nature portfolio

Peer Review File

Focal Liver Lesion Diagnosis with Deep Learning and Multistage CT Imaging

Editorial Note: Figures on page 16 of this Peer Review File have been redacted as indicated to remove third-party material where no permission to publish could be obtained.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author): Expert in liver cancer clinical research and pathology, digital pathology, and artificial intelligence

The study built a deep-learning method to identify benign and malignant liver tumors. They performed external validation in a large population, demonstrated the high performance of their algorithm. The system could potentially be implemented in clinical practice for the diagnosis of liver lesions. I have the following suggestions.

1.The "Deep learning analysis" results provided class activation maps for each image. In order to improve the trust of clinical users, it is suggested to enhance the interpretability of the model by analyzing the class activation maps by expert doctors or other methods.
2.Pathological diagnosis is considered the gold standard for tumor diagnosis. In cases where AI systems and doctors disagree on the diagnosis, it is important to verify whether there are corresponding pathological diagnosis results available for these tumor cases. It is suggested to compare inconsistent results with the pathological diagnosis to determine whether the AI judgment is a true positive.
3.Combined hepatocellular-cholangiocarcinoma is a distinct type of malignant liver tumor different from hepatocellular carcinoma, cholangiocarcinoma, and metastatic carcinoma. How are these cases typically managed in this research?
4.The language of the article needs further refinement. For example, the content from "lines 233 to 234" , "lines 323 to 325" is difficult to understand.
5.The third and fourth paragraphs in the “discussion” is repetitive with the “results”. It is suggested to condense or delete them.

Reviewer #2 (Remarks to the Author): Expert in cancer digital pathology, artificial intelligence, and deep learning

It is indeed a shortcoming that previous studies have only looked at HCC detection and not at the multiclass problem. This study thereby addresses a crucial topic and it uses a large and very nice dataset.

Also, the authors should be commended for having used CNNs with self-supervised learning and not hand-crafted radiomics, because the latter are outdated.

The performance is generally very good but still probably too bad for immediate clinical use. How could this be improved?

The multi-reader study is very good and the external validation is very good.

However, there are some concerns which should be fixed in a revision.

Major
1. Code availability on GitHub under an open access license and with adequate documentation
2. Data availability — the data must be provided in an anonymized way.
3. Availability of the resulting models under open access licenses should be clearly pointed out.
4. Figure 4: How was the threshold selected? To go from a ROC curve to a confusion matrix, you need a threshold. This threshold should not be determined on the test set, but on the training set or a validation set.
5. Throughout the article, point out how the training set and test set were split to make clear that there was no contamination of the training set with test set samples whatsoever.
6. The website http://66.135.22.51/ is not accessible for me.

Minor
1. Typo line 109: It should be "in the radiomics method," not "in the radiology method."
2. Line 206: What do the authors mean by "images"? Do they mean slices or series or examinations? Be more precise in this throughout the article, please.
3. English language could be improved.
4. The figure quality could be improved; color maps are missing, figures should be concatenated into multi-panel figures.
5. Figure 1 comes after Figure 5; please renumber.
6. Please provide analyses broken down by sex/gender as mandated by the Nature reporting guidelines.
7. Please declare adherence to the STARD guidelines.

Reviewer #3 (Remarks to the Author): Expert in liver cancer clinical research and pathology, and digital pathology

In this paper, a model called LiLNet is constructed using ResNet50 as the backbone network for CT images of liver diseases. The model has achieved better accuracy than radiologists in selecting target areas by object detection, determining malignant/benign tumors by LiLNet_BM, types of malignant tumors by LiLnet_M, and determining benign tumors by LiLNet_B. 1580 cases were used for training, 1151 cases for validation, 1308 cases for testing, and 221 cases were used separately for comparison with Radiologist. Although the study is based on a large number of cases, there are several critical issues to be addressed.

Comment
1. There is no description of the final pathological diagnosis of the case. Especially for surgical cases, the accuracy of the pathological diagnosis should be described. For malignant tumors without pathology samples as a result of TACE or radiofrequency treatment, the accuracy of the system should be described for the results of the final definitive diagnosis using angiography or tumor markers.
2. Table 1 shows the comparison with the Radiologist's results, and Figure 4 shows the comparison with the clinical doctor. The clinical doctor should not make a judgment based on CT images alone. In the

absence of surgery or needle biopsy, the clinical doctor's diagnosis should be the final diagnosis, but in Figure 4, there are several "wrong" cases, and it is not stated what these mistakes were based on. Also, there are many CT images (multi-phase) for each case, but there is no description of how the selection of which image to use is made. Also, there is no description of the conditions under which the CT images were taken.

3. There is no description of the accuracy of automatic detection of the target area. The accuracy of this automatic detection should be described.

4. In liver disease, the size of the target area is also an important factor. For example, it is necessary to indicate the degree of accuracy for each size, e.g., 1 cm or less 1-3 cm 3 cm-5 cm, etc.

5. It should be noted how the background liver condition (cirrhosis, fibrosis, or inflammation) may have made a difference in lesion extraction or in the change in AI detection on CT.

6. In multi-phase contrast-enhanced CT images, it is necessary to describe how the results changed in phases such as arterial phase, venous phase, and equilibrium phase.

7. If the results are different in the same case with different Phase images, it is necessary to describe how the results of the case were determined.

8. There seems to be an improvement in accuracy especially in Benign compared to the original ResNet50 accuracy, please discuss why.

Responds to the reviewer's comments:

Reviewer #1 (Remarks to the Author): Expert in liver cancer clinical research and pathology, digital pathology, and artificial intelligence

The study built a deep-learning method to identify benign and malignant liver tumors. They performed external validation in a large population, demonstrated the high performance of their algorithm. The system could potentially be implemented in clinical practice for the diagnosis of liver lesions. I have the following suggestions.

Author Response: We greatly appreciate your recognition and valuable suggestions for our research. Your suggestions has offered valuable insights, contributing to further refinement of our study. We have taken your suggestions into serious consideration and have supplemented relevant experiments accordingly.

1.The "Deep learning analysis" results provided class activation maps for each image. In order to improve the trust of clinical users, it is suggested to enhance the interpretability of the model by analyzing the class activation maps by expert doctors or other methods.

Author Response: Thank you for the valuable suggestion to enhance the interpretability of our "Deep learning analysis" results. To better explain the deep learning model, we conducted two experiments, including analysis by professional radiologists on activation maps and gradient analysis. Specifically:

Gradient Statistical Analysis: Model interpretability refers to the process of explaining the outputs generated by a machine learning model, elucidating which features and how they influence the actual output of the model. In the realm of deep learning, particularly in computer vision classification tasks, where features are essentially pixels, model interpretability aids in identifying pixels that have either positive or negative impacts on predicting categories. To achieve this, we employ the Gradient Shapely Additive exPlanations (SHAP) library for interpreting deep learning models. This process primarily involves analyzing the gradients within the model to

gain a deeper understanding of how decisions are made. By inspecting gradients, we can determine which features contribute most significantly to the model's predictions. In Figure 1, we present plots for HCC, ICC, MET, FNH, HEM, and CYST. Each SHAP plot comprises the original image alongside grayscale images corresponding to the number of output classes predicted by the model. Each grayscale image represents the model's contribution to the output class. In these images, blue pixels indicate a negative effect, while red pixels indicate a positive effect. Conversely, white pixels denote areas where the model ignores input features. Below the images, there is a color scale ranging from negative to positive, illustrating the intensity of SHAP values assigned to each relevant pixel. For instance, in the case of correct HCC category prediction, the SHAP plot for HCC reveals that red activations are predominantly concentrated in the lesion area. However, in SHAP plots for other categories such as ICC and MET, although some red pixels are present, they are not concentrated in the lesion area. This suggests that the appearance of red activations outside the lesion area in other categories may indicate a misjudgment or confusion by the model during prediction. Meanwhile, the activation in the lesion area remains one of the key factors for accurate prediction.

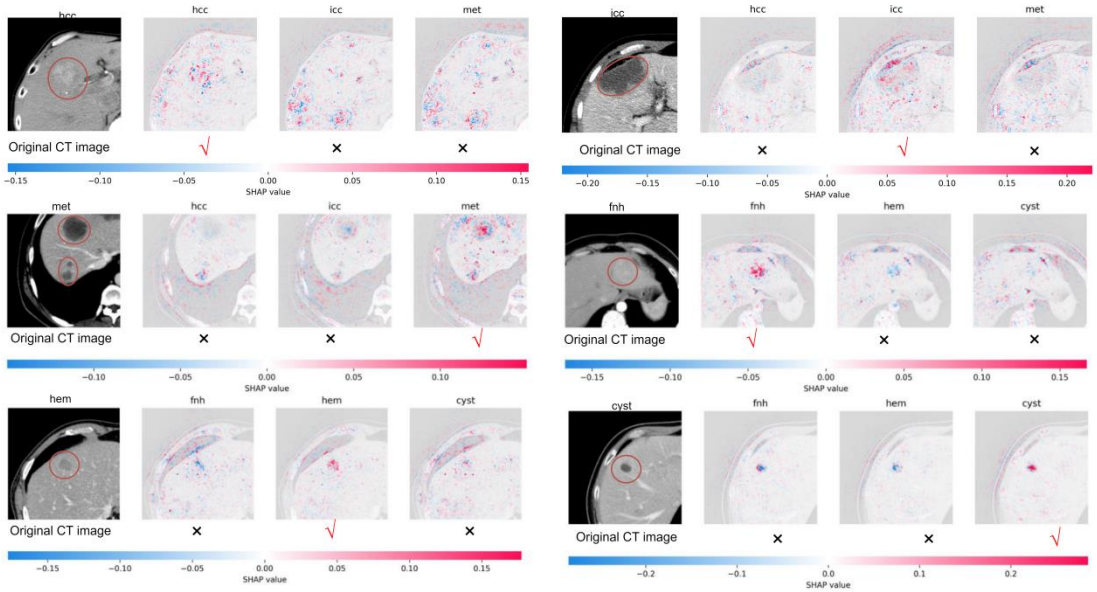


Figure 1: SHAP Plots Revealing Pixel Influence on Model Predictions for HCC, ICC, MET, FNH, HEM, and CYST.

Analysis by Professional Radiologists on Activation Maps: CAMs are generated by computing the activation level of each pixel in the image by the model, revealing the areas of focus within the image. From CAM images, it can be observed that the model pays more attention to lesion areas relative to normal liver tissue to distinguish between different subtypes. We have invited a radiology expert to review these activation maps. Their clinical expertise may offer valuable insights into the relevance and accuracy of highlighted regions.

