Abstract #1240

AI(H): Deep Learning Model for Staging and Grading Autoimmune Hepatitis from Histology

Authors: Caner Ercan*#1, Kattayoun Kordy*2