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
2. 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. (JHEP Reports P 3-4) (JOH P1352). 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%. 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. 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”. (JHEP Reports P4) (JOH P1355). Shi et al. conducted a fine work to explore prognostic indicators in the pathological images of HCC via 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. 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.

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. 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 (WJH P2044)

1. 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. The 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. 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 (WJH P2044). 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. Exploring predictive indicators for therapy response

Recent pioneering studies have aimed to predict molecular signatures/alterations predictive of response to systemic therapies (JOH P1352).

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

(One paragraph highlighting the urgent need to explain the “black box” of deep learning)

1. Standardization of image analysis (JHEP p4)
2. 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. 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
4. **Model-based explanation**
5. 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. Then, a clustering analysis by an unsupervised learning approach identify three distinct histologic subtypes (MDPI). 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 (PeerJ). However, these authors did not provide further in-depth interpretation of the underlying biological implications of relevant features.

1. 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. 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, 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. However, ...

however,

1. Textual explanation
2. Image captioning
3. Image captioning with visual explanation
4. Example-based explanation

a. Triplet network

b. Prototypes

1. **Post hoc explanation**
2. Visual explanation (saliency mapping, pathologist-in-the-loop)
3. Backpropagation-based approaches

Including class activation mapping (CAM) and gradient-weighted class activation mapping (Grad-CAM)

CAM (Class Activation Mapping) and Grad-CAM (Gradient-weighted Class Activation Mapping) are both techniques used in computer vision to visualize and interpret the decision-making process of deep learning models. CAM highlights the important regions of an image by analyzing the weights of the final convolutional layer of the model. It generates a heatmap that shows which parts of the image contribute the most to the model's prediction. 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 in their study to visualize the pathological phenotypes 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.

**Chen et al.**

**Other studies used multi-modal approaches to rational the outputs from deep learning model. 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 .... Although this conclusion was based on a small number of patients and lack validation, it provides a unique ....**

1. Perturbation-based approaches

Including Occlusion sensitivity map (OSM), local interpretable model-agnostic explannations (LIME)

LIME was based on assumption that models behave linearly on a local scale. It aims to explain individual predictions by generating sparse linear models around an individual prediction in its local vicinity of inputs. In other words, LIME could tell, in a local sense, what is the most important attribute around the data point of interest.

1. Multiple instance learning-based approaches
2. Textual explanation

Testing with concept activation vectors (TCAV)

1. Example-based explanation
2. Influence functions

4. Conclusion and future applications.