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

Artificial intelligence (AI)-based pathological analysis of malignancies has provided profound insights into cancer prognostication and treatment efficacy. In recent years, numerous studies have been published on the application of AI in managing liver cancer, including hepatocellular carcinoma (HCC), intrahepatic cholangiocarcinoma (ICC), and other primary or metastatic liver cancers. These pioneering studies have demonstrated that AI-based approaches can extract critical pathological features that serve as morphological determinants of the underlying molecular background and prognostic indicators.

However, due to the “black box” nature of mainstream AI approaches, such as neural networks, significant limitations have emerged. These include a tendency towards shortcut learning, poor generalizability, and limited interpretability, which have become major obstacles to their widespread clinical adoption. Furthermore, the ability to explain, justify, and understand the decisions made by deep learning methods is crucial to establishing a trust relationship between AI systems and pathologists before these systems can assist in clinical routines.

Researchers in the interdisciplinary field of computational pathology and liver cancer have already begun to address this issue, increasingly using multiple approaches to explain their algorithms. A well-constructed explanation not only ensures the reliability of the method by comparing its attention regions with expert knowledge but also provides new insights into the biological imprints of this malignancy. However, due to the intrinsic complexity of the model, understanding the mechanism remains a bottleneck in current studies, presenting both challenges and promising future research areas.

In this review, we will first outline the current technical developments for AI-based pathological analysis of liver cancer. We will then focus on the strategies these studies have adopted to unravel the “black box”.

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

1.1 AI-based diagnosis and segmentation of liver cancer

AI-based diagnosis was the first implementation of computer vision in pathology. Many pioneering studies have demonstrated that AI can approach, and even surpass, pathologists in specific tasks with reduced inter-observer variability. The auto-diagnosis of liver cancer via biopsy or surgical specimens was the first application of AI-based techniques in this field. 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%. Liao et al. used a

CNN to distinguish HCC from adjacent normal tissues with AUCs above 0.90. Kiana et al. developed a tool that can classify image patches as HCC or ICC with an accuracy of 0.88. In summary, the automated identification and diagnosis of tumor tissue in medical images is an effective replication of human pathologists' jobs and helps pave the way for more advanced tasks for AI, such as prognostication.

1.2 AI-based prognostication of liver cancer

In 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 have tried to infer clinical endpoints directly from pathological images in the form of a "risk score". Shi et al. conducted a fine work to explore prognostic indicators in the pathological images of HCC via a weakly supervised deep learning framework. 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. They compared two different algorithms, CHOWDER and SCHMOWDER, on the same task. CHOWDER 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 an attention mechanism. Although they both outperformed the composite score of all other clinicopathological variables, SCHMOWDER had significantly better performance than CHOWDER, highlighting 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, hoping to guide more personalized adjuvant therapy. Other attempts tried using AI to infer recognized pathological prognosticators from pathological images, such as MVI, tumor cell nuclei grading, and differentiation, with the hope of relieving pathologists from dull, redundant routines. Chen et al. developed a deep learning model called MVI-DL to evaluate the presence of MVI in HCC from WSIs, achieving an AUC of 0.904. Another group conducted a study using a neural network to classify well, moderate, and poor tumor differentiation of HCC, with 89.6% accuracy. To date, these prognostication studies have been restricted to HCCs, while other less common histological types have not been involved.

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 models can also help identify and analyze complex features or patterns that are related to specific molecular alterations.

A pioneering study by Fu et al. conducted a comprehensive analysis using deep transfer learning to analyze histopathological patterns covering 28 different cancer types. 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 included TP53,

BRAF, and PTEN. Gene expression levels also profoundly influenced the morphological fluctuations of cancer, reflecting various tumor compositions or the extent of tumor-infiltrating lymphocytes. Overall, this state-of-the-art study demonstrated the potential of computer vision in characterizing the molecular basis of tumor histopathology on a pan-cancer level.

In the research area of liver cancer, similar attempts have also been reported. 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. Seven mutations were found to be 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 reaching the highest value at 0.903 (CTM). Chen et al. also predicted the presence of specific genetic mutations. 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. Compared to prognostic studies, the cohorts used to infer molecular alterations were much smaller, especially the validation set, thus limiting the reliability of the findings.

1.4 Exploring predictive indicators for therapy response

Recent studies have also 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.

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. This finding was further confirmed by Haber et al., who observed increased sensitivity to immunotherapy in HCCs with upregulated interferon gamma and gene sets associated with antigen presentation. Deep convolutional neural networks (DCNNs) can easily identify immune cells, suggesting that deep learning may be able to predict such gene expression profiles.