HCC typically exhibits heterogeneity in internal structure and cellular composition, resulting in significant variation within the tumor. Rapid proliferation of tumor cells leads to increased cell density and richer vascularity in the central region, often manifested as arterial phase enhancement on imaging. Conversely, the surrounding area may display lower density and vascularity due to compression by normal hepatic tissue or arrangement of tumor cells in a nest-like pattern, presenting as low density on imaging. Consequently, in this CAM image, the central region may exhibit deep activation, while the surrounding area may show secondary activation. Additionally, the irregular spiculated margins commonly observed in HCC are a critical feature, often encompassed within the activated regions.

ICC is characterized by tumor cells primarily distributed in the peripheral regions, with fewer tumor cells and immune-related lymphocytes in the central area. Imaging typically reveals higher density and vascularity in the tumor periphery, contrasting with lower density and vascularity in the central region. These imaging features are reflected in the CAM image.

Metastatic tumors, arising from either intrahepatic primary tumors or extrahepatic malignancies, often exhibit necrosis and uneven vascularity in tissue composition. This results in the characteristic imaging appearance of indistinct margins and multifocal lesions. CAM images frequently depict this by demonstrating areas of diffuse and poorly defined activation, with uneven depth and distribution of activation regions.

FNH typically arises from abnormal arrangement of normal hepatic cells and contains abundant vascular tissue with high density. On imaging, it typically presents as

homogeneous enhancement of focal lesions, while surrounding normal hepatic tissue appears relatively hypoenhanced due to compression. In CAM images, the lesion often exhibits uniform overall activation, while the compressed normal hepatic parenchyma demonstrates relatively lower activation. FNH is characterized by richer vascularity compared to other lesions, resulting in higher overall activation.

Hem usually contain abundant vascular tissue and manifest as focal lesions with significant enhancement during the contrast-enhanced phase on imaging. In CAM images, they typically appear as locally activated areas, exhibiting higher activation compared to other non-vascular lesions, with a more uniform distribution.

CYST typically consist of fluid or semi-solid material, with uniform internal tissue distribution and clear borders. On imaging, they appear as circular or oval-shaped low-density areas with clear borders. In CAM images, cystic regions display as circular areas with deep activation, and the activation intensity within the cyst is usually uniform, without significant differences.

CAM images reveal the areas of focus and activation patterns of the model in identifying tumors, aiding doctors in gaining a deeper understanding of the model's decision-making process.

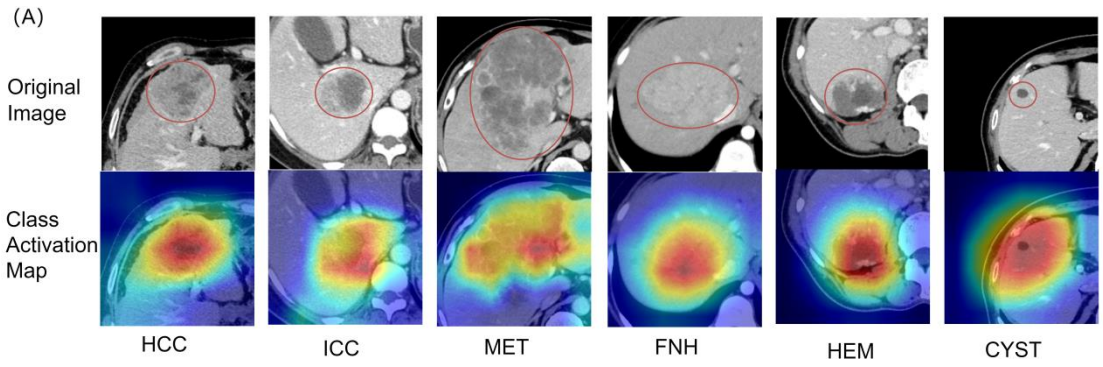


Figure 2: Visualization of the class activation map generated by the last convolution layer. We presented activation maps for liver lesions. The first line displays the original image, while the second line displays the corresponding activation map. Red denotes higher attention values, the color blue denotes lower values, and the red circle represents the tumor area.

Corresponding modifications in the paper:

(1) Lines 413-480

(2) Figure: 5

2.Pathological diagnosis is considered the gold standard for tumor diagnosis. In cases where AI systems and doctors disagree on the diagnosis, it is important to verify whether there are corresponding pathological diagnosis results available for these tumor cases. It is suggested to compare inconsistent results with the pathological diagnosis to determine whether the AI judgment is a true positive.

Author Response: The pathological diagnosis is indeed regarded as the gold standard for tumor diagnosis, crucially validating the performance of artificial intelligence systems. All data rely on labels obtained from this gold standard as the ground truth. Consequently, all results, including any inconsistencies, were compared with pathological diagnoses to assess the performance of the AI judgments

For example, in *Compared with Radiologists* section, the benchmark for comparing the results of AI and radiologists is the true labels. The confusion matrix in Figure 4 compares the AI and radiologists' predictions with the true labels. In the "Benign" of Figure 4, AI and radiologists agree on 92 cases. Among these, 87 cases are confirmed correct by pathology, while 5 cases are incorrect. Additionally, there are 12 cases of disagreement: 5 are incorrect AI judgments (false negative), 7 are correct (true positive), and 7 are incorrect radiologist diagnoses (false negative), with 5 being correct (true positive). Consequently, when AI and radiologists differ, AI achieves a 58.34% true positive rate for benign diagnoses, compared to the radiologists' 41.67%. In the "Malignant", AI and radiologists agree on 96 cases. Among these, 95 cases are confirmed correct by pathology, while 1 cases are incorrect. Additionally, there are 21 cases of disagreement: 9 are incorrect AI judgments (false negative), 12 are correct (true positive), and 12 are incorrect radiologist diagnoses (false negative), with 9 being correct (true positive). Consequently, when AI and radiologists differ, AI achieves a 57.14% true positive rate for benign diagnoses, compared to the radiologists' 42.85%. Furthermore, additional diagnostic information for HCC, ICC, MET, FNH, HEM, and CYST can be found in Figure 4.

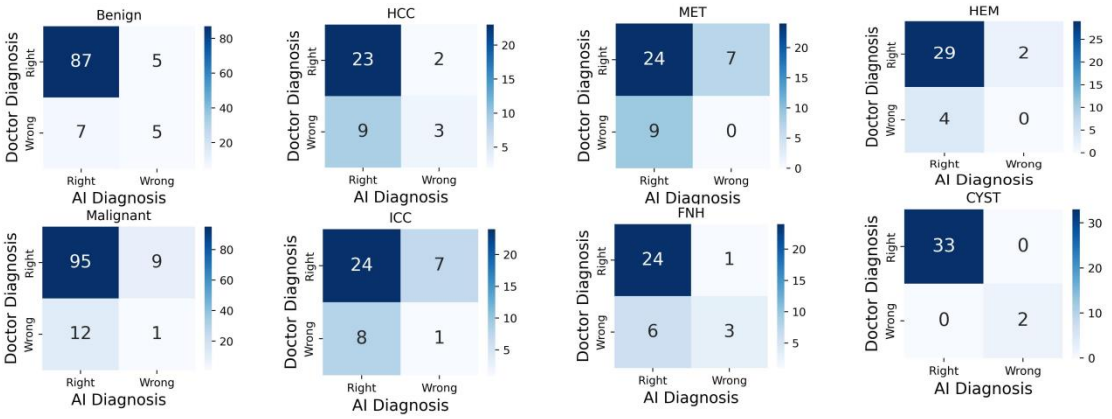


Figure 4: The confusion matrix is employed to depict the classification of lesions for patients, categorizing them into four groups based on the diagnoses provided by AI systems and radiologists. 'Radiologist Right' and 'AI Right' indicate instances where both the AI system and the doctor correctly diagnosed liver tumors. 'Radiologist Right' and 'AI Wrong'refer to cases where the AI system incorrectly diagnosed a liver tumor but the radiologist's diagnosis was accurate. 'Radiologist Wrong' and 'AI Right' pertain to situations in which the AI system made a correct diagnosis of liver tumors but the radiologist's diagnosis was incorrect. 'Radiologist Wrong' and 'AI Wrong' represent instances where neither the AI system nor the doctor diagnosed liver tumors correctly.

Corresponding modifications in the paper:

(1) Lines 317-348

(2) Figure: 4e

3.Combined hepatocellular-cholangiocarcinoma is a distinct type of malignant liver tumor different from hepatocellular carcinoma, cholangiocarcinoma, and metastatic carcinoma. How are these cases typically managed in this research?

Author Response: We appreciate the constructive feedback from the reviewers. We apologize for not including combined hepatocellular-cholangiocarcinoma (cHCC-CCA) in our study. cHCC-CCA is a rare variant of liver cancer, with an incidence rate ranging from 0.4% to 14.2% compared to other primary liver cancers[1,2,3]. This posed significant challenges to our data collection, with relatively

few samples available and difficulties in data collection and annotation. Therefore, without a sufficient number of cHCC-CCA samples in our research dataset, it was not feasible to conduct relevant analysis and evaluation. We regret not including cHCC-CCA in our study due to constraints in resources and time. We had to focus our efforts on clinically common liver lesions. Nevertheless, we recognize the significance of cHCC-CCA in the field of hepatocellular carcinoma and plan to prioritize it as a key area for future research. We hope to collaborate with expert pathologists to collect relevant data and incorporate cHCC-CCA into our future studies, thereby enhancing the scope of our research findings.

[1] Beaufrère, Aurélie, Julien Calderaro, and Valérie Paradis. "Combined hepatocellular-cholangiocarcinoma: An update." Journal of hepatology 74.5 (2021): 1212-1224. https://doi.org/10.1016/j.jhep.2021.01.035

[2] D. Ramai, A. Ofosu, J.K. Lai, M. Reddy, D.G. Adler Combined hepatocellular cholangiocarcinoma: a population-based retrospective study Am J Gastroenterol, 114 (2019), pp. 1496-1501, 10.14309/ajg.0000000000000326

[3] Calderaro, J., Ghaffari Laleh, N., Zeng, Q. et al. Deep learning-based phenotyping reclassifies combined hepatocellular-cholangiocarcinoma. Nat Commun 14, 8290 (2023). https://doi.org/10.1038/s41467-023-43749-3

Corresponding modifications in the paper:

(1) Lines 564-574

4.The language of the article needs further refinement. For example, the content from "lines 233 to 234"，"lines 323 to 325" is difficult to understand.

Author Response: We apologize for any confusion our sentence may have caused, and we have made some adjustments to clarify it.

- lines 233 to 234: We Change “It's worth noting that there is a high level of agreement in the diagnosis of MET，HEM and CYST” to “Based on Figure 4, it's evident that AI and radiologists attained congruent outcomes in cyst diagnosis,

accurately identifying 32 cases while misdiagnosing 2 cases. Upon analyzing the misdiagnosed CYST images, we discovered one case with a mixed lesion, showing characteristics of both HEM and CYST. In this instance, the CYST was positioned near a blood vessel, resulting in a misdiagnosis as HEM. Another misdiagnosis was due to the lesion's size being less than 1cm, presenting challenges in identification. However, there are some differences in diagnosis between AI and radiologists in other categories, indicating differences in diagnostic approach or focus. These findings highlight the potential for our AI-assisted software to collaborate with radiologists in enhancing the diagnostic accuracy of liver lesions."

