 2021; Received in revised form 6 July 2021; Accepted 29 July 2021

Available online 28 August 2021

0720-048X/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license.
making, treatment regimen optimization, and prognostic prediction.

Tumor biopsy or postoperative histopathological examination of the resected tumor samples is the gold standard for HCC diagnosis and pathological grading [12–14]. However, biopsy may be associated with sampling error and pathologist assessment subjectivity, and it may be difficult to manage a negative biopsy result [15,16]. Currently, noninvasive imaging techniques are oftentimes used for diagnosis of HCC, but it is still needed a noninvasive radiological method for HCC grading [17–19]. Ultrasound has been recommended by several guidelines as the most commonly screening method for populations at a high risk for HCC due to its low cost, ease of operation, fast result interpretation, real-time reproducibility, and absence of radiation exposure [12–14,20,21]. However, conventional ultrasound images may not reveal the rich information of tumor heterogeneity [22]. Therefore, conventional ultrasound images cannot be directly used for prediction of the HCC grade.

Radiomics is an emerging science, which extracts a massive volume of information from radiographic images with high throughput for deeper digging, prediction, and analysis to assist physicians in making more accurate diagnoses [23–25]. Many reports have shown that radiomics is applicable to preoperative pathological grading of HCC. Mao et al. employed the radiomics signatures extracted from contrast-enhanced computed tomography (CECT) to predict the pathological grading of HCC [26]. Wu et al. studied the correlation between the radiomics signatures based on MRI and the pathological grading of HCC [27]. However, there has been no multicenter reports on using ultrasound features to predict the pathological grading of HCC. Given the facts above, this study was intended to investigate the value of ultrasound signatures in preoperative prediction of pathological grading of HCC, which was verified on an independent validation set.

2. Materials and methods

2.1. Study population

This study adopted a retrospective design involving three hospitals. The present study was approved by the ethics committee, and informed consent was waived. Clinical and ultrasound data were collected from 1370 patients pathologically confirmed as HCC from January 2019 to March 2021 at three hospitals. Among them, 193 patients (70 high-grade patients and 123 low-grade patients) were included in the final analysis. The inclusion criteria were as follows: (1) Being pathologically confirmed as HCC; (2) Liver ultrasound scan within one month before surgery and the ultrasound data were complete; (3) No history of anti-tumor treatments, including liver transplantation (LT), microwave ablation (MWA), radiofrequency ablation (RFA), and transcatheter arterial chemoembolization (TACE); (4) The ultrasound images satisfied the analytical requirements (clear without artifacts, and the target lesions were completely visible on the ultrasound images); (5) No history of other malignancies.

First, the patients from two hospitals (Henan Provincial People’s Hospital and the First Affiliated Hospital of Zhengzhou University, n = 160) were mixed and randomly divided into training set (n = 128) and test set (n = 32) at a 8:2 ratio. The patients from the third hospital (Henan Cancer Hospital, n = 33) were used as an independent validation set. The processes of inclusion and exclusion of study population are shown in Fig. 1.
2.2. Clinicopathological features of patients

The clinical data were acquired from electronic health records of the patients, including demographics (gender, age, history of hepatitis), laboratory test results (alpha-fetoprotein (AFP), alanine aminotransferase (ALT), aspartate transaminase (AST), total bilirubin (TB), and prothrombin time (PT)), and ultrasound features (maximum lesion diameter and position of lesions). The data of laboratory tests and ultrasound imaging were collected within one month before surgery.

Pathology information of the patients (the pathological grading of tumors) was acquired from the pathology information system. In the present study, low-grade HCC referred to tumor lesions of Edmondson-Steiner (ES) grade I, I-II and II; high-grade HCC referred to those of ES grade II–III, III, III-IV and IV[6].

2.3. Imaging acquisition and tumor segmentation

All ultrasound data of liver tumors were collected using the ultrasound machines with a convex array transducer (frequency range 2.5–6 MHz), including GE Logiq E9, GE Vivid E9, HI VISION Ascendus, HI ALOK ProSound A5, Philips EPIQ 5, and ALOKA EZU-MT28-S1. Before the liver ultrasound scan, all patients were fasted for over 8 h. All ultrasound scans were performed by ultrasound physicians with over nine years of experience in liver ultrasound. First, the entire liver was scanned by grayscale ultrasound. Next, the maximum diameter, echo signals, blood flow signals, shape, boundaries, and margins of the target lesions were evaluated and recorded. At least one original ultrasound image showing the lesion and the same image containing the measurement parameters were stored in the DICOM format.

