Unraveling the “black-box” of artificial intelligence-based pathological analysis of liver cancer

Artificial intelligence-based pathological analysis of malignancies has invoked tremendous insights into cancer prognostication and treatment efficacy. In recent years, the application of artificial intelligence regarding the management of liver cancer has published numerous studies. These pioneering studies showed that AI-based approaches could extract essential pathological features that were morphological determinants of the underlying molecular background and prognostic indicators. However, derived from the “black box” nature of the mainstream AI-approaches (ie, neural network), prominent limitations such as the tendency of shortcut learning, poor generalizability, and limited interpretability become the major obstacles.

1. Current advances of AI-based approaches for clinical management of liver cancer

1.1 AI-based diagnosis and segmentation of liver cancer

The AI-based diagnosis was the first implementation of computer vision in pathology. Many pioneering studies had demonstrated that AI could approach even surpass pathologists on same specific tasks with reduced inter-observer variability. The auto-diagnosis of liver cancer via biopsy or surgical specimens were the first application of AI-based techniques in this filed[1][2]. Kriegsmann et al. implemented deep learning algorithms in liver pathology to optimize the diagnosis of benign lesions and adenocarcinoma metastasis, which showed high prediction capability with a case accuracy of 94%[3]. In summary, the automated identification and diagnosis of tumor tissue in medical images is an effective replication of human pathologists’ jobs and help paving the path for more advanced tasks for AI, such as prognostification.

1.2 AI-based prognostication of liver cancer

For most studies using AI technology for prognostication of liver cancer, the auto-detection and segmentation of tumor tissue was the prior step. To date, all attempts tried to infer clinical endpoints directly from pathological images in the forms of “risk score”[1] [2]. Shi et al. conducted a fine work to explore prognostic indicators in the pathological images of HCC via weakly supervised deep learning framework[4]. They established a “tumor risk score (TRS)” to evaluate patient outcomes which had superior predictive ability compared to clinical staging systems. Saillard et al. also discussed the use of deep-learning algorithms on histological slides to predict survival after HCC resection[1]. They compared two different algorithms, CHOWDER and SCHMOWDER, on the same task. CHOWER directly predicts a risk score from WSIs without annotations while SCHMOWDER determined tumoral or non-tumoral regions in a supervised manner and then generated risk prediction based on attention mechanism. Although they both

outperformed the composite score of all other clinicopathological variables, SCHMOWDER had a significantly better performance than CHOWDER, highlighted the importance of combining expert knowledge with machine learning processes.

Qu et al. and Yamashita et al. both used deep learning to explore pathological signatures to predict recurrence of HCC after resection or liver transplantation[5] [6].

Other attempts tried using AI to infer recognized pathological prognosticators from pathological images, with the hope to relief pathologists from dull redundant routines. Chen et al. developed a deep learning model called MVI-DL to evaluated the presence of MVI in HCC from WSIs, achieved an AUC of 0.904[7]. Another group conducted a study using neural network to classifying well, moderate, and poor tumor differentiation of HCC, with 89.6% accuracy, such as MVI, tumor cell nuclei grading, HCC differentiation.

1.3 Molecular profiling of liver cancer via AI

Histological appearances of human cancers contain a massive amount of information related to their underlying molecular alterations. DL model can also help identify and analyze complex features or patterns which are related to specific molecular alterations.

Pioneering study by Fu et al. conducted a comprehensive study using deep transfer learning to analyze histopathological patterns covering 28 different cancer types[8]. They used a computational histopathological algorithm called PC-CHiP, which was trained on over 17,000 slide images. They found that the computational histopathological features learned by the algorithm were associated with various genomic alterations, including whole-genome duplications, chromosomal aneuploidies, focal amplifications and deletions, and driver gene mutations. The most predictable gene mutations including TP53, BRAF, PTEN. Gene expression levels also profoundly influenced the morphological fluctuations of cancer, reflecting various tumor composition or the extent of tumor-infiltrating lymphocytes. Overall, this state-of-art study demonstrated the potiential of computer vision in characterization of the molecular basis of tumor histopathology on a pan-cancer level.

Liao et al. used two datasets (one from TCGA and one from West China Hospital) to predict and validate the presence of specific somatic mutations[9]. Seven mutations were found be to accurately predicted by the deep-learning based platform, including ALB, CSMD3, CTNNB1, MUC4, OBSCN, TP53, and RYR2. The AUCs for these predictions were above 0.70, with CTNNB1 reached the highest value at 0.903 (CTM). Chen et al. also predicted the presence of specific genetic mutations[7]. Another study showed that DL could predict a subset of recurrent HCC genetic defects (CTNNB1, FMN2, TP53, and ZFX4) with AUCs ranging from 0.71 to 0.89 (NPJ).

