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 (including HCC, ICC, and other primary or metastatic cancers of liver) 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 that restrict their widespread clinical adoption. Also, the ability to explain, justify and understand the decisions made by deep learning methods is essential to establish a relationship of trust between AI systems and pathologists before it could assist in clinical routines. Researchers in this interdisciplinary field of computational pathology and liver cancer have already focused on this issue, and increasingly used multiple approaches to explain their algorithms. A good explanation could not only ensure the reliability of the method by comparing its attention regions with expert knowledge, but also provide new insights into the biological imprints of this malignancy. However, owing to the intrinsic complexity of the model, understanding the mechanism is still the bottleneck of current studies, and providing challenging and promising future research areas.

In this review, we will first outline the current technical developments for AI-based pathological analysis of liver cancer, and then emphasized on the strategies that these studies 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

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]. Liao et al used a CNN to distinguish HCC from adjacent normal tissues with AUCs above 0.90. Kiana et al. developed a tool able to 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 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, hoping to guide more personalized adjuvant therapy[5] [6]. 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 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. To date, these prognostication studies were restricted to HCCs, while other less common histological types were not 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 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 included TP53, BRAF, 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-art study demonstrated the potential of computer vision in characterization of 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[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). Compare to prognostic studies, the cohorts used to infer molecular alterations were much smaller, especially the validation set, thus limited 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[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.

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

Although numerous studies have depicted the future envision of AI-dominated medical management of liver cancer, none of which have made a clinical impact. The creators of these AI-systems must overcome multifarious hurdles before they can be approved by clinicians. One major limitation of deep learning approaches is the tendency of shortcut learning, which means the deep neural networks tend to establish connections by taking shortcuts instead of learning the intended solution, leading to a lack of generalisation and unintuitive failures. In pathological scenario, the shortcuts include data artifacts, non-universal features, and other irrelevant information that could obscure the true relationships. However, if the transformation from inputs to outputs could be comprehended by human experts, the learned relationships would be more rationalized and authentic, reducing the risk of overfitting. Also, the stakes of medical decision making are high; any auxiliary medical system should be inspected thoroughly by human experts before it could provide trustful opinions to the users.

The strategies for unraveling the “black box” could be summarized into two distinct kinds, model-based explanations and post hoc explanations. 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.

Although hepatologists have increasingly recognized the importance of achieving the “transparency” of their AI algorithms, fewer studies have sufficiently achieved this goal. In the next chapter, we will summarize the mainstream approaches that have been or could been applied in liver cancer 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

Commonly used statistical models include linear regression, logistic regression, and Cox-proportional hazards regression, which are relatively intuitive to interpret. Similarly, classical machine learning techniques (such as support vector machine or random forest) 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. Initial computational pathology based on hand-crafted, human-interpretable features (HIFs) extracted from regions of interest provided valuable diagnostic and prognostic information. In hepatology, 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]. 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. These frameworks step-wisely combined neural network with statistical methods or classical machine learning techniques, thus were interpretable by design. However, these authors did not provide further in-depth interpretation of the underlying biological implications of the relevant features. Also, these frameworks did not fully utilized 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 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 long 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]. In hepatology, the declared fully interpretable model established by Saillard et al. was based on this approach. The interpretability of this model was relied on pathologist assessment of the image tiles that the network defined as the most significant associated with patient outcomes. Some features identified here (including macrotrabecular-massive subtype and cellular atypia ) were previously shown to be the predictors of dismal outcome, which endorsed the rationality of the model. Other feature, i.e. the presence of vascular spaces were also identified as indicators of poor survival. This elegant 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 leveraged on the similar strategy[15], and some histological features associated with 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 providing post hoc explanations. 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. However, CAM can only be applied to models with global average pooling layers, which limits its applicability. On the other hand, 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 wider range of models, including those without global average pooling layers. In summary, 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. modified CAM to risk activation mapping (RAM) in their study to visualize the pathological phenotypes of HCC associated with patient risk. The heatmaps generated through RAM indicated regions of the tissue that are potentially associated with increased risk 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, thus the researchers concluded that their deep learning framework could effectively decode pathological images and provide unnoticed biological information for HCC.

### 3.4 Feature extraction by attention mechanism

Attention mechanism is inspired by the biological systems of humans that tend to focus on the distinctive parts when processing large amounts of information. In the field of medical imaging, the attention layer of in a neural network could highlights where the model focuses on an image, and determines the proportion of attention paid to different areas of the image for classification[49]. 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[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.

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

### 3.5 Other promising approaches

There are many other approaches to explore the key features and novel biomarkers from AI-based pathological analysis, which have not been applied in liver cancer. Among them, perturbation-based approaches and textual explanation seem to be the most promising.

Perturbation-based approaches involve altering input images to assess the importance of specific areas for a given task. This strategy includes occlusion sensitivity map (OSM) and local interpretable model-agnostic explanations (LIME). OSM 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. 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.

Different from the visual explanations listed above, AI models using textual explanation could directly generate human-understandable semantics of their internal states. As an example, Testing with Concept Activation Vectors (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 feasibility of TCVA has been demonstrated in a medical image processing example, relating physician annotations like ‘microaneurysm’ to diabetic retinopathy in fundus imaging. 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.

These approaches might be applied in the future studies, and shed 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 strides, from diagnosis to prognostication and molecular profiling. 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.

However, the field is still in its infancy, with many challenges to overcome, such as standardization of image analysis and addressing limitations such as poor generalizability, low sensitivity to staining protocols and lack of prospective validation. Among all the challenges, the urge for interpretability will be increasingly significant, since it is the key to solving many other limitations. XXXXXX

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