The grayscale ultrasound images were loaded into the ITK-SNAP v.3.6.0 for manual segmentation (open-source software: www.itk-snap.org)[28]. Next, one reader with over 9 years of abdominal ultrasound experience manually annotated the regions of interest (ROI). To assess the reproducibility of the features, another reader with 30 years of abdominal ultrasound experience delineated ROI alone. Intra-class correlation coefficient (ICC) was calculated for each feature. Only the features with an ICC value equal to or higher than 0.80 that indicates excellent reproducibility were included in the further feature selection process. The ROI segmentation results in the representative liver lesions are shown in Fig. 2.

2.4. Image preprocessing

Before feature extraction, a researcher with five years of experience undertook image preprocessing as follows. First, the ultrasound images were normalized based on the mean and standard deviation. Next, the images were resampled by B-spline interpolation to 1 mm × 1 mm pixel. Finally, gray-level discretization was performed in the histogram with a fixed bin width of 25[29].

2.5. Feature extraction and selection

Ultrasound features were extracted by using the Open source Python package Pyradiomics v.2.1.2 for each patient[30]. The extracted features were divided into the following seven categories[31]: I. first-order statistical properties; II. two-dimensional shape features; III. gray-level co-occurrence matrix (GLCM); IV. gray-level run length matrix (GLRLM); V. gray-level size-zone matrix (GLSZM); VI. gray-level dependence matrix (GLDM); VII. neighborhood gray-tone-difference matrix (NGTDM). Fourteen filters were run on the original images to obtain the derivative images of each patient. Except for the shape features, the features of all categories were calculated from the original and derivative images. Detailed information about the feature extraction method and filters can be found in the supplementary information 1.

Before feature selection, the features were first normalized (Z-score normalization) to ensure a relatively uniform distribution of the image features. It consisted of transformation of each feature by subtracting the mean value for centralization and dividing by the standard deviation for scaling.

However, all of the extracted features were high-dimensional. The use of high-dimensional features might have the problems of low computational efficiency and overfitting. First, the features with zero variance were excluded by using the variance filtering method. Next, lasso method was performed for further dimensionality reduction of the features and the most valuable features were selected. The 10-fold cross-validation process was repeated 1,000,000 times to obtain the optimal value of parameter λ, which was introduced into the lasso method to calculate the regression coefficients of each feature. Finally, the features with non-zero coefficient were selected. The overall flowchart of the study is outlined Fig. 3.

2.6. Model construction based on machine learning

Python scikit-learn 0.23.2 package was used for SVM modeling and evaluation. The patients from two hospitals were randomly divided into the training set and the test set at 8:2 ratio. The patients from the third hospital were used as an independent validation set. The learning curve and the grid search were employed to select the optimized parameter combination of the kernel function, coefficient of the kernel function, penalty coefficient, and class_weight. The specific process of parameter tuning can be found in the supplementary information 3.

Three models were constructed in this paper. First, the clinical model was constructed using the patients’ clinical features, including gender, age, history of hepatitis, AFP, ALT, AST, TB, PT, maximum lesion diameter, and position of lesions. Next, the ultrasonomics model was constructed using ultrasonomics signatures extracted and screened from the skewed ROI on the ultrasound images of HCC. Finally, the combined model was constructed by integrating the clinical features and the ultrasonomics signatures.

The three models built upon the training set were evaluated by the test set and the independent validation set. The predictive performance of the three models was evaluated by plotting the Receiver Operating Characteristic (ROC) curve and calculating the area under the curve (AUC). The results were compared using the Wilcoxon rank-sum test. The model with the highest AUC and the smallest p-value was considered to have the best performance.

![Fig. 2. Example of delineating region of interest (ROI) on grayscale ultrasound images. A and B is a patient with high grade HCC, C and D is a patient with low grade HCC.](image-url)
2.7. Statistical analysis

SPSS 25.0 software was used for statistical analysis. The normality of continuous variables was tested using the Kolmogorov-Smirnov test. Continuous variables obeying a normal distribution were analyzed by independent samples t-test. Those not obeying a normal distribution were analyzed by Wilcoxon’s rank-sum test. Categorical variables were compared by using the chi-square test. Unless otherwise specified, the continuous variables were expressed as median (interquartile range, IQR); the categorical variables were expressed as n (%). P < 0.05 indicated a significant difference.

Reproducibility of feature extraction was evaluated by intra-class correlation coefficient (ICC). ICC ≥ 0.8 indicated a high consistency, 0.5–0.79 medium consistency, and < 0.5 low consistency[33,34].

3. Results

3.1. Clinicopathological features of the patients

The clinicopathological features in the training set, test set and independent validation test are shown in Table 1. The percentages of intermediate- and high-grade HCC patients in the training set, test set, and independent set were 39.8% (51/128), 21.9% (7/32) and 36.4% (12/33), respectively. The percentages of HCC patients with a history of hepatitis were 78.9% (101/128), 71.9% (23/32), and 90.9% (30/33), respectively. The average age of patients was 56.8 ± 10.6 years, 56.4 ± 8.5 years, and 57.6 ± 11.4 years in the training set, test set, and independent validation test, respectively. The three sets were not statistically different in demographics, laboratory test results, and ultrasound features (p > 0.05).