2. Explainable/Interpretable AI could pave the way to clinical implementation

Despite numerous studies forecasting a future where AI dominates the medical management of liver cancer, none have yet made a significant clinical impact. The developers of these AI systems face a multitude of challenges that must be overcome before their solutions can gain acceptance from clinicians. A key limitation of deep learning approaches is their propensity for shortcut learning. This means that deep neural networks often establish connections by taking shortcuts, bypassing the intended solution. This leads to a lack of generalization and can result in unexpected failures. In a

pathological context, these shortcuts may include data artifacts, non-universal features, and other irrelevant information that can obscure the true relationships. However, if the transformation from inputs to outputs can be understood by human experts, the learned relationships would be more rational and authentic, thereby reducing the risk of overfitting. Given the high stakes of medical decision-making, any auxiliary medical system must be thoroughly inspected by human experts before it can provide reliable advice to users.

Strategies for unraveling the "black box" of AI can be categorized into two distinct types: model-based explanations and post hoc explanations. The primary difference between them lies in their approach to achieving explainability. Model-based explanations rely on inherently interpretable models, such as linear regression or support vector machines. These models are designed to be simple enough for humans to understand, yet 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. Conversely, 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 require the model to be explainable from the outset, post hoc explanations seek 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.

While hepatologists are increasingly recognizing the importance of achieving "transparency" in their AI algorithms, few studies have successfully met this objective. In the following chapter, we will summarize the mainstream approaches that have been, or could be, applied in liver cancer research to deconstruct the model and identify key pathological features.

3. Strategies for unraveling the "black-box" of AI-based liver cancer models

3.1 Classical machine learning techniques have model-based explanations

Statistical models such as linear regression, logistic regression, and Cox-proportional hazards regression are frequently employed due to their relative ease of interpretation. Similarly, classical machine learning techniques, including support vector machines and random forests, depend on handcrafted features. These features, assembled by human investigators, include aspects such as tumor size, roundness, symmetry, and intensity. In essence, these classical techniques can mimic and simulate the processing routine typically performed by human experts. Initial computational pathology, based on these hand-crafted, human-interpretable features (HIFs), extracted from regions of interest, has provided valuable diagnostic and prognostic information. In the field of hepatology, Lu et al. utilized three pre-trained CNN models to extract imaging features from hepatocellular carcinoma (HCC) histopathology. They then performed supervised classification using a

linear support vector machine (SVM) classifier to delineate tumor regions and conducted survival analysis using Cox proportional hazards (CoxPH) regression models. Similarly, Wang et al. trained a CNN for automated segmentation and classification of individual nuclei at single-cell levels on H&E sections of HCC. They then performed feature extraction to identify 246 quantitative image features. Following this, an unsupervised learning approach was used for clustering analysis, which identified three distinct histologic subtypes. These frameworks combined neural networks with statistical methods or classical machine learning techniques in a stepwise manner, making them inherently interpretable. However, the authors did not provide an in-depth interpretation of the underlying biological implications of the relevant features. Moreover, these frameworks did not fully leverage the potential of deep learning, which can automatically identify and extract relevant morphological features from high-dimensional input data.

3.2 Model explanation by visual inspection through backpropagation and deconvolution

Backpropagation and deconvolution are early techniques that generate saliency maps by emphasizing pixels with the most significant impact on the output of the analysis. These techniques provide local, model-specific (exclusive to CNNs), post hoc explanations. Examples include the visualization of partial derivatives of the output at the pixel level (Simonyan et al., 2013), deconvolution (Zeiler and Fergus, 2014), and guided backpropagation (Springenberg et al., 2014). These methods have been widely 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).

In hepatology, the fully interpretable model proposed by Saillard et al. was based on this approach. The interpretability of this model relied on a pathologist's assessment of the image tiles that the network identified as most significantly associated with patient outcomes. Some features identified in this process, including the macrotrabecular-massive subtype and cellular atypia, were previously shown to be predictors of poor outcomes, which validated the model's rationality. Another feature, the presence of vascular spaces, was also identified as an indicator of poor survival. This insightful work underscores the necessity of human-machine interactions and highlights the importance of model deconstruction.

The interpretability of the deep learning algorithm used by Liu et al. also hinged on a similar strategy. Some histological features associated with a high risk of post-resection recurrence of HCC were manually identified, including the presence of stroma and nuclear hyperchromasia.

3.3 Class activation mapping and Gradient-weighted Class Activation Mapping

Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (Grad-CAM) are both backpropagation-based techniques that provide post hoc explanations. CAM modifies the architecture of a CNN by replacing the fully connected layers at the end 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. However, CAM's applicability is limited as it can only be applied to models with global average pooling layers.

In contrast, Grad-CAM overcomes this limitation by using the gradients of the model's output with respect to the input image. It calculates the importance of each pixel by considering the gradient values, which indicate how much each pixel influences the final prediction. Grad-CAM can be applied to a broader range of models, including those without global average pooling layers. In essence, while CAM relies on the weights of the final convolutional layer, Grad-CAM utilizes the gradients of the model's output to generate heatmaps that highlight important regions in an image. Grad-CAM is a more versatile and widely applicable technique compared to CAM.