1.4 Exploring predictive indicators for therapy response

Recent studies have focused on predicting molecular signatures and alterations that can indicate response to systemic therapies in cancer patients. In gastrointestinal cancers, neural networks (NNs) have been used to process digital slides, achieving high performance in predicting microsatellite instability, which is strongly associated with

sensitivity to immunomodulating therapies. Pan-cancer studies by Kather et al. (2020) and Fu et al. (2020) have also shown that NN models can predict a wide range of molecular alterations or signatures related to therapy response[8] [10].

For hepatocellular carcinoma (HCC), no molecular feature is currently used to predict response to systemic therapies. However, Sangro et al. reported that responses to the anti-PD1 antibody nivolumab were more frequently observed in patients with tumors showing overexpression of specific immune gene signatures[11]. This finding was further confirmed by Haber et al. observed increased sensitivity to immunotherapy in HCCs with upregulated interferon gamma and gene sets associated with antigen presentation[12]. Deep convolutional neural networks (DCNNs) can easily identify immune cells, suggesting that deep learning may be able to predict such gene expression profiles.

These studies share common limitations, such as limited patient numbers, sensitivity to staining protocols, and lack of prospective validation. Standardization of slide encoding and processing, as well as automated quality control of slides, will be crucial for comparing model performance and addressing artifacts like tissue folds or stains.

2. Current challenges limiting AI-based approaches in the management of liver cancer

a. Lack of standardization of image analysis[1].
b. Most of these different studies share the same limitations, including the limited number of patients, sensitivity to staining protocols and lack of prospective validation. The standardisation of slide encoding and processing will also be key to enable comparisons of model performance[2]. Finally, it will be critical to determine how predictions are impacted by artifacts such as tissue folds or stains. Automated quality control of slides may help to overcome these issues.

3. Strategies for unraveling the “black-box” of AI-based

Model-based explanations and post hoc explanations are two distinct strategies to understand and interpreter the “black-box” of machine learning models. The primary difference between them lies in the way they achieve explainability.

Model-based explanations focus on using inherently interpretable models, such as linear regression or support vector machines. These models are designed to be simple enough for humans to understand while still being capable of capturing the relationship between input and output variables. Model-based explanations often enforce sparsity or simulatability, limiting the number of features used or ensuring that the model’s decision-making process can be internally reasoned by humans.

In contrast, post hoc explanations analyze an already trained model, such as a deep neural network, to gain insights into the learned relationships. Unlike model-based explanations, which force the model to be explainable from the outset, post hoc explanations attempt to decipher the behavior of a complex, “black box” model after it has been trained. This

approach is particularly relevant for deep learning models, which typically have thousands to millions of weights and are not inherently interpretable.

Methods for post hoc explanations include examining learned features, feature importance, and feature interactions, as well as visual explanations through saliency maps. These techniques help to shed light on the inner workings of complex models, making them more understandable and accessible to humans.

Model-based explanations prioritize interpretability from the beginning by using simpler, more transparent models, while post hoc explanations focus on deciphering the behavior of complex, pre-trained models to provide insights into their decision-making processes.

3.1 Model-based explanation

3.1.1 Support vector machine or random forests vs. deep learning

Commonly used statistical models include linear regression, logistic regression, and Cox-proportional hazards regression, which are relatively intuitive to interpret. Classical machine learning techniques (such as random forest or support vector machine) rely on handcrafted features (assembled by human investigators, such as tumor size, roundness, symmetry and intensity). In other words, classical techniques can recapitulate and simulate the processing routine normally performed by human experts.

Deep learning methods (such as neural networks) usually have enormous amount of free parameters which was automatically found by the machine in the process of associating inputs with outputs. They can extract subtle features from complex data which are not immediately obvious to the human eye, thus would be defined as “Black box”.

DL methods usually outperform classical techniques and consequently dominate the field of AI in hepatology.

Other model-based explanation had stepwise framework design. Wang et al. first trained a CNN for automated segmentation and classification of individual nuclei at single-cell levels on HE sections of HCC, and performed feature extraction to identify 246 quantitative image features[13]. Then, a clustering analysis by an unsupervised learning approach identify three distinct histologic subtypes. Lu et al applied three pretrained CNN models to extract imaging features from HCC histopathology, then they performed supervised classification using a linear support vector machine (SVM) classifier to delineate tumor regions, and also conducted survival analysis using Cox proportional hazards (CoxPH) regression models[14]. However, these authors did not provide further in-depth interpretation of the underlying biological implications of relevant features.