3.2. Reproducibility and feature selection

From each patient, 1,409 features were extracted from the ultrasound images. Among the extracted features, the fourteen 2D shape features were acquired only from the original images. Features of the other six categories were all acquired from one original image plus 14 derivative images. There were 18 first-order statistical properties, 24 GLCM features, 16 GLRLM features, 16 GLSZM features, 14 GLDM features, and 5 NGTDM features.

Among the extracted features, 207 features with ICC < 0.8 were excluded and 16 features with zero variance were excluded using the variance filtering method. After dimensionality reduction with lasso regression, 10 features were finally selected (Fig. 4). These 10 features were used for modeling, and detailed information can be found in the supplementary information 2 and Fig. S1 and Fig. S2.

3.3. Predictive ability of the clinical model and ultrasomics model

Table 2 displays the performance of the clinical model and the ultrasomics model in differentiating between low- and high-grade HCC on the training set, test set, and independent validation set. On the test set, AUC (along with 95% CI), sensitivity, specificity, and accuracy of the clinical model and the ultrasomics model were 0.720(0.534–0.863), 0.571, 0.720, 0.688 vs 0.789(0.608–0.912), 0.714, 0.840, 0.813, respectively. The overall flowchart of the study is outlined Fig. 3. The Radiomics Quality Score (RQS) for our ultrasomics process can be found in the supplementary information 5[32].

Fig. 3. Overall flowchart of the study, including image acquisition and segmentation, feature extraction and feature selection, and model construction and evaluation.
Table 1

| The clinicopathological features in the training set, test set and validation set. |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                 | Training set (n = 128) | Test set (n = 32) | P value | Validation set (n = 33) | P value |
| Gender                          |                   |                   |         |                   |         |
| Male                            | 102(79.7)         | 24(75.0)          | 0.56    | 26(78.8)          | 0.91    |
| Female                          | 26(20.3)          | 8(25.0)           |         | 7(21.2)           |         |
| Age (years)                     | 56.8 ± 10.6       | 56.4 ± 8.5        | 0.85    | 57.6 ± 11.4       | 0.70    |
| Liver diseases                  |                   |                   |         |                   |         |
| Hepatitis                       | 101(78.9)         | 23(71.9)          | 0.39    | 30(90.9)          | 0.11    |
| Other                           | 27(21.1)          | 9(28.1)           |         | 3(9.1)            |         |
| AFP (ng/ml)                     | * 38.5 ± 24.2     | 0.77              | 74.7    | 0.31              |         |
| ALT (IU/L) *                    | 29.0 ± 11.3       | 0.96              | 28.0    | 0.82              |         |
| AST (IU/L) *                    | 35.0 ± 11.3       | 0.69              | 34.8    | 0.99              |         |
| TB (umol/L) *                   | 12.9 ± 3.2        | 0.50              | 15.1    | 0.09              |         |
| PT(s) *                         | 12.2 ± 11.3       | 0.70              | 11.9    | 0.32              |         |
| Dmax (mm) *                     | 44.5 ± 29.0       | 0.94              | 46.0    | 0.63              |         |
| Position of lesions             |                   |                   |         |                   |         |
| Right lobe                      | 96(75.0)          | 26(81.2)          | 0.46    | 24(72.7)          | 0.79    |
| Left lobe                       | 32(25.0)          | 6(18.8)           |         | 9(27.3)           |         |
| Pathological grade              |                   |                   |         |                   |         |
| Low grade                       | 77(60.2)          | 25(78.1)          | 0.06    | 21(63.6)          | 0.72    |
| High grade                      | 51(39.8)          | 7(21.9)           |         | 12(36.4)          |         |

Note: Except where indicated, data are numbers of patients, with percentages in parentheses.

*Data are expressed as mean ± standard deviation.

†Data are medians, with interquartile range in parentheses.

P < 0.05 indicates there are significant differences in clinicopathological features of patients in the training set, test set and validation set.

4. Discussion

As a branch of radiomics, ultrasomics has been studied in several fields. However, there has been no multicenter report on the use of ultrasomics features to predict pathological grading of HCC. In the present study, ultrasound features were extracted from the sketched ROI in HCC lesions. Feature preprocessing and selection were performed by using normalization, variance filtering, and lasso regression. SVM was combined with the learning curve and grid search for parameter tuning to train the models. The results showed that grayscale ultrasound features could be effectively used to differentiate between high- and low-grade HCC on the training set, test set, and the independent validation set (p < 0.05). Our study indicated that ultrasomics offers a pathway for preoperative prediction of HCC pathological grade by mining the information of tumor heterogeneity that cannot be recognized by the naked eye in grayscale ultrasound images. It can be seen from Table 2, on the training set, test set and independent validation set, the combined model outperformed the ultrasomics model, while the ultrasomics model outperformed the clinical model (p < 0.05). These findings were consistent with the previous research using CT- and MRI-based radiomics [26,27,35,36]. This indicates that the combined model combining ultrasomics signatures and clinical factors is more robust.