Corresponding modifications in the paper:

(1) Lines: 339-347

- "lines 323 to 325"："Secondly, some centers had limited data with only HCC and ICC samples. LiLNet can distinguish between benign, malignant, and benign lesions more easily than between HCC and ICC. While there may be some misclassifications among malignancies, this doesn't notably impact the accurate prediction of benign lesions" mean "HCC and ICC together account for the majority of liver cancer cases, representing approximately 70-90% of primary liver malignancies globally. And these two types present more significant challenges for classification compared to other types of liver focal lesions. Therefore, we collected additional HCC and ICC cases from three validation centers for further validation. However, the limited number of benign samples may have impacted the model's comprehensiveness and its ability to generalize in benign diagnosis to some extent. To address this, we intend to incorporate more data in our future efforts to enhance the model's generalization performance." However, we conducted validation in two clinical centers, and the performance of benign data did not decrease. Therefore, we have modified our limitations in this paper to "cHCC-CCA refers to combined hepatocellular-cholangiocarcinoma, a rare type of liver tumor that exhibits both hepatocellular carcinoma and cholangiocarcinoma characteristics. cHCC-CCA is a rare variant of liver cancer,

with an incidence rate ranging from 0.4% to 14.2% compared to other primary liver cancers28,29,30. The performance of the LiLNet system in diagnosing cHCC-CCA is currently unclear due to the rarity of this tumor and limited research and data availability. Recognizing the importance of cHCC-CCA in the field of hepatocellular carcinoma, we plan to focus on this topic as a key area of future research. We aim to collaborate with pathology experts to collect relevant data and incorporate cHCC-CCA into our future studies, thereby expanding the scope of our research findings."

Corresponding modifications in the paper:

(1) Lines: 562-572

5.The third and fourth paragraphs in the "discussion" is repetitive with the "results". It is suggested to condense or delete them.

Author Response: We appreciate the reviewer's feedback regarding the redundancy in the discussion section. After carefully revising the manuscript, we have condensed the third and fourth paragraphs in the discussion to avoid repetition with the results section. Instead, we focused on providing a more in-depth analysis and interpretation of the methods used and the significance of the results obtained. We believe these revisions have enhanced the clarity and coherence of the discussion, and we thank the reviewer for bringing this to our attention.

The revised dicussion is "Our model exhibits robust performance in both the test set and external validation, primarily owing to the integration of extensive datasets and advanced AI technology. The training data is comprehensive, encompassing a wide array of patterns, which include diverse imaging devices, variations in image window widths and levels, and adjustments in target area sizes. These factors are meticulously considered to accommodate differing background liver conditions, such as cirrhosis, fibrosis, inflammation, fatty liver, and abdominal fluid. Our model has shown excellent performance on test and external validation sets, primarily owing to the integration of big data and AI technology. Our training dataset is extensive and diverse, comprising images acquired from a variety of CT device models, each with

unique specifications for window width and level settings. Additionally, the dataset includes samples representing a wide range of background liver conditions, including cirrhosis, fibrosis, inflammation, fatty liver, and the presence of abdominal fluid. Moreover, it encompasses varying sizes of target areas for comprehensive coverage. The richness of data in our training set significantly contributes to enhancing the model's generalization. We adopt a two-stage approach, starting with detection-then-recognition technology. Initially, through object detection methodologies, we extract ROI to minimize irrelevant background information and direct the model's attention to the tumor. Subsequently, we segment the liver tumor classification task into benign and malignant stages,then proceed with subtype classification. This strategic classification approach not only reduces the complexity and difficulty of subsequent classifiers but also enhances the overall accuracy and stability of our classification system. Compared to the popular deep learning classification algorithm ResNet50 and pre-trained ResNet50, our proposed model demonstrates better performance on both the test set and external validation. This is primarily attributed to several key enhancements. Firstly, our model introduces an enhanced supervised signal, which selectively discards irrelevant regions in the feature maps and expands original labels into joint labels during training. This additional supervision signal enables the model to better comprehend image content and learn more robust feature representations. Given the challenging nature of the liver tumor classification task, characterized by significant confusion or overlap between categories, our approach provides clearer supervisory signals to differentiate categories effectively, thereby reducing confusion and enabling the model to focus on recognizing detailed features. Additionally, leveraging self-distillation technology empowers our model to learn from its own generated responses, further improving its performance. This self-distillation process allows the model to refine its understanding and generalization ability over time, leading to enhanced performance in practical applications.

The LiLNet model outperforms clinical radiologists in third-tier cities due to its utilization of a vast dataset from a reputable comprehensive hospital in China for

training, offering broader coverage and greater sample diversity. The AI algorithms have undergone extensive standardization and optimization, ensuring consistent and accurate diagnosis. Conversely, radiologists in third-tier cities may face challenges such as limited medical resources and variations in personnel quality, hindering the level of professionalism and standardization in diagnosis. While there's high consistency between AI and radiologists in diagnosing straightforward cases like CYST, discrepancies arise in easily confused cases like HCC and ICC, HCC and FNH, MET and HEM, suggesting differing diagnostic methods or focuses. The partnership between AI-assisted software and radiologists holds promise for enhancing the accuracy of liver disease diagnosis.

There are a few limitations. Firstly, compared to other liver diseases, the higher incidence of HCC results in data imbalance, which may slightly affect the diagnostic performance of ICC. Comparative studies with clinical doctors have shown that the most accurate results are achieved when artificial intelligence collaborates with doctors to diagnose HCC and ICC. cHCC-CCA stands for combined hepatocellular-cholangiocarcinoma, a rare type of liver tumor that exhibits both hepatocellular carcinoma and cholangiocarcinoma characteristics. The performance of the LiLNet system in diagnosing cHCC-CCA is currently unclear due to the rarity of this tumor and limited research and data availability. Recognizing the importance of cHCC-CCA in the field of hepatocellular carcinoma, we plan to focus on it as a key area of future research. We aim to collaborate with pathology experts to collect relevant data and incorporate cHCC-CCA into our future studies, thereby expanding the scope of our research findings."

Corresponding modifications in the paper:

(1) Lines: 510-559

Reviewer #2 (Remarks to the Author): Expert in cancer digital pathology, artificial intelligence, and deep learning

It is indeed a shortcoming that previous studies have only looked at HCC detection and not at the multiclass problem. This study thereby addresses a crucial topic and it

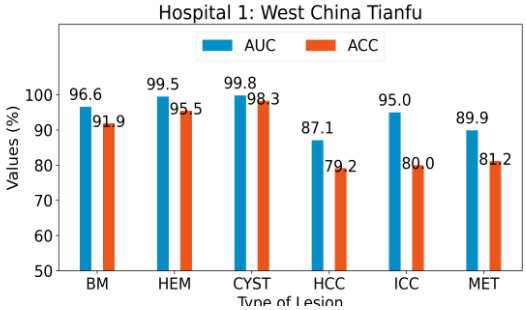
uses a large and very nice dataset. Also, the authors should be commended for having used CNNs with self-supervised learning and not hand-crafted radiomics, because the latter are outdated. The performance is generally very good but still probably too bad for immediate clinical use. How could this be improved?

Author Response: We appreciate your interest and evaluation of our research. We have conducted additional validation at two additional medical centers to ensure the performance and reliability of our model. Our system is presently suitable for routine clinical diagnoses, encompassing outpatient, emergency and inpatient scenarios with patients undergoing AP and PVP sequences. To authenticate the actual clinical efficacy of the system, we seamlessly integrated the system into the established clinical infrastructure and workflow at West China Tianfu Hospital and Sanya People's Hospital in China, as shown in Figure 1.

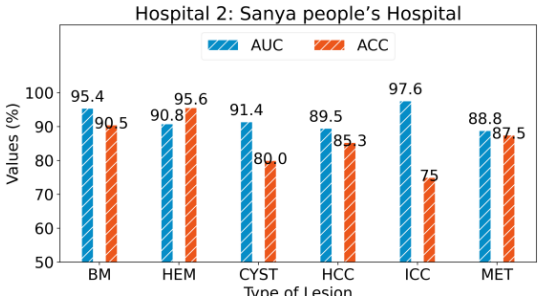
[Figure Redacted]

Hospital 1： At West China Tianfu Hospital, we assessed outpatient and inpatient data from February 29th to March 7th, comprising 117 CYST, 22 HEM, and 16 MET. To improve the evaluation of the model's diagnostic ACC for malignancies, we included 24 HCC and 5 ICC from January 2022 to February 2024. All malignant tumors were

pathologically confirmed, while benign tumors were diagnosed by three senior radiologists. The results of our system at Tianfu center indicated an AUC of 96.6% and ACC of 91.9% for diagnosis of benign and malignant lesions. For HEM, the AUC was 99.54% with an ACC of 95.45%, while CYST showed an AUC of 99.8% and ACC of 98.3%. HCC had an AUC of 87.1% with an ACC of 79.2%, ICC achieved an AUC of 95.0% and ACC of 80%, and MET had an AUC of 89.9% with an ACC of 81.2%.



Hospital 2： We assessed outpatient and inpatient data at Sanya people’s Hospital from March 15th to March 29th, comprising 68 CYST, 23 HEM, 121 Normal , 1 ICC, and 3 MET. Additionally, we retrospectively collected 34 HCC, 3 ICC, and 5 MET from April 2020 to February 2024. All malignant tumors were pathologically confirmed, while benign tumors were diagnosed by three senior radiologists. The results of our system at Sanya center indicated an AUC of 95.4% and ACC of 90.5% for diagnosis of benign and malignant lesions. For HEM, the AUC was 90.8% with an ACC of 95.6%, while CYST showed an AUC of 91.4% and ACC of 80.0%. HCC had an AUC of 89.5% with an ACC of 85.3%, ICC achieved an AUC of 97.6% and ACC of 75%, and MET had an AUC of 88.8% with an ACC of 87.5%.



While our study has made some progress, we recognize that the model's performance may still need further improvement to meet the demands of clinical practice. In the

future, we plan to incorporate additional clinical trial data to further validate the performance and reliability of our model. This data will be sourced from multiple medical centers or diverse regions to ensure the generalizability and applicability of our model. We are committed to actively soliciting feedback from medical experts and clinical practitioners regarding the system, aiming to enhance the comprehensiveness of our research.