3.1.2 Supervised learning vs. weakly supervised learning vs. unsupervised learning

Supervised learning perform training on a dataset that is labeled in relation to the class of interest, and this label is available to the algorithm while the model is being created, unsupervised learning involves training on a dataset that lacks class labels, yielding clusters of output data that subsequently require additional human inspection. In aspects of interpretation, supervised learning needs to answer the question of “how” the network

come to the output, whereas the unsupervised learning require human to comprehend "why" the network inferred its clustering results.

Some studies declared that the model was interpretable-by-design, such as the CHOWDER model established by Saillard et al to predict post-resection HCC prognosis[1]. However, this declared interpretability was based on pathologist assessment of the image tiles that the model defined as the most significantly associated with patient outcomes. Some features including vascular spaces and the macrotrabecular architectural pattern were identified as indicators of poor survival. The interpretability of the deep learning algorithm used by Liu et al. also leveraged on the similar strategy[15], and some histological features associated with high risk of post-resection recurrence of HCC were manually identified by pathologists, including the presence of stroma and nuclear hyperchromasia.

3.1.3 Textual explanation

There are two types of textual explanation: (1) image captioning, and (2) image captioning with visual explanation.

Image captioning focuses on generating textual descriptions for images using neural networks, while image captioning with visual explanation extends this approach by incorporating visual attention maps. The main difference between the two lies in the additional context and understanding provided by visual explanations.

Image captioning generates text based on the features extracted from images, but it does not explicitly show the relationship between the generated text and specific image features. On the other hand, image captioning with visual explanation provides visual attention maps that highlight the areas of the image that are most relevant to the generated text. This additional information makes the explanations more interpretable and helps users understand the reasoning behind the generated text.

3.1.3.1 Image captioning

Vinyals et al. (2015) proposed an end-to-end image captioning framework that combined a convolutional neural network (CNN) for image encoding and a long-short term memory (LSTM) network for textual encoding[16]. They used human-generated sentences for training and the BLEU metric for evaluation, which measures the precision of word N-grams between generated and reference sentences.

Singh et al. (2019) applied this framework to chest X-rays, using word-embedding databases GloVe and RadGloVe to train the LSTM[17]. GloVe is a general-purpose word-embedding database, while RadGloVe is a radiology-specific variant. They evaluated the performance using BLEU, METEOR, CIDER, and ROUGE metrics, which are different evaluation metrics for generated text. They found that using both RadGloVe and GloVe led to better performance in generating radiology reports compared to using only GloVe.

3.1.3.2 Image captioning with visual explanation

Zhang et al. (2017) introduced a framework using dual attention for both text and imaging[18]. They employed a similar approach to image captioning, with an image encoder and an LSTM for text, but added dual attention to facilitate high-level interactions between image and text predictions. This resulted in visual attention maps corresponding to textual explanations in histology images, providing a better understanding of the relationship between image features and generated text.

Wang et al. (2018) used a similar approach for chest X-rays, showing that different parts of the textual explanation led to different areas of saliency mapping in the image[19]. This demonstrated that the generated text was closely related to specific image features, making the explanations more interpretable. Lee et al. (2019a) applied this approach to breast mammograms, adding a visual word constraint loss to the text-generating LSTM to ensure that the provided explanations followed the correct jargon of breast mammography reports[20]. They found that adding this loss improved textual explanation quality and linked radiology reports to visual saliency maps, making the generated explanations more accurate and relevant to the domain.

3.1.4 Example-based explanation

1. Triplet network. A triplet network consists of three identical networks with shared parameters. It calculates the L2 distances between the representations in the latent space of input samples, allowing for unsupervised comparison of samples. This approach has been used in colorectal cancer histology (Peng et al., 2019)[21] and radiological picture archiving and communication systems (Yan et al., 2018)[22]. Codella et al. (2018) combined triplet loss with global average pooling to provide both example-based explanation and visual explanation in dermatology images of melanoma[23].

2. Prototypes. Chen et al. (2019) proposed using typical examples (prototypes) as explanation to reflect case-based reasoning that humans perform[24]. A prototype layer is added to the neural network, grouping training inputs according to their classes in the latent space. During testing, the method utilizes parts of the test image that resemble trained prototypes. Uehara et al. (2019) used prototypes to explain the classification of histology image patches as cancer or non-cancer[25].