Shi et al. adapted CAM to create Risk Activation Mapping (RAM) in their study, aiming to visualize the pathological phenotypes of Hepatocellular Carcinoma (HCC) associated with patient risk. The heatmaps generated through RAM indicated regions of the tissue potentially associated with increased or reduced risk. By analyzing these heatmaps, pathologists identified specific features such as sinusoidal capillarization, prominent nucleoli and karyotheca, the nucleus/cytoplasm ratio, and inflammatory cells that were relevant to patient prognosis. Consequently, the researchers concluded that their deep learning framework could effectively decode pathological images and provide previously unnoticed biological information for HCC.

3.4 Feature extraction by attention mechanism

The attention mechanism is inspired by human biological systems that tend to focus on distinctive parts when processing large amounts of information. In the field of medical imaging, the attention layer in a neural network can highlight the areas of an image that the model focuses on and determines the proportion of attention paid to different areas of the image for classification. This method amplifies relevant areas and suppresses irrelevant ones. 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. 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) and a variant of VGG (Simonyan and Zisserman, 2014). The attention coefficients were used to explain which areas of the image the network focused on.

Through the attention mechanism, the deep learning algorithm by Qu et al. identified immune cells as the most significant tissue category for predicting HCC recurrence post-transplantation. The researchers then performed multiplex immunofluorescence to explore the immune landscape and identified intratumoral NK cells as the most relevant subgroup. Although this conclusion was based on a small number of patients and lacked validation, it provides a multi-modal explanation approach to rationalize the outputs from the deep learning model.

3.5 Other promising approaches

Numerous innovative approaches are being explored to identify key features and novel biomarkers through AI-based pathological analysis, many of which have yet to be applied to liver cancer. Among these, perturbation-based methods and textual explanation techniques appear to hold significant promise.

Perturbation-based methods involve modifying input images to evaluate the significance of specific regions for a given task. This strategy encompasses techniques such as the occlusion sensitivity map (OSM) and local interpretable model-agnostic explanations (LIME). The OSM technique visualizes the most crucial parts of an image for classification by obscuring certain areas and observing the effect on classification results. A study by Zeiler and Fergus (2014) revealed that occluding a dog's face could lead to misclassification of the breed as a tennis ball, underscoring the importance of understanding which image parts contribute to the classification decision. LIME, on the other hand, offers local explanations by approximating complex models with simpler ones, such as substituting a CNN with a linear model. Ribeiro et al. (2016) developed LIME, which perturbs input data and learns the correlation between perturbed input and output changes using a simpler model. This method has found applications in various fields, including medical image analysis, where it has been used to identify bloody regions in gastric endoscopy images, thereby aiding clinicians in understanding the model's decision-making process.

In contrast to the visual explanations mentioned above, AI models employing textual explanation can directly generate human-understandable semantics of their internal states. For instance, Testing with Concept Activation Vectors (TCAV) uses concept activation vectors (CAVs) to gauge a model's sensitivity to high-level concepts, such as 'stripes' for zebras or 'spiculated mass' for cancer. These concepts can be provided after the neural network's training as a post hoc analysis. The TCAV algorithm uses user-defined sets of concept examples and random non-concept examples. The feasibility of TCAV has been demonstrated in a medical image processing example, linking physician annotations like 'microaneurysm' to diabetic retinopathy in fundus imaging. Building on TCAV, Graziani et al. (2020) introduced regression concept vectors, which represent continuous-valued measures of a concept, such as tumor size. This can be particularly 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 varied between healthy and cancerous regions.

These promising approaches may be utilized in future studies, providing valuable insights into the key features and biological behavior of liver cancer.

4. Conclusions and outlook

The application of AI in liver cancer management has made significant advancements, spanning from diagnosis to prognostication and molecular profiling. As we look to the

future, the integration of AI into clinical practice holds the potential to revolutionize liver cancer management. AI-based methodologies could automate routine tasks, thereby reducing the workload for pathologists and enhancing diagnostic accuracy. Furthermore, AI could assist in prognostication by extracting critical pathological features that serve as indicators of underlying molecular backgrounds. Additionally, AI could play a pivotal role in predicting therapy responses. By identifying molecular signatures and alterations indicative of systemic therapy response, AI could guide the development of personalized treatment plans for liver cancer patients.

However, the field is still nascent, with numerous challenges to surmount, such as the standardization of image analysis and addressing limitations such as poor generalizability, low sensitivity to staining protocols, and lack of prospective validation. Among these challenges, the need for interpretability will become increasingly significant, as it is the key to resolving many other limitations. Interpretability ensures transparency in AI predictions, fosters trust among clinicians, and facilitates meaningful collaboration between AI systems and medical professionals. Thus, interpretability is not merely a technical necessity, but also a clinical one, serving as the key to fully harness the potential of AI in liver cancer management.

Looking ahead, 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 appears promising. With continued research and development, AI has the potential to significantly improve liver cancer diagnosis, prognostication, and treatment, ultimately leading to improved patient outcomes.

References