Corresponding modifications in the paper:

(1) Lines: 378-479

(2) Figure: 5

1. Code availability on GitHub under an open access license and with adequate documentation

Author Response: Thank you for your suggestion. We have already made the code available on GitHub under an open access license, along with comprehensive documentation to facilitate understanding and usage. This ensures transparency and reproducibility of our research, allowing others to validate our findings and potentially build upon our work. We encourage you to explore the code and documentation, and we welcome any feedback or contributions.

Additionally, we have included more detailed explanations of the code in the Code availability section. The detailed description is as follows: Our system architecture incorporates innovative technologies, including the YOLOv8 model for lesion detection, the 3D UNet segmentation model for post-processing, and our proposed LiLNet classification model. To uphold transparency and reproducibility, we offer access to the source code and models for each component through open-source platforms. The respective repositories are as follows: YOLOv8 can be stored in https://github.com/ultralytics/ultralytics; The website for 3D UNet is https://github.com/ellisdg/3DUnetCNN; The website for ReNet50 is https://github.com/weiaicunzai/pytorch-cifar100/blob/master/models/resnet.py.

In addition, our system's complete custom code can be found in https://github.com/yangmeiyi/Liver/tree/main Obtained from an open-source

repository, including a repository for lesion detection: https://github.com/yangmeiyi/Liver/tree/main/Detection; Our classification model is stored in: https://github.com/yangmeiyi/Liver/tree/main/Classification; We have written detailed operational documentation for each module.

Corresponding modifications in the paper:

(1) Lines:678-691

2. Data availability — the data must be provided in an anonymized way.

Author Response: Thank you for your thoughtful suggestion regarding the anonymization of our research data. Regarding the request for data availability, we fully understand the importance of providing access to data for transparency and reproducibility in research. We have provided sample data on link https://github.com/yangmeiyi/Liver/tree/main/Web%20testing%20data for testing the web version system (http://66.135.22.51/). All analytical data underpinning the findings of this study are incorporated within this paper in the designated Source Data files (Source_data_Figure_2.xlsx to Source_data_Figure_4.xlsx and Source_data_Table_2.xlsx to Source_data_Table_3.xlsx). All data has been anonymized. The original datasets used in this study are subject to access control, as they were obtained under institutional permission and specifically approved by the

institutional review committee. Therefore, they cannot be publicly accessed. However, However, we are open to discussing potential collaborations or agreements that would allow for private academic access to the data under mutually agreeable terms. Please feel free to contact us via email at cjr.songbin@vip.163.com or csmliu@uestc.edu.cn if you would like to explore this further or have any additional questions. We appreciate your understanding and interest in our work. We are more than willing to accommodate any inquiries and provide the necessary data to ensure the integrity and reproducibility of our research.

Corresponding modifications in the paper:

(1) Lines:661-675

3. Availability of the resulting models under open access licenses should be clearly pointed out.

Author Response: We appreciate your interest in the accessibility of the models generated by our research. While we acknowledge the importance of open access licenses in fostering scientific collaboration and progress, determining appropriate licensing strategies for our models involves consideration of various factors, including data privacy concerns. At present, we are unable to provide these models under open access licenses due to the complexities involved in obtaining multi-center ethical permissions and usage agreements. However, we are committed to facilitating academic collaboration and remain open to discussing potential collaborations or agreements that would enable access to these models under suitable conditions. Additionally, we are actively collecting more data to enrich and enhance the system. If you are interested, we warmly welcome you to contact the corresponding author (csmliu@uestc.edu.cn) for further discussion.

4. Figure 4: How was the threshold selected? To go from a ROC curve to a confusion matrix, you need a threshold. This threshold should not be determined on the test set, but on the training set or a validation set.

Author Response: We appreciate the reviewers' thorough evaluation and valuable recommendations. I agree with your point. Typically, to avoid overfitting on the test

set, it's advisable to select the threshold on the training or validation set. In our study, the threshold for classification was not explicitly selected because we used softmax for classification instead of choosing an optimal threshold. This method assigns each class probability scores, and the class with the highest probability is chosen as the predicted class.

(1) For binary classification problems, we employ the default threshold of 0.5, a widely accepted standard. Accordingly, if the model's output probability exceeds 0.5, we predict the sample as positive; conversely, if the probability is less than or equal to 0.5, we predict the sample as negative.

(2) For multi-classification problems, we adopt a straightforward approach where we select the category with the highest probability as the prediction. This method is both intuitive and suited for scenarios where the softmax classifier yields the highest probability as the prediction. As a result, manual threshold selection is unnecessary in multi-classification problems, as we simply rely on the category associated with the maximum probability value in the model's output probability vector.

Corresponding modifications in the paper:

(1)Lines:651-654

5. Throughout the article, point out how the training set and test set were split to make clear that there was no contamination of the training set with test set samples whatsoever.

Author Response: In the manuscript, we meticulously outlined the training and test set partitioning procedure within the *Training strategy* section. This meticulous approach ensures the complete isolation of the training and test sets, thereby eliminating any potential data contamination. Specifically, we employed the following measures to achieve this:

- **Obtain the unique patient identifier:** Each patient is assigned a distinct identification ID, termed as the patient ID. And we use name_ID as the unique

identifier to ensure the uniqueness and consistency of the data.

- **Eliminate duplicate unique identification numbers:** We have systematically excluded redundant samples sharing the same unique identification number to ensure the exclusive presence of each sample in either the training or testing set.
- **Randomly allocate patients:** This step involves the random partitioning of the dataset, guaranteeing that all slices pertaining to a particular patient are allocated to the same dataset. This mitigates the risk of disparate slices from the same patient being erroneously distributed across both the training and testing sets.

Corresponding modifications in the paper:

(1)Lines:631-633

6. The website http://66.135.22.51/ is not accessible for me.

Author Response: Apologies for the inconvenience you experienced in accessing the website http://66.135.22.51/. Our computing nodes are deployed on cloud servers in the United States to ensure easy access for users worldwide. The reasons for the inability to access may include:

- Server network failure: From December 20 to 29, 2023, the website was inaccessible due to server network failure. Although we were unable to change the nodes during the review process, the network has since been restored to normal, and there have been no further issues reported.
- Local network issues for users: Another possibility is network issues in your area, which may affect your access to the website. We suggest checking your network connection or trying to access it using a different network environment.

We have arranged for an additional backup website: http://217.69.1.59. If you encounter continued difficulties accessing the website, we can offer a offline version for you to test .

Minor

1. Typo line 109: It should be "in the radiomics method," not "in the radiology method."

Author Response: Thank you for your valuable feedback. We have addressed this issue and made the appropriate revisions in the revised manuscript.

2. Line 206: What do the authors mean by "images"? Do they mean slices or series or examinations? Be more precise in this throughout the article, please.

Author Response: They are slices. We have revised the term "images" to "slices" throughout the article to ensure accuracy and clarity.

3. English language could be improved.

Author Response: We appreciate your feedback. We will carefully review and revise the text to ensure that it meets the highest standards of readability and accuracy. Additionally, the manuscript has undergone final polishing by Springer Nature to ensure correct English language usage, grammar, punctuation, spelling, and overall style. This polishing certificate was issued on April 26, 2024, and may be verified on the SNAS website using the verification code FF23-19E7-FCCB-6BE5-0FEC.

4. The figure quality could be improved; color maps are missing, figures should be concatenated into multi-panel figures.

Author Response: Thank you for your suggestion. We have already improved the quality of the figures by adding color maps and have concatenated multiple panels into multi-panel figures to enhance clarity and presentation.

5. Figure 1 comes after Figure 5; please renumber.

Author Response: Thank you for bringing this issue to our attention. We have renumbered the figures accordingly.

6. Please provide analyses broken down by sex/gender as mandated by the Nature reporting guidelines.

Author Response: We utilized retrospective data exclusively collected through clinical practice. Gender assignment was based on government-issued IDs. The datasets utilized in the internal training and test cohorts, as well as the external multi-center test cohorts, have reported sex distributions as outlined in the paper (See Table 1 for more details on data distribution and statistics). No sex-based analysis was conducted as gender was unrelated to model implementation or deployment. Patient self-identification of gender was not collected. There were no adverse events in this

study.

Table 1: Baseline characteristics

	Tumor type	Internal Train (n=1580)	Internal Test (n=1308)	Validation HN(n=636)	Validation CD(n=94)	Validation GZ(n=205)	Validation LS(n=216)
Age,years (mean,std)	HCC	53.06±11.90	52.34±12.69	55.56±10.34	59.79±12.20	54.04±11.00	57.5±10.88
	ICC	57.16±12.10	57.27±12.28	59.29±10.38	-	59.18±11.70	59.92±12.14
	MET	55.07±14.37	56.24±13.44	58.61±12.89	-	-	-
	FNH	35.12±13.49	33.65±13.21	35.46±15.22	-	-	-
	HEM	50.13±15.63	47.96±11.06	50.77±10.71	-	-	-
	CYST	58.53±12.93	56.66±12.36	59.13±11.12	-	-	-
Sex (Female/Male)	HCC	155/548	196/750	42/259	20/74	31/142	21/135
	ICC	157/166	43/37	18/26	0/0	23/20	31/18
	MET	70/79	60/73	39/59	0/0	0/0	0/0
	FNH	59/41	19/16	20/19	0/0	0/0	0/0
	HEM	84/48	33/30	63/31	0/0	0/0	0/0
	CYST	77/96	23/28	31/29	0/0	0/0	0/0

Note: Std Standard Deviation,HN Henan Provincial People's Hospital, CD The First Affiliated Hospital of Chengdu Medical College, GZ Guizhou Provincial People's Hospital, LS Leshan People's Hospital

Corresponding modifications in the paper:

(1)Lines:134-154

7. Please declare adherence to the STARD guidelines.

Author Response: Thank you for the reminder. We have added the following statement to the paper: "Reporting of the study adhered to the STARD guidelines."

Corresponding modifications in the paper:

(1)Lines:135

Reviewer #3 (Remarks to the Author): Expert in liver cancer clinical research and pathology, and digital pathology

The study built a deep-learning method to identify benign and malignant liver tumors. They performed external validation in a large population, demonstrated the high performance of their algorithm. The system could potentially be implemented in clinical practice for the diagnosis of liver lesions.

Author Response: Thank you for the valuable feedback. Your suggestions have further strengthened the experimental rigor of the article, rendering the research more comprehensive and reliable. We have deployed and tested clinical applications in two hospitals during the review period, and supplemented relevant experiments in *Real-world clinical evaluation* section of the article.

1. There is no description of the final pathological diagnosis of the case. Especially for surgical cases, the accuracy of the pathological diagnosis should be described. For malignant tumors without pathology samples as a result of TACE or radiofrequency treatment, the accuracy of the system should be described for the results of the final definitive diagnosis using angiography or tumor markers.