3. Examples from the latent space. Sarhan et al. (2019) proposed learning disentangled representations of the latent space using a residual adversarial VAE with a total correlation constraint[26]. Biffi et al. (2020) provided a framework for explainable anatomical shape analysis using a ladder VAE[27]. Silva et al. (2018) proposed example-based explanation that showed similar and dissimilar cases for aesthetic results of breast surgery on photos and skin images on dermoscopy[28]. In later work, Silva et al. (2020) combined example-based explanation with saliency mapping for chest X-rays classification[29]. Sabour et al. (2017) showed that by replacing scalar feature maps with vectorized representations (i.e., capsules), they were able to encode high-level features of images[30]. LaLonde et al. (2020) used capsules for lung cancer diagnosis while also predicting visual attributes[31].

3.2 Post hoc explanation

3.2.1 Visual explanation (saliency mapping, pathologist-in-the-loop)

3.2.1.1 Backpropagation-based approaches

Backpropagation-based approaches include (Guided) backpropagation and deconvolution, Class Activation Mapping (CAM), Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), Deep SHapley Additive exPlanations (Deep SHAP), and trainable attntion.

1. (Guided) Backpropagation and Deconvolution. These early techniques create saliency maps by highlighting pixels with the highest impact on the analysis output. They provide local, model-specific (only for CNNs), post hoc explanations. Examples include visualization of partial derivatives of the output on pixel level (Simonyan et al., 2013)[32], deconvolution (Zeiler and Fergus, 2014)[33], and guided backpropagation (Springenberg et al., 2014)[34]. These methods have been used in medical image analysis, such as estimating the amount of coronary artery calcium per cardiac or chest CT image slice and visualizing the decision basis (de Vos et al., 2019)[35].

2. Class Activation Mapping (CAM). Introduced by Zhou et al. (2016)[36], CAM replaces the fully connected layers at the end of a CNN with global average pooling on the last convolutional feature maps. The class activation map is a weighted linear sum of the presence of visual patterns at different spatial locations. This technique provides local, model-specific, post hoc explanations and has been used in medical imaging by various researchers. For instance, Jiang et al. (2019) constructed an ensemble of Inception-V3, ResNet-152, and Inception-ResNetV2 to distinguish fundus images of healthy subjects or patients with mild diabetic retinopathy from those with moderate or severe diabetic retinopathy, providing a weighted combination of the resulting CAMs for localization of diabetic retinopathy[37]. Lee et al. (2019) constructed CAMs of the output of an ensemble of four CNNs: VGG-16, ResNet-50, Inception-V3, and InceptionResNet-V2, for the detection of acute intracranial hemorrhage[38]. Additionally, multi-scale CAMs have been proposed to better identify small structures, as demonstrated by Liao et al. (2019)[39] and Shinde et al. (2019)[40] in their respective studies.

3. Gradient-weighted Class Activation Mapping (Grad-CAM): Introduced by Selvaraju et al. (2017), Grad-CAM, is a generalization of CAM that can work with any type of CNN to produce post hoc local explanations[41]. Unlike CAM, Grad-CAM does not specifically require global average pooling. The authors also introduced guided Grad-CAM, an element-wise multiplication between guided backpropagation and Grad-CAM. Grad-CAM and guided Grad-CAM have been used in medical image analysis for various applications. For example, Ji (2019) used Grad-CAM to show the areas of histology lymph node sections on which a classifier based its decision for metastatic tissue[42]; Kowsari et al. (2020) employed it to pinpoint small bowel enteropathies on histology[43]; and Windisch et

al. (2020) utilized Grad-CAM to reveal the areas of brain MRI that influenced the classifier's decision on the presence of a tumor[44].

4. Layer-wise Relevance Propagation (LRP). Introduced by Bach et al. (2015), LRP uses the output of the neural network and iteratively backpropagates it throughout the network, assigning a relevance score to each input neuron from the previous layers[45]. LRP has been used in medical image analysis for various applications. For instance, Böhle et al. (2019) employed LRP to identify regions responsible for Alzheimer's disease from brain MR images, comparing the saliency maps provided by LRP with those generated by guided backpropagation[46]. They found that LRP was more specific in identifying known regions associated with Alzheimer's disease than guided backpropagation.

5. Deep SHapley Additive exPlanations (Deep SHAP). It is a unified approach proposed by Lundberg and Lee (2017) for explaining predictions using SHapley Additive exPlanations (SHAP)[47]. This model-agnostic approach uses Shapley values to determine the marginal contribution of every feature to the model's output individually. Deep SHAP has been used in medical image analysis for various applications. For example, van der Velden et al. (2020) employed a regression CNN to estimate the volumetric breast density from breast MRI and used Deep SHAP to explain which parts of the image had a positive contribution and which parts had a negative contribution to the density estimation[48].