It has been shown that radiomics based on multimodal imaging techniques can preoperatively predict the pathological grading of HCC to varying degrees. Mao et al. were concerned with the application of the machine learning algorithm XGBoost combined with contrast-enhanced CT (CECT)-based radiomics signatures in preoperative prediction of pathological grading of HCC [26]. AUC of the combined model was 0.801 on the test set. Wu et al. analyzed the value of MRI-based radiomics features in preoperative prediction of pathological grading of HCC [27]. On the test set, the AUC of the radiomics score established by combining radiomics signatures and clinical features with lasso regression was 0.800. In our study, on both the test set and the independent validation set, AUC of the combined model was 0.874 and 0.849, respectively. This result indicated that the ultrasomics-based approach was comparable to or even outperformed the CT- or MRI-based radiomics. More importantly, ultrasound has the advantages of low cost, ease of operation, fast result interpretation, and absence of radiation, which is hardly paralleled by other radiographic techniques. The ultrasomics-based approach seems to be a promising alternative.

However, our study had certain limitations. Firstly, we only divided the HCC patients into two types, high- and low-grade. In the future, the sample size will be expanded to further study the value of ultrasomics in multicategory classification. Secondly, most of the enrolled patients had a history of viral hepatitis. The reproducibility of our approach is needed to verify in other heterogeneous liver diseases. Thirdly, the features were extracted from ROI instead of the entire tumor. Therefore, more sections need to be covered in future studies. Fourthly, we only used the grayscale ultrasound images. Multimodal ultrasound images will be included in future studies. Fifthly, Ultrasound images come from many different devices. The difference of ultrasound parameters between patients may have potential influence on the experimental results.

Taken together, machine learning-based ultrasomics offers a great promise in preoperative prediction of pathological grading of HCC. The combined model integrating ultrasomics signatures and clinical factors had a better predictive performance for both high- and low-grade HCC compared with the single clinical model and the ultrasomics model. The combined model also had a stronger generalization ability.
Table 2
Performance of training set, test set and validation set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
<th>AUC (95 %CI)</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>Clinical</td>
<td>64.71</td>
<td>80.52</td>
<td>74.22</td>
<td>0.789(0.707–0.856)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>Ultrasomics</td>
<td>78.43</td>
<td>93.51</td>
<td>87.50</td>
<td>0.878(0.808–0.929)</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>90.20</td>
<td>93.51</td>
<td>92.19</td>
<td>0.977(0.934–0.995)</td>
<td>0.0013</td>
</tr>
<tr>
<td>Test set</td>
<td>Clinical</td>
<td>57.14</td>
<td>72.00</td>
<td>68.75</td>
<td>0.720(0.534–0.863)</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>Ultrasomics</td>
<td>71.43</td>
<td>84.00</td>
<td>81.25</td>
<td>0.789(0.608–0.912)</td>
<td>0.0100</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>85.71</td>
<td>84.00</td>
<td>84.38</td>
<td>0.874(0.795–0.964)</td>
<td>0.0029</td>
</tr>
<tr>
<td>Validation set</td>
<td>Clinical</td>
<td>75.00</td>
<td>61.90</td>
<td>66.67</td>
<td>0.770(0.591–0.988)</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>Ultrasomics</td>
<td>75.00</td>
<td>76.19</td>
<td>75.76</td>
<td>0.825(0.654–0.935)</td>
<td>0.0025</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>75.00</td>
<td>85.71</td>
<td>81.82</td>
<td>0.849(0.682–0.949)</td>
<td>0.0099</td>
</tr>
</tbody>
</table>

p value < 0.05 indicates a significant difference in the discrimination of low-grade HCC and high-grade HCC

Fig. 5. The ROC curves of the modes in the training set, test set and validation set. A is the ROC curve of the clinical model based on clinical factors. B is the ROC curve of the ultrasomics model based on ultrasomics signatures. C is the ROC curve of the combined model based on clinical factors and ultrasomics signatures.

CRediT authorship contribution statement

Shanshan Ren: Conceptualization, Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Data curation, Supervision. Qinghua Qi: Resources, Data curation. Shunhua Liu: Software, Investigation, Visualization. Shaobo Duan: Resources, Data curation. Bing Mao: Resources, Data curation. Zhiyang Chang: Resources, Data curation. Ye Zhang: Resources, Data curation. Shuaiyang Wang: Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejrad.2021.109891.

References