Author Response: We appreciate the reviewer's valuable feedback regarding the need for more information on the final pathological diagnosis of the cases included in our study. The patients included in our retrospective study did not have a history of hepatectomy, transarterial chemotherapy (TACE), or radiofrequency ablation (RFA) before undergoing CT examination. Additionally, malignant tumors were pathologically confirmed, while benign tumors were confirmed either by consensus among three radiologists or by follow-up of at least six months using two imaging modalities. We have emphasized this information in the *Patient Characteristics* section of our manuscript.

Corresponding modifications in the paper:

(1)Lines:140-144

2. Table 1 shows the comparison with the Radiologist's results, and Figure 4 shows

the comparison with the clinical doctor. The clinical doctor should not make a judgment based on CT images alone. In the absence of surgery or needle biopsy, the clinical doctor's diagnosis should be the final diagnosis, but in Figure 4, there are several "wrong" cases, and it is not stated what these mistakes were based on. Also, there are many CT images (multi-phase) for each case, but there is no description of how the selection of which image to use is made. Also, there is no description of the conditions under which the CT images were taken.

Author Response: We apologize for the confusion resulting from erroneously referring to radiologists as "clinical doctors" in the article. The comparison depicted in Figure 4 is indeed between the performance of AI systems and radiologists. We sincerely regret this oversight and appreciate your feedback. Corrections will be made in the article to ensure the accuracy and clarity of all content.

- **About "wrong" cases in Figure 4:** The data we collected is retrospective and adheres to the inclusion criteria. Malignant tumors were pathologically confirmed, and Benign tumors were confirmed either by consensus among three radiologists or by follow-up of at least six months using two imaging modalities. The "wrong" cases in Figure 4 is based on true labels (obtained through gold standard).
- **About multi-phase CT images:** we utilize images from both the arterial phase and portal venous phase. In the *Study design and participants* section, we have revised the original statement “multi-phase contrast-enhanced CT series” to “With ethical committee approval, a total of 4039 patients' multi-phase contrast-enhanced CT series (arterial phase and portal venous phase) from six centers in China were included.”
- **CT image acquisition conditions：** As shown in Table 1, CT imaging was performed by using multidetector CT scanners (Revolution, GE Healthcare, Milwaukee, USA; SOMATOM definition, Siemens Healthcare, Erlangen, Germany；Brilliance, Philips Healthcare, Amsterdam, Netherlands; uCT780, United Imaging Healthcare, Shanghai, China). Precontrast images were first obtained before contrast agent (iodine concentration, 300-370 mg/mL; volume,

1.5–2.0 ml/kg of body weight; contrast type, iopromide injection, Bayer Pharma AG) injection. Then, the arterial phase and portal venous phase were obtained with the following parameters: For GE Healthcare, tube voltage, 100-120 kVp; tube current, 450 mA; pitch, 0.992:1; rotation speed: 0.5 s/rot; and ASIR-V: 30%. For Siemens Healthcare, tube voltage, 100-120 kVp; tube current, 210 mA; pitch, 1.0:1; rotation speed: 0.5 s/rot; and Kernel：B30f medium smooth. For United Imaging Healthcare, tube voltage, 100-120 kVp; tube current,150 mA; pitch, 0.987:1; rotation speed: 0.5 s/rot; and Interative reconstruction: KARL 3D. For Philips Healthcare, tube voltage, 100-120 kVp; tube current, 109 mAs; pitch,1.386:1; rotation speed: 0.27 s/rot. The arterial phase and portal venous phase were obtained at 25 s and 60-90 s after contrast injection. The slice thickness for non-contrast images were 5mm, and 1-3 mm for arterial and portal venous phase.

Table 1. CT image acquisition conditions.

Parameters	**GE Healthcare**	**Siemens**	**Philips Healthcare**	**United Imaging**
kV	100-120	100-120	100-120	100-120
mAs	NA	210	109	180
mA	430	NA	NA	NA
Pitch	0.992	1.0	1.386	0.993
Rotation time	0.5 s/rot	0.5 s/rot	0.5 s/rot	0.5 s/rot
Reconstruction of thick slices (CT)	5mm	5mm	5mm	5mm
Reconstruction of thick slices (AP)	1-3mm	1-3mm	1-3mm	1-3mm
Reconstruction of thick slices (PVP)	1-3mm	1-3mm	1-3mm	1-3mm

Corresponding modifications in the paper:

(1)Lines:138

(2)Appendix Table 4

3． There is no description of the accuracy of automatic detection of the target area.

The accuracy of this automatic detection should be described.

Author Response: Thank you for the reviewer's suggestion. We've added object detection results to Figure 2 in the *Performance of LiLNet* section. We filter out bounding boxes with a confidence level above 0.25 and compare them with the actual ground truth boxes. Boxes with an IoU greater than the threshold are true positives, while those with less IoU or repeats are false positives. Undetected boxes are false negatives. As shown in Figure 2, we analyzed F1 score, recall, and precision at different IoU thresholds. At IoU 0.1, we achieved F1 of 94.2%, recall of 95.1%, and precision of 93.3%. At IoU 0.3, F1 was 92.8%, recall 93.7%, and precision 91.3%. IoU 0.5 yielded 87.4% F1, 88.3% recall, and 86.6% precision. These results demonstrate the robust performance of our model across different IOU thresholds. In the LiLNet system, we chose an IOU threshold of 0.1. Despite the minimal overlap between the detection box and the true bounding box at this IoU value, the subsequent classification images are extended to a 224×224 detection box, ensuring coverage of a portion of the lesion.

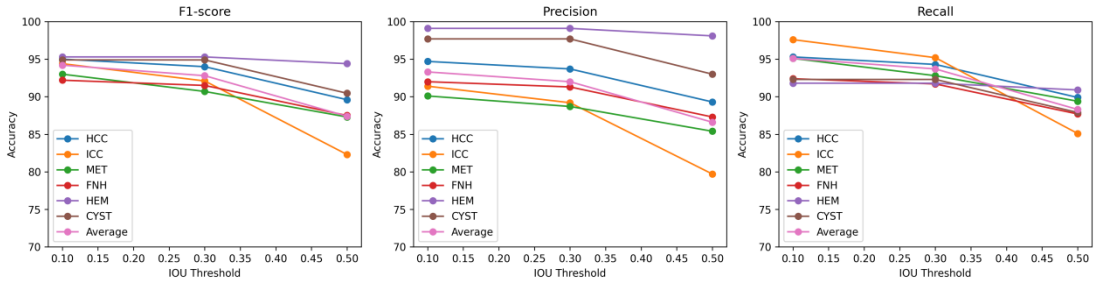


Figure 2: The performance of the proposed model in the testing cohort. This is the outcome of lesion detection under various IOU thresholds.

Corresponding modifications in the paper:

(1) Lines:165-177

(2) Figure: 2a

4．In liver disease, the size of the target area is also an important factor. For example, it is necessary to indicate the degree of accuracy for each size, e.g., 1 cm or less 1-3 cm 3 cm-5 cm, etc.

Author Response: Thank you for your suggestion. In our study, we also recognize the importance of considering the size of the target area in liver disease diagnosis. Therefore, we have supplemented the results obtained from diagnosing cases of different size categories in different centers. Table 2 presents tumor accuracy across different sizes in both the test set and Henan (HN) external validation set. Each cell shows the accuracy percentage for a tumor type within its size range, along with the total sample number. For instance, in the test set, tumors smaller than 1 cm in size achieved 100% accuracy for the HCC type, with a total of 4 samples. However, in the HN validation set, there were no samples in this size range, resulting in a 0% accuracy. Accuracy varies across size ranges, and specific tumor types show differing accuracies within these ranges. Hence, tumor size isn't the sole factor influencing classification accuracy. The result shows varying accuracy levels across tumor sizes, with no consistent trend. Some size ranges display high accuracy, while others show lower accuracy in both the test set and HN validation set. Accuracy also varies for specific tumor types within different size ranges, indicating that tumor size alone doesn't determine classification accuracy. Other factors, such as tumor type, likely contribute to these variations.

Table 2: Accuracy of tumor with differnet size. The numerical values in each cell represent the accuracy of tumor types within the corresponding size range, presented as percentages, and annotated with the total number of samples.

	Size	HCC	ICC	MET	FNH	HEM	CYST	Average
Accuracy of tumor with differnet size on testing sets								
Accuracy /Number	<1 cm	100%/4	none/0	100%/7	0%/1	60.0%/5	100%/20	91.9%/37
	1-3 cm	83.4%/429	92.8%/28	85.8%/106	89.7%/29	89.6%/48	89.7%/29	85.2%/669
	3-5 cm	91.9%/349	76.2%/42	78.9%/19	80.0%/5	90.0%/10	50.0%/2	89.4%/427
	>5 cm	97.5%/164	90.0%/10	100%/1	none/0	none/0	none/0	97.1%/175
Accuracy of tumor with differnet size on HN validation sets								
Accuracy	<1 cm	none/0	none/0	66.6%/4	none/0	83.3%/6	100%/4	85.7%/14

/Number	1-3 cm	69.7%/132	87.5%/8	94.3%/53	73.3%/30	91.3%/46	81.8%/44	79.5%/313
	3-5 cm	89.1%/129	80.7%/26	73.5%/34	62.5%/8	97.3%/37	66.7%/9	85.6%/243
	>5 cm	92.5%/40	60%/10	28.6%/7	100%/1	100%/5	100%/3	81.8%/66

Corresponding modifications in the paper:

(1) Lines: 231-243

(2) Table: 2

5. It should be noted how the background liver condition (cirrhosis, fibrosis, or inflammation) may have made a difference in lesion extraction or in the change in AI detection on CT.

Author Response: Thank you for raising this important point. To assess the potential impact of background liver conditions, such as fibrosis or inflammation, on the performance of our proposed system in analyzing CT images, we have recently collected data from an additional centers like Tianfu, including 3 cases of HCC without hepatitis and liver fibrosis, 21 cases of HCC with hepatitis and liver fibrosis, 5 cases of ICC with similar liver conditions and 16 cases of MET without hepatitis and liver fibrosis. We observed that the system achieved an AUC of 88.1% and an ACC of 80.9% for HCC with liver fibrosis caused by hepatitis, while for ICC, the AUC is 96.4% with an ACC of 80%. Comparing these results to the system's performance at Tianfu Center, where "HCC had an AUC of 87.1% with an ACC of 79.2%, ICC achieved an AUC of 95.0% and ACC of 80%, and MET had an AUC of 89.9% with an ACC of 81.2%," Our experiments have shown that the background liver condition has minimal impact on lesion extraction and imaging examination. This is because our data originates from real clinical events, where liver lesions often coexist with conditions such as cirrhosis, hepatitis, and liver fibrosis. During data collection, we did not exclude background liver diseases. And, the distinct imaging features of liver diseases, such as cirrhosis, fibrosis, or inflammation, on CT images

typically differ from those of liver lesions, making it relatively straightforward for the model to differentiate between them.