6. Trainable Attention. This is a mechanism proposed by Jetley et al. (2018) that not only highlights where a network focuses on an image, but also determines the proportion of attention paid to different areas of the image for classification[49]. This method amplifies relevant areas and suppresses irrelevant ones. In the field of medical imaging, Schlemper et al. (2019) applied this concept and introduced grid attention, based on the observation that most objects of interest in medical images are highly localized[50]. The grid attention captured the anatomical information in medical images, demonstrating high performance for both segmentation and localization. They incorporated the attention gates into a UNET (Ronneberger et al., 2015)[51] and a variant of VGG (Simonyan and Zisserman, 2014)[52]. The attention coefficients were used to explain which areas of the image the network focused on.

3.2.1.2 Perturbation-based approaches

Perturbation-based approaches involve altering input images to assess the importance of specific areas for a given task. The approaches include occlusion sensitivity, LIME, meaningful perturbation, and prediction difference analysis.

1. Occlusion Sensitivity. This approach visualizes the most important parts of an image for classification by occluding certain areas and observing the impact on classification outcomes. Zeiler and Fergus (2014) demonstrated that a dog's breed could be misclassified as a tennis ball when the dog's face was occluded[33]. Their work highlighted the importance of understanding which parts of an image contribute to the classification decision.

2. Local Interpretable Model-agnostic Explanations (LIME). LIME provides local explanations by approximating complex models with simpler ones, such as replacing a CNN with a linear model. Ribeiro et al. (2016) developed LIME, which perturbs input data and learns the mapping between perturbed input and output changes using the simpler model[53]. This method has been applied in various domains, including medical image analysis, where it has been used to identify bloody regions in gastral endoscopy images, helping clinicians understand the model's decision-making process.

3. Meaningful Perturbation. Fong and Vedaldi (2017) introduced this approach, which involves altering input images to detect changes in neural network predictions using naturalistic or plausible effects[54]. Uzunova et al. (2019) argued that Fong and Vedaldi's perturbations were not suitable for medical images and proposed replacing pathological regions with healthy tissue equivalents using a variational autoencoder (VAE)[55]. This approach demonstrated better localization of pathology in various imaging studies, such as optical coherence tomography images of the eye and MRI of the brain, providing more meaningful explanations for clinicians.

4. Inpainting-based Perturbation. Lenis et al. (2020) used inpainting to replace pathological regions with healthy tissue equivalents[56], similar to Uzunova et al.'s approach. They showed that inpainting-based perturbations outperformed backpropagation and Grad-CAM in pinpointing masses in breast mammography and tuberculosis on chest X-rays. By providing more accurate localization of pathological regions, this approach helps clinicians better understand the model's decision-making process and potentially improve patient care.

5. Prediction Difference Analysis. Zintgraf et al. (2017) adapted this approach for generating saliency maps by assigning relevance values to each pixel based on how the prediction changes when the pixel is considered unknown[57]. They expanded this approach by adding conditional sampling and multivariable analysis, allowing for more accurate and informative saliency maps. Seo et al. (2020) combined prediction difference analysis with superpixels (or supervoxels for 3D) on multiple scales to create visually pleasing saliency maps that follow image edges[58]. These maps provided explanations for distinguishing between Alzheimer's disease patients and normal controls, offering valuable insights into the model's decision-making process.

3.2.1.3 Multiple instance learning-based approaches

Multiple instance learning involves training sets with labeled bags of instances, where the instances themselves are unlabeled.

Schwab et al. (2020) developed a patch-based approach for localizing critical findings in chest X-ray images[59]. They used multiple instance learning to assign predictions to individual image patches. By overlaying these predictions on the original image, they were able to visualize the areas on which the classifier based its decision.

Araújo et al. (2020) applied this approach to explain which areas of a fundus photograph were important for diabetic retinopathy diagnosis[60]. They assessed the severity of the disease using an ordinal scale with grades from 0 to 5. Utilizing a patch-based approach, they generated visual explanation maps for each diabetic retinopathy grade.

3.2.2 Textual explanation

1. Testing with Concept Activation Vectors (TCAV). Kim et al. (2018) introduced a technique for making neural networks more interpretable by generating human-understandable explanations of their internal states[61]. TCAV uses concept activation vectors (CAVs) to measure a model's sensitivity to high-level concepts, such as 'stripes' for zebras or 'spiculated mass' for cancer. These concepts can be provided after training of the neural network as a post hoc analysis. The TCAV algorithm uses user-defined sets of examples of a concept and random non-concept examples. The authors demonstrated the feasibility of TCAV in a medical image processing example, relating physician annotations like 'microaneurysm' to diabetic retinopathy in fundus imaging.