Corresponding modifications in the paper:

(2) Lines: 248-262

6. In multi-phase contrast-enhanced CT images, it is necessary to describe how the results changed in phases such as arterial phase, venous phase, and equilibrium phase.

Author Response: In our study, we conducted experiments using multi-phase contrast-enhanced CT images, specifically focusing on the Arterial Phase (AP) and Portal Venous Phase (PVP). The selection of these two phases was based on their significance in tumor and vascular lesion assessment, as well as their widespread clinical use. In clinical practice, lesions exhibit varying characteristics across different phase, each offering different features. Radiologists commonly leverage multiple phases for lesion diagnosis. In accordance with this user habit, our system will detect lesions across multiple phases simultaneously, thereby augmenting support for medical professionals. To assess the benefits of incorporating different phases, we conducted ablation experiments on a dataset comprising 931, 553 patients from both the test set and the external validation set of Henan People's Hospital, encompassing data from multiple phases. The results are depicted in Figures 4A, 4B, 4C and 4D respectively.

As depicted in Figure 4A, for the malignant triple classification in the test set, the diagnostic performance of using both AP and PVP is superior to using either AP or PVP alone, while the results for using AP or PVP alone are comparable. However, in the benign triple classification, the AUC performance is optimal when utilizing both AP and PVP images simultaneously, followed by using AP alone, and finally PVP alone; other performance indicators show that AP outperforms AP and PVP, which outperforms PVP. As illustrated in Figure 4C, in the validation set, the diagnostic performance of AP and PVP surpasses that of using AP or PVP alone, regardless of malignant or benign classification. Through analysis of the confusion matrices of the

test set and external validation set (Figure 4B and Figure 4D), we observed that employing images from both AP and PVP phases simultaneously yields superior results compared to using a single phase. Although the diagnostic outcomes of the two phases align in approximately 90% of cases, there are still instances where lesions exhibit better performance in the AP than in the PVP phase, and vice versa. This discrepancy may be attributed to the inherent characteristics of the data. In summary, integrating information from multiphase CT-enhanced images enables a comprehensive and accurate assessment of liver lesion characteristics and properties, thereby offering a more reliable basis for clinical diagnosis and treatment.

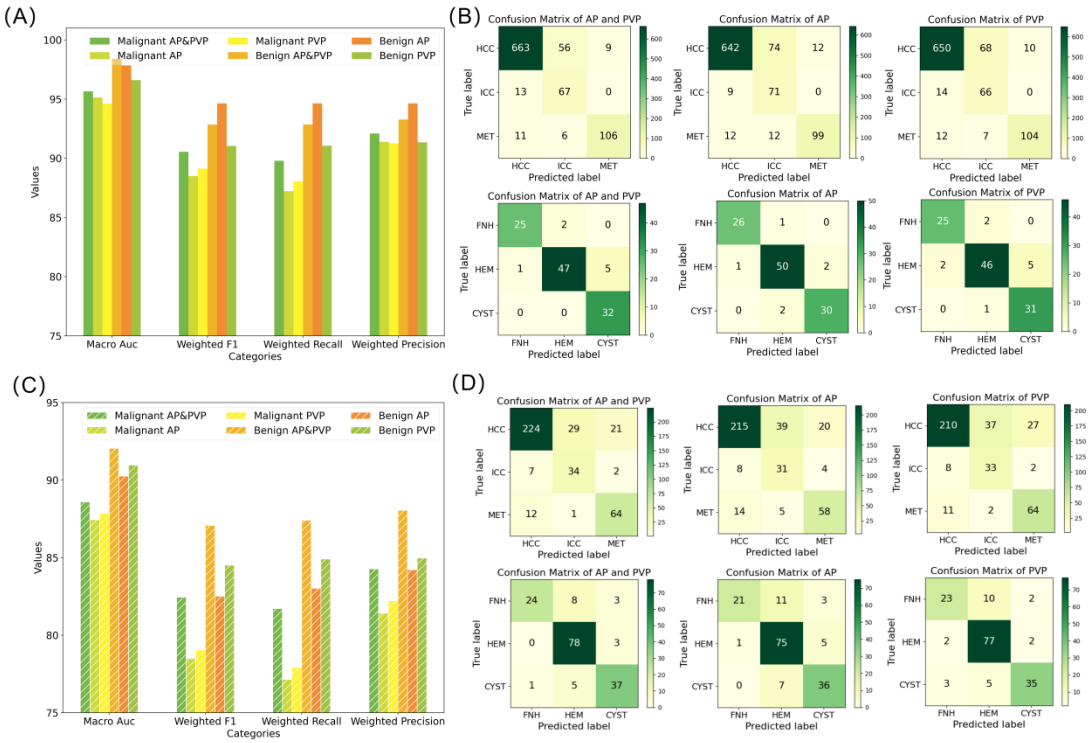


Figure 4: (A) Comparison of AUC, F1, Recall, and Precision for the malignant 3-classification (HCC, ICC, MET) and benign 3-classification (FNH, HEM, CYST) using different phases in the test set. "Malignant AP&PVP" indicates the simultaneous use of AP and PVP for diagnosing malignant lesions, "Malignant AP" indicates the use of only AP, and "Malignant PVP" indicates the use of only PVP for diagnosing malignant lesions. Similarly, "Benign AP&PVP" indicates the simultaneous use of AP and PVP for diagnosing benign lesions, "Benign AP" indicates the use of only AP, and "Benign PVP" indicates the use of only PVP for diagnosing benign lesions. (B) Confusion matrices for the malignant 3-classification and benign 3-classification using different phases in the test set. (C) Comparison of AUC, F1, Recall, and Precision for the malignant

3-classification and benign 3-classification using different phases in the HN external validation. (D) Confusion matrices for the malignant 3-classification and benign 3-classification using different phases in the validation set.

Corresponding modifications in the paper:

(1) Lines: 263-288

(2) Figure:4a, 4b, 4c, 4d

7. If the results are different in the same case with different Phase images, it is necessary to describe how the results of the case were determined.

Author Response: We compute the prediction probability for each image for all individuals, encompassing both AP and PVP images. Subsequently, we calculate the average prediction probability for each individual by averaging the prediction probabilities of all their images. If a patient i has n_i images with corresponding prediction probabilities $p_{i1}, p_{i1}, \ldots, p_{in_i}$, their average prediction probability is obtained by:

$$Average\ \ Probabilit\ \ y_i = \frac{i}{n_i}\sum_{j=1}^{n_i} p_{ij}$$

Then, we apply softmax processing to these average prediction probabilities and determine the category with the highest probability as the final result. The softmax function calculation formula is:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

where *N* is the total number of categories.

When different phase images of the same individual yield disparate results, we address this by averaging the prediction probabilities, thereby assigning higher weight to the image with the most confident prediction. This approach allows us to maximize

the utilization of information from each image, rather than solely relying on the outcomes of a few images. By balancing the influence of each phase image, this method mitigates the impact of abnormal results from one phase image, thereby reducing misjudgments.

Corresponding modifications in the paper:

(1) This method is detailed in the *Method for Calculating Prediction Results* section on page 3 of the appendix.

8． There seems to be an improvement in accuracy especially in Benign compared to the original ResNet50 accuracy, please discuss why.

Author Response: Our proposed method outperformed the model with directly loaded pre-trained parameters, which in turn outperformed the model trained from scratch in our experiment.

- **The reason why the proposed model shows superior performance compared to directly loaded pre-trained models:** The pretext task serves as an effective technique for enhancing the performance of the target task by constructing semantically meaningful image representations. Leveraging the similarity of semantic information between advanced features and images, feature-based pretext tasks offer significant advantages for representation learning. Our model, in comparison to the original ResNet50, introduces an enhanced supervised signal that transforms feature maps by discarding various regions. Subsequently, the original labels are expanded into joint labels to identify the discarded parts during the training process. This additional supervision signal aids the model in better comprehending image content and acquiring more robust feature representations. Given the complexity of the liver tumor classification task, there is considerable confusion or overlap between categories, making it more susceptible to misclassification. By providing clearer supervisory signals, our method assists the model in better discerning the distinctions between categories, reducing the likelihood of confusion, and enabling a sharper focus on recognizing detailed features. Consequently, our approach enhances the model's discriminative ability.

As illustrated in Figure 5, we utilize a t-SNE (t-distributed Stochastic Neighbor Embedding) plot to visualize the feature representations learned by our model. Our proposed model demonstrates superior capability in separating features of different categories and expanding the distance within the feature space.

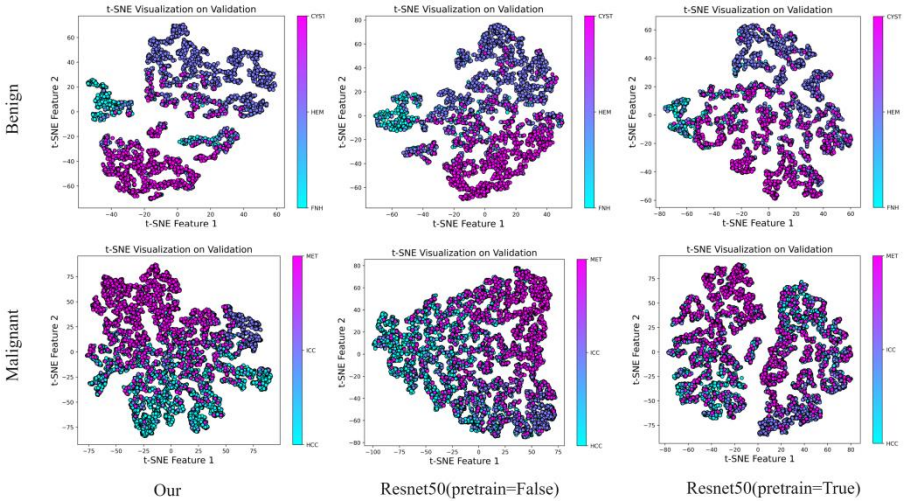


Figure 5: Visual comparison of t-SNE between the proposed model (Our), loaded with pre-trained ResNet50 (Resent50(pretrain=True)) and the standard ResNet50 (Resent50(pretrain=False)) on the Henan validation set.

- The reason why our model performs significantly in benign classification on the external validation set is:

(1) We notice that our proposed method consistently performs well on the test set, showing particularly strong adaptability to benign cases during external validation. The confusion matrix in Figure 6 underscores a 7% increase in ACC compared to transfer learning and an 18% improvement over ResNet50. This significant enhancement is mainly attributed to the utilization of preloaded parameters and supervised information, which provide the model with more user-friendly cues for classification.

(2) Moreover, The effectiveness of benign classification is closely tied to the unique characteristics of benign data. With significantly fewer training samples for benign classification compared to malignant cases, learning from scratch becomes challenging. Transfer learning alleviates this challenge by requiring less target task data for strong generalization. Models loaded with pre-trained parameters thus

outperform those trained from scratch, especially in malignant classification where the abundance of data reduces the relative benefit of pre-training. Pre-trained models leverage existing general feature representations, enhancing their ability to utilize limited target task data effectively.

(3) Comparing our method's performance on the test set and external validation (our column in Figure 6), we observed that our approach prevents the model from excessively focusing on specific categories, which could otherwise lead to subpar performance on other categories. In the case of loading pre-trained models (ResNet50 (pretrain=True) column in Figure 6), we noted challenges in distinguishing FNH and HEM, as well as HEM and CYST. The original ResNet50 tends to prioritize FNH and CYST, which compromises its ability to discern HEM accurately. Since HEM has the poorest performance and there are relatively more HEM cases in the external validation set, this category's performance is notably lower by 7% compared to the test set. Conversely, examining ResNet50 trained from scratch (ResNet50 (pretrain=False) column in Figure 6), we observed a stronger emphasis on FNH and CYST, potentially indicating overfitting and underfitting issues for these categories. While this led to good performance on the test set, it resulted in poor performance on external validation.

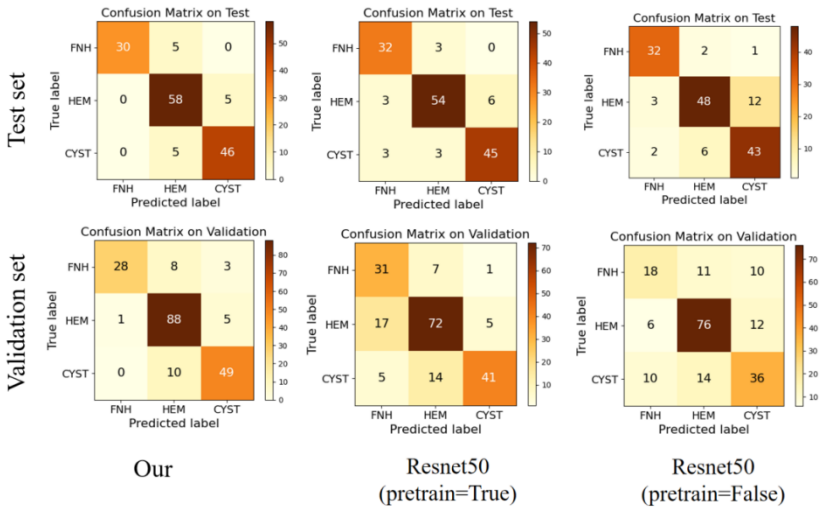


Figure 6: The confusion matrices of t-SNE between the proposed model (Our), loaded with pre-trained ResNet50 (Resent50(pretrain=True)) and the standard ResNet50 (Resent50(pretrain=False)) on the Test and Henan validation sets.

Corresponding modifications in the paper:

(1)Figure: Figure 5 is in *Comparison of methods through t-SNE section* on page 4 of the appendix.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The study built a deep-learning method to identify benign and malignant liver tumors. They performed external validation in a large population, demonstrated the high performance of their algorithm. The system could potentially be implemented in clinical practice for the diagnosis of liver lesions. The manuscript has been revised according to the stipulated requirements and is recommended for acceptance.

Reviewer #2 (Remarks to the Author):

The authors have made progress in addressing the majority of the comments previously provided. However, there are still some critical points that require further attention and clarification before the manuscript can be considered for publication.

Specific Comments

Release of Trained Models:

The authors have addressed many of the initial concerns, but they have not adequately responded to the request for the release of the trained models. The justification provided, citing data privacy concerns, is not fully convincing. It is important to note that this specific request is not for the release of the raw data but for the trained models. Typically, the trained models do not contain individual patient data and thus do not pose a privacy risk. The authors should provide a more detailed explanation of how the trained models could potentially compromise patient data privacy, or ideally, proceed with the release of the models to ensure transparency and reproducibility of their work.

Collaboration and Data Exchange:

The authors mention that they are open to collaborations and can be contacted via email for data exchange. However, considering Chinese regulations, it is generally understood that Chinese data cannot be transferred outside the country unless anonymized. This regulatory aspect needs to be clarified by the authors. If the data cannot be shared internationally, this should be explicitly stated, and an explanation should be provided on how they intend to facilitate collaboration under these constraints.

Training and Test Set Splitting:

A significant methodological issue is the random splitting of the training and test datasets. This practice is not recommended as it can lead to data leakage and overestimation of the model's performance. A more robust approach would be to split the data based on time (e.g., training on data from earlier

periods and testing on data from later periods) or ideally by geographical location (e.g., different centers). This would provide a more realistic evaluation of the model's performance and should be implemented in the study.

Web Interface Accessibility:

The provided web interface is currently not functioning as expected. This issue needs to be resolved to ensure that potential users can access and interact with the tool. The suggestion to invest in a virtual server with a proper domain is practical and should be considered. This relatively small investment could markedly enhance the accessibility and usability of the web interface, thereby supporting wider adoption and validation of the tool by the research community.

The manuscript should be revised to address the above comments before it can be accepted for publication.

Reviewer #3 (Remarks to the Author):

Authors responded to the comments properly and the manuscript appears to be much improved.No further comments from me at this time.

Responds to the reviewer's comments:

Reviewer #1 (Remarks to the Author): Expert in liver cancer clinical research and pathology, digital pathology, and artificial intelligence

The study built a deep-learning method to identify benign and malignant liver tumors. They performed external validation in a large population, demonstrated the high performance of their algorithm. The system could potentially be implemented in clinical practice for the diagnosis of liver lesions. The manuscript has been revised according to the stipulated requirements and is recommended for acceptance.

Author Response: We are extremely grateful for your review and valuable feedback on our manuscript. We are delighted to learn that our study has been recognized and recommended for acceptance. Thank you for the constructive comments and suggestions during the review process, which have greatly helped us improve the quality of our manuscript.

Reviewer #2 (Remarks to the Author): Expert in cancer digital pathology, artificial intelligence, and deep learning

The authors have made progress in addressing the majority of the comments previously provided. However, there are still some critical points that require further attention and clarification before the manuscript can be considered for publication.

Author Response: Thank you for your continued review and feedback on our manuscript. We appreciate your recognition of the progress we have made in addressing the majority of the comments previously provided. We are committed to thoroughly addressing these remaining issues to ensure that our manuscript meets the high standards required for publication.

1. Release of Trained Models: The authors have addressed many of the initial concerns, but they have not adequately responded to the request for the release of the trained models. The justification provided, citing data privacy concerns, is not fully convincing. It is important to note that this specific request is not for the release of the raw data but for the trained models. Typically, the trained models do not contain individual patient data and thus do not pose a

privacy risk. The authors should provide a more detailed explanation of how the trained models could potentially compromise patient data privacy, or ideally, proceed with the release of the models to ensure transparency and reproducibility of their work.

Author Response: Thank you for your feedback and for highlighting the importance of releasing our trained models for transparency and reproducibility. We understand the significance of this request and would like to provide a more detailed explanation of our concerns regarding data privacy.

While trained models typically do not contain direct individual patient data, they can be vulnerable to model inversion attacks. These attacks allow adversaries to approximate the original training data, posing a significant risk for sensitive medical data. Attackers can use intermediate model outputs to recover input medical images through black-box attacks. Additionally, trained model parameters might inadvertently encode specific patterns or outliers from the training data, leading to indirect data leakage. For instance, an attacker might use intermediate outputs from a model to recover an input medical image, leveraging a process known as black-box attack. The adversary does not need to know the model's structure or parameters but can still achieve this by querying the model and observing the outputs. This type of attack has been successfully demonstrated in scenarios where hospitals share their models via APIs for training and inference services[1-3]. Therefore, for the first time, we did not share the trained model parameters.

To balance transparency and privacy, we have removed all identifiable patient information. We have uploaded the model parameters to Google Drive and provided download links and explanations on our GitHub code repository. We have provided parameters for four models, along with specific links as follows:

- Detection model (best.pt): https://drive.google.com/file/d/1y7TzrwmhK6vb1BeXQFty0FqNKXcgVTBU/view?usp=sharing
- Diagnosis model for benign and malignant lesions (BM.pth.tar): https://drive.google.com/file/d/1SclQhkmfspgpZgtqq4WnNqK--DqOt467/view?usp=sharing
- Benign lesion diagnostic model (B.pth.tar):

https://drive.google.com/file/d/1fRdwKXlfEX2h87vW6Mv6C7XcZKRXJnRF/view?usp=sharing

- Malignant lesion diagnostic model (M.pth.tar): https://drive.google.com/file/d/1-ti8Fyugc4djyGpQ4Qs92Ek_CnXoGmfd/view?usp=sharing

However, access requires agreeing to a data usage agreement that stipulates the parameters are for academic research only, prohibits commercial use, and forbids reverse engineering or attempts to reconstruct the original training data.

References:

[1]Mehnaz S, Dibbo SV, De Viti R, Kabir E, Brandenburg BB, Mangard S, Li N, Bertino E, Backes M, De Cristofaro E, Fritz M. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In31st USENIX Security Symposium (USENIX Security 22) 2022 (pp. 4579-4596).

[2]Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., Chen, B., Wang, X. and Xia, S.T., 2024. Privacy Leakage on DNNs: A Survey of Model Inversion Attacks and Defenses. arXiv preprint arXiv:2402.04013.

[3]Wu, M., Zhang, X., Ding, J., Nguyen, H., Yu, R., Pan, M. and Wong, S.T., 2020. Evaluation of inference attack models for deep learning on medical data. arXiv preprint arXiv:2011.00177.

2. Collaboration and Data Exchange: The authors mention that they are open to collaborations and can be contacted via email for data exchange. However, considering Chinese regulations, it is generally understood that Chinese data cannot be transferred outside the country unless anonymized. This regulatory aspect needs to be clarified by the authors. If the data cannot be shared internationally, this should be explicitly stated, and an explanation should be provided on how they intend to facilitate collaboration under these constraints.

Author Response: Thank you for your understanding. As you mentioned, sharing data outside the hospital is indeed restricted. Considering your suggestion, we have made

extensive efforts to communicate with multiple departments, including the Information Technology Department, the Research Office, and the International Cooperation Office. Although these data do not involve blood or other biological samples, the hospital's strict data management policy prohibits the external sharing of any research data. These regulations aim to ensure that all data, regardless of its nature, are securely protected to prevent any unauthorized use or disclosure. However, to facilitate further international cooperation and exchange, it is recommended to jointly apply for an international multi center cooperation project. This would allow us to conduct further research within the hospital while ensuring compliance with hospital policies, data security, and patient privacy.

By jointly applying for international collaboration projects and using secure communication channels, the hospital can engage in research cooperation with foreign institutions while safeguarding patient privacy and data security. This approach not only ensures the lawful and compliant use of data but also promotes scientific research and data sharing across institutions. We welcome interested researchers to contact us via email to jointly apply for collaboration projects.