2. Cardiac Disease Identification. In an application of TCAV for medical diagnosis, Clough et al. (2019) identified cardiac disease in cine-MRI by classifying the latent space of a VAE[62]. They used TCAV to show which clinically known biomarkers were related to cardiac disease. Furthermore, they reconstructed images with low peak ejection rate – a characteristic that might be related to cardiac disease – by adding the CAV to the latent space.

3. Regression Concept Vectors. Building upon TCAV, Graziani et al. (2020) introduced regression concept vectors, which indicate continuous-valued measures of a concept, such as tumor size[63]. This can be useful when investigating a continuous concept like tumor size. They demonstrated that regression concept vectors could explain why a network classified different areas of a breast histopathology image as cancerous or healthy based on the concepts 'contrast' and 'nuclei area'. The concept 'nuclei area' refers to a clinically used system for evaluating cell size, which was different between healthy and cancerous regions.

3.2.3 Example-based explanation

1. Influence Functions. Koh and Liang (2017) proposed using influence functions to explain which inputs from a training set a decision was based on[64]. They provided an efficient approximation using influence functions (Cook and Weisberg, 1980)[65]. Wang et al. (2019) used influence functions to explain classifications of liver lesions on multiphase MRI and their association with radiological characteristics[66].

4. Conclusion and future applications

Artificial Intelligence (AI) has shown immense potential in the pathological analysis of liver cancer, providing valuable insights into cancer prognostication and treatment efficacy. Despite the challenges posed by the "black-box" nature of AI, various strategies

such as model-based explanations and post hoc explanations have been developed to interpret these complex models.

The application of AI in liver cancer management has made significant strides, from diagnosis and segmentation to prognostication and molecular profiling. However, the field is still in its infancy, with many challenges to overcome, including standardization of image analysis and addressing limitations such as sensitivity to staining protocols and lack of prospective validation.

Looking forward, the integration of AI in clinical practice could revolutionize liver cancer management. AI-based approaches could potentially automate routine tasks, reducing workload for pathologists and improving diagnostic accuracy. Furthermore, AI could aid in prognostication by extracting essential pathological features that serve as indicators of underlying molecular backgrounds.

Moreover, AI could play a crucial role in therapy response prediction. By identifying molecular signatures and alterations indicative of systemic therapy response, AI could guide personalized treatment plans for liver cancer patients.

In the future, we anticipate further advancements in AI technology that will enhance its interpretability and applicability in liver cancer management. This includes the development of more sophisticated models that can provide more accurate and interpretable predictions, as well as the integration of AI with other technologies such as genomics and proteomics for a more comprehensive understanding of liver cancer.

In conclusion, while there are challenges to be addressed, the future of AI in liver cancer management looks promising. With continued research and development, AI has the potential to significantly improve liver cancer diagnosis, prognostication, and treatment, ultimately leading to better patient outcomes.

References

[1] Saillard C, Schmauch B, Laifa O, et al Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides[J] Hepatology, 2020, 72(6): 2000-2013.

[2] Calderaro J, Seraphin T P, Luedde T, et al Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma[J] Journal of Hepatology, 2022, 76(6): 1348-1361.

[3] Kriegsmann M, Kriegsmann K, Steinbuss G, et al Implementation of deep learning in liver pathology optimizes diagnosis of benign lesions and adenocarcinoma metastasis[J] Clinical and Translational Medicine, 2023, 13(7): e1299.

[4] Shi J Y, Wang X, Ding G Y, et al Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning[J] Gut, 2021, 70(5): 951-961.

[5] Qu W F, Tian M X, Lu H W, et al Development of a deep pathomics score for predicting hepatocellular carcinoma recurrence after liver transplantation[J] Hepatology International, 2023: 1-15.

[6] Yamashita R, Long J, Saleem A, et al Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images[J] Scientific reports, 2021, 11(1): 2047.

[7] Chen Q, Xiao H, Gu Y, et al Deep learning for evaluation of microvascular invasion in hepatocellular carcinoma from tumor areas of histology images[J] Hepatology International, 2022, 16(3): 590-602.

[8] Fu Y, Jung A W, Torne R V, et al Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis[J] Nature cancer, 2020, 1(8): 800-810.

[9] Liao H, Long Y, Han R, et al Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma[J] Clinical and translational medicine, 2020, 10(2).

[10] Kather J N, Heij L R, Grabsch H I, et al Pan-cancer image-based detection of clinically actionable genetic alterations[J] Nature cancer, 2020, 1(8): 789-799.

[11] Sangro B, Sarobe P, Hervás-Stubbs S, et al Advances in immunotherapy for hepatocellular carcinoma[J] Nature reviews Gastroenterology & hepatology, 2021, 18(8): 525-543.

[12] Haber P K, Castet F, Torres-Martin M, et al Molecular markers of response to anti-PD1 therapy in advanced hepatocellular carcinoma[J] Gastroenterology, 2023, 164(1): 72-88 e18.

[13] Wang H, Jiang Y, Li B, et al Single-cell spatial analysis of tumor and immune microenvironment on whole-slide image reveals hepatocellular carcinoma subtypes[J] Cancers, 2020, 12(12): 3562.

[14] Lu L, Daigle Jr B J Prognostic analysis of histopathological images using pre-trained convolutional neural networks: application to hepatocellular carcinoma[J] PeerJ, 2020, 8: e8668.

[15] Liu Z, Liu Y, Zhang W, et al Deep learning for prediction of hepatocellular carcinoma recurrence after resection or liver transplantation: a discovery and validation study[J] Hepatology international, 2022, 16(3): 577-589.

[16] Vinyals O, Toshev A, Bengio S, et al Show and tell: A neural image caption generator[C]//Proceedings of the IEEE conference on computer vision and pattern recognition 2015: 3156-3164.

[17] Singh S, Karimi S, Ho-Shon K, et al From chest x-rays to radiology reports: a multimodal machine learning approach[C]//2019 Digital Image Computing: Techniques and Applications (DICTA) IEEE, 2019: 1-8.

[18] Zhang Z, Chen P, Sapkota M, et al Tandemnet: Distilling knowledge from medical images using diagnostic reports as optional semantic references[C]//Medical Image Computing and Computer Assisted Intervention− MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20 Springer International Publishing, 2017: 320-328.

[19] Wang X, Peng Y, Lu L, et al Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays[C]//Proceedings of the IEEE conference on computer vision and pattern recognition 2018: 9049-9058.

[20] Lee H, Kim S T, Ro Y M Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis[C]//Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9 Springer International Publishing, 2019: 21-29.

[21] Peng T, Boxberg M, Weichert W, et al Multi-task learning of a deep k-nearest neighbour network for histopathological image classification and retrieval[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22 Springer International Publishing, 2019: 676-684.

[22] Yan K, Wang X, Lu L, et al Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion data-base[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018: 9261-9270.

[23] Codella N C F, Lin C C, Halpern A, et al Collaborative human-AI (CHAI): Evidence-based interpretable melanoma classification in dermoscopic images[C]//Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1 Springer International Publishing, 2018: 97-105.

[24] Chen C, Li O, Tao D, et al This looks like that: deep learning for interpretable image recognition[J] Advances in neural information processing systems, 2019, 32.

[25] Uehara K, Murakawa M, Nosato H, et al Prototype-based interpretation of pathological image analysis by convolutional neural networks[C]//Asian Conference on Pattern Recognition Cham: Springer International Publishing, 2019: 640-652.

[26] Sarhan M H, Eslami A, Navab N, et al Learning interpretable disentangled representations using adversarial vaes[C]//Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1 Springer International Publishing, 2019: 37-44.

[27] Biffi C, Cerrolaza J J, Tarroni G, et al Explainable anatomical shape analysis through deep hierarchical generative models[J] IEEE transactions on medical imaging, 2020, 39(6): 2088-2099.

[28] Silva W, Fernandes K, Cardoso M J, et al Towards complementary explanations using deep neural networks[C]//Understanding and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1 Springer International Publishing, 2018: 133-140.

[29] Silva W, Poellinger A, Cardoso J S, et al Interpretability-guided content-based medical image retrieval[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 Springer International Publishing, 2020: 305-314.

[30] Sabour S, Frosst N, Hinton G E Dynamic routing between capsules Advances in Neural Information Processing Systems 30[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 3856-3866.

[31] LaLonde R, Torigian D, Bagci U Encoding visual attributes in capsules for explainable medical diagnoses[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 Springer International Publishing, 2020: 294-304.

[32] Simonyan K, Vedaldi A, Zisserman A Deep inside convolutional networks: Visualising image classification models and saliency maps[J] arXiv preprint arXiv:13126034, 2013.

[33] Zeiler M D, Fergus R Visualizing and understanding convolutional networks[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13 Springer International Publishing, 2014: 818-833.

[34] Springenberg J T, Dosovitskiy A, Brox T, et al Striving for simplicity: The all convolutional net[J] arXiv preprint arXiv:14126806, 2014.

[35] De Vos B D, Wolterink J M, Leiner T, et al Direct automatic coronary calcium scoring in cardiac and chest CT[J] IEEE transactions on medical imaging, 2019, 38(9): 2127-2138.

[36] Zhou B, Khosla A, Lapedriza A, et al Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 2921-2929.

[37] Jiang H, Yang K, Gao M, et al An interpretable ensemble deep learning model for diabetic retinopathy disease classification[C]//2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC) IEEE, 2019: 2045-2048.