3. Training and Test Set Splitting: A significant methodological issue is the random splitting of the training and test datasets. This practice is not recommended as it can lead to data leakage and overestimation of the model's performance. A more robust approach would be to split the data based on time (e.g., training on data from earlier periods and testing on data from later periods) or ideally by geographical location (e.g., different centers). This would provide a more realistic evaluation of the model's performance and should be implemented in the study.

Author Response: Thank you for your suggestion. While we believe that random partitioning typically does not result in data leakage, we recognize that dividing data by time or geographic location can indeed better simulate real-world scenarios and provide a more realistic evaluation of model performance. Therefore, based on your suggestion, we will conduct relevant experiments. Here's a detailed explanation:

Random Splitting: Temporal splitting is suitable for time series data or datasets with a temporal relationship. Imaging data of cancer is typically not strictly considered a time series

dataset. These imaging data are usually spatial rather than temporal. Each patient's image represents the liver structure or lesion at a single point in time, rather than continuous observations over a time series. We believe that random splitting of the training and test datasets is appropriate because each patient's data is collected at a single point in time without a direct temporal relationship. Additionally, random splitting does not typically lead to data leakage, as the training and test sets contain imaging data from different patients, representing distinct entities. Therefore, this method effectively evaluates the model's performance on unseen data.

Validation Across Different Centers: We fully agree with the reviewers' suggestion that validating the model's generalizability with data from different centers is essential. We have implemented this in our study. Specifically, our training and test sets come from the same center, while the external validation data comes from different centers (with varied geographical distribution). This approach comprehensively assesses the model's generalizability across different geographical locations and data sources.

Temporal Splitting Experiments: We conducted a time-based data partitioning experiment to further validate the model's generalization ability on the test set. Following the reviewer's suggestion, we sorted the data used for model development chronologically, using early data for training and later data for testing (with the same test set size as random partitioning). We compared the results of random partitioning with those of time-based partitioning, as shown in Figure 6. Using the time-based partitioning method, we achieved an AUC of 98.7% and an ACC of 93.3% for benign and malignant results. The diagnostic AUC for benign data was 98.0%, with an ACC of 91.3%, while for malignant diagnosis, the AUC was 97.5% with an ACC of 90.9%. In external validation, the AUC for benign and malignant diagnosis was 94.6%, with an ACC of 88.3%. The AUC for benign data diagnosis was 88.3%, with an ACC of 82.4%, while for malignant diagnosis, the AUC was 87.9% with an ACC of 85.1%. In external validation for CD, the accuracy of malignant diagnosis was 91.5%. For GZ, the accuracy of malignant diagnosis was 80.1%, and for LS, it was 82.8%.

According to the comparison results in Figure 6, the time-based method consistently outperforms the random partition-based method, showing a 1-2% increase in the AUC on the test set. Performance remains similar on the external validation set. While the time-based

approach slightly decreases ACC for malignancy prediction on the test set and the HN external validation set, it notably enhances prediction accuracy on the test set and multiple external validation sets, notably achieving an 8% increase on the CD validation set. However, performance varies for benign prediction across validation sets, indicating fluctuations in time-based method performance across diverse datasets and tasks. Nonetheless, our time-based approach demonstrates robustness and generalization, adaptable to diverse data features and task requirements.

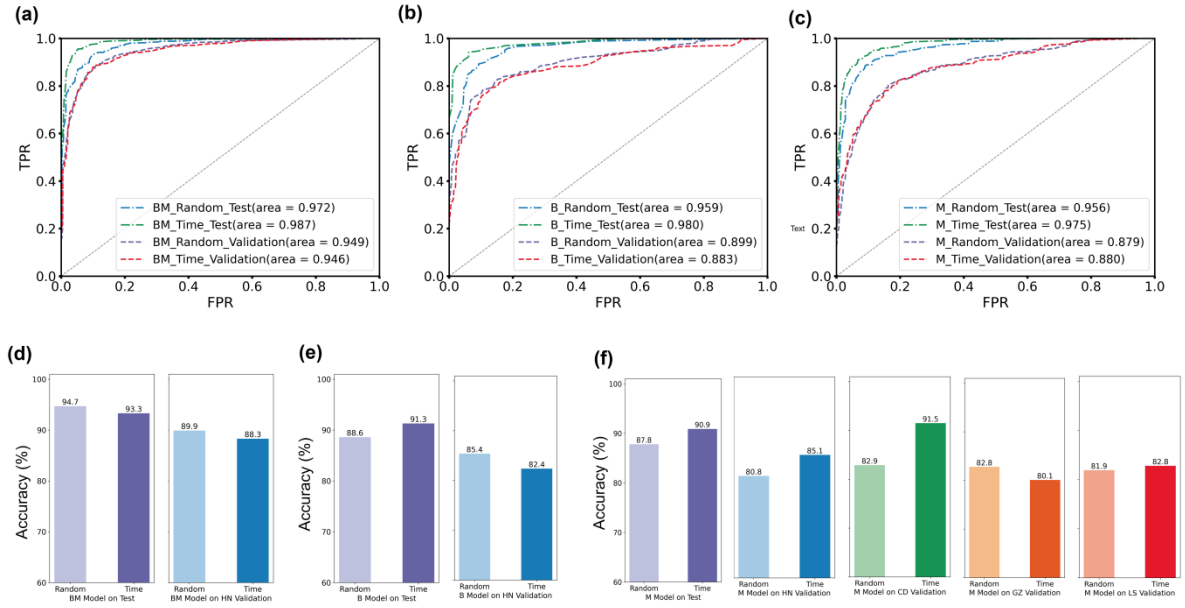


Figure 6: **Comparison of Results between randomly and time-divided data. a** displays ROC curves comparing the differentiation of benign and malignant tumors in the Test and Henan external validation sets. **b** shows ROC curves comparing the differentiation of benign tumors in the Test and Henan external validation sets. **c** presents ROC curves comparing the identification of malignant tumors. **d** displays ACC for distinguishing between benign and malignant tumors in the Test and Henan external validation sets. **e** demonstrates ACC for distinguishing benign tumors in the Test and Henan external validation sets. **f** provides ACC for identifying malignant tumors in the HN, CD, GZ, and LS validation sets.

Corresponding modifications in the paper:

(1)Lines: 497-531

4. The provided web interface is currently not functioning as expected. This issue needs to be resolved to ensure that potential users can access and interact with the tool. The suggestion to invest in a virtual server with a proper domain is practical and should be considered. This

relatively small investment could markedly enhance the accessibility and usability of the web interface, thereby supporting wider adoption and validation of the tool by the research community.

Author Response: Thank you for your feedback. We agree that having a dedicated domain name enhances the professional appearance and credibility of our tools. Therefore, we have purchased a domain name to make it easier for users to find, remember, and access our interface. Our updated web page address is: http://www.liver.services. After checking our logs, we found that during the review period, the web page was running normally without any crashes. If you encounter issues such as the homepage easily refreshing after logging in, it may be due to network restrictions. To ensure that users can access and use the tool normally, we suggest trying the following methods to solve this problem: (1) Use different network environments, such as different Wi-Fi networks or mobile data connections; (2) Ensure that the browser and device settings do not have automatic refresh enabled. (3) Try clearing the browser cache and cookies. Additionally, to prevent access issues with the web due to internet problems, we provide portable software that allows users to experience the main functions of the tool. I have tested this software on Windows 10 and Windows 11, and it runs smoothly. The usage instructions are as follows:

Step 1:

Download software:

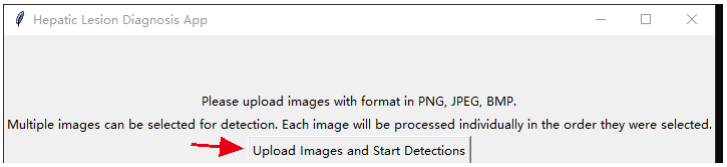
https://drive.google.com/file/d/1A0MvEwP0IhNRotAoKpZ5RMZYf7ktw_C8/view?usp=sharing

Download test data:

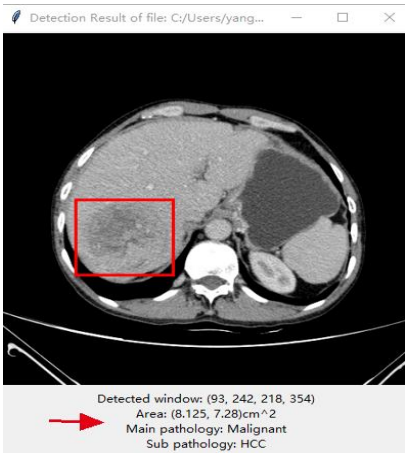
https://drive.google.com/file/d/1fvdQpuR2TZrUf6EyhhISHslsNKtTsHOe/view?usp=sharing

Step2: After decompressing the file, double-click run.bat

Step3: Click "Upload Images and Start Detections" to load one or more images for detection.



Step4: The results are shown



Reviewer #3 (Remarks to the Author): Expert in liver cancer clinical research and pathology, and digital pathology

Authors responded to the comments properly and the manuscript appears to be much improved.No further comments from me at this time.

Author Response: Thank you very much for your positive feedback and for acknowledging the improvements in our manuscript. We are delighted to hear that our revisions have addressed your comments satisfactorily. We greatly appreciate your time and effort in reviewing our manuscript and providing valuable suggestions. Your insights have significantly contributed to enhancing the quality of our work.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

The authors have addressed most of my comments. I do not think that Google Drive is an appropriate platform to disseminate the trained model. The state of the art is to upload the model to Zenodo or Huggingface (like here https://github.com/mahmoodlab/UNI). The non-commercial license is not an issue (like here https://github.com/mahmoodlab/UNI). Overall the approach to making the models accessible is not very professional here.

Responds to the reviewer's comments:

Reviewer #2 (Remarks to the Author): Expert in cancer digital pathology, artificial intelligence, and deep learning

Author The authors have addressed most of my comments. I do not think that Google Drive is an appropriate platform to disseminate the trained model. The state of the art is to upload the model to Zenodo or Huggingface (like here https://github.com/mahmoodlab/UNI). The non-commercial license is not an issue (like here https://github.com/mahmoodlab/UNI). Overall the approach to making the models accessible is not very professional here.

Response: Thank you for your feedback. We appreciate your suggestions on the appropriate platforms for disseminating the trained model. We uploaded the model to Zenodo [https://zenodo.org/records/12646854] to ensure it met the state-of-the-art standards for accessibility and professionalism, as you recommended. Additionally, we have linked the code to Zenodo and obtained a DOI, which is https://doi.org/10.5281/zenodo.12655750. The non-commercial license was included as demonstrated in the provided examples. Thank you for guiding us towards improving the accessibility and professional presentation of our models.