[38] Lee H, Yune S, Mansouri M, et al An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets[J] Nature biomedical engineering, 2019, 3(3): 173-182.

[39] Liao W M, Zou B J, Zhao R C, et al Clinical interpretable deep learning model for glaucoma diagnosis[J] IEEE journal of biomedical and health informatics, 2019, 24(5): 1405-1412.

[40] Shinde S, Chougule T, Saini J, et al HR-CAM: Precise localization of pathology using multi-level learning in CNNS[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22 Springer International Publishing, 2019: 298-306.

[41] Selvaraju R R, Cogswell M, Das A, et al Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision 2017: 618-626.

[42] Ji J Gradient-based interpretation on convolutional neural network for classification of pathological images[C]//2019 International Conference on Information Technology and Computer Application (ITCA) IEEE, 2019: 83-86.

[43] Kowsari K, Sali R, Ehsan L, et al Hmic: Hierarchical medical image classification, a deep learning approach[J] Information, 2020, 11(6): 318.

[44] Windisch P, Weber P, Fürweger C, et al Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices[J] Neuroradiology, 2020, 62: 1515-1518.

[45] Bach S, Binder A, Montavon G, et al On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation[J] PloS one, 2015, 10(7): e0130140.

[46] Böhle M, Eitel F, Weygandt M, et al Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification[J] Frontiers in aging neuroscience, 2019, 11: 194.

[47] Lundberg S M, Lee S I A unified approach to interpreting model predictions[J] Advances in neural information processing systems, 2017, 30.

[48] van der Velden B H M, Janse M H A, Ragusi M A A, et al Volumetric breast density estimation on MRI using explainable deep learning regression[J] Scientific reports, 2020, 10(1): 18095.

[49] Jetley S, Lord N A, Lee N, et al Learn to pay attention[J] arXiv preprint arXiv:180402391, 2018.

[50] Schlemper J, Oktay O, Schaap M, et al Attention gated networks: Learning to leverage salient regions in medical images[J] Medical image analysis, 2019, 53: 197-207.

[51] Ronneberger O, Fischer P, Brox T Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Conference Proceedings 2022.

[52] Simonyan K, Zisserman A Very deep convolutional networks for large-scale image recognition[J] arXiv preprint arXiv:14091556, 2014.

[53] Ribeiro M T, Singh S, Guestrin C 《 Why should i trust you?》 Explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016: 1135-1144.

[54] Fong R C, Vedaldi A Interpretable explanations of black boxes by meaningful perturbation[C]//Proceedings of the IEEE international conference on computer vision 2017: 3429-3437.

[55] Uzunova H, Ehrhardt J, Kepp T, et al Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders[C]//Medical Imaging 2019: Image Processing SPIE, 2019, 10949: 264-271.

[56] Lenis D, Major D, Wimmer M, et al Domain aware medical image classifier interpretation by counterfactual impact analysis[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 Springer International Publishing, 2020: 315-325.

[57] Zintgraf L M, Cohen T S, Adel T, et al Visualizing deep neural network decisions: Prediction difference analysis[J] arXiv preprint arXiv:170204595, 2017.

[58] Seo D, Oh K, Oh I S Regional multi-scale approach for visually pleasing explanations of deep neural networks[J] IEEE Access, 2019, 8: 8572-8582.

[59] Schwab E, Gooßen A, Deshpande H, et al Localization of critical findings in chest X-ray without local annotations using multi-instance learning[C]//2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) IEEE, 2020: 1879-1882.

[60] Araujo T, Aresta G, Mendonça L, et al DR| GRADUATE: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images[J] Medical Image Analysis, 2020, 63: 101715.

[61] Kim B, Wattenberg M, Gilmer J, et al Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]//International conference on machine learning PMLR, 2018: 2668-2677.

[62] Clough J R, Oksuz I, Puyol-Antón E, et al Global and local interpretability for cardiac MRI classification[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention Cham: Springer International Publishing, 2019: 656-664.

[63] Graziani M, Andrearczyk V, Marchand-Maillet S, et al Concept attribution: Explaining CNN decisions to physicians[J] Computers in biology and medicine, 2020, 123: 103865.

[64] Koh P W, Liang P Understanding black-box predictions via influence functions[C]//International conference on machine learning PMLR, 2017: 1885-1894.

[65] Cook R D, Weisberg S Characterizations of an empirical influence function for detecting influential cases in regression[J] Technometrics, 1980, 22(4): 495-508.

[66] Wang C J, Hamm C A, Savic L J, et al Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features[J] European radiology, 2019, 29: 3348-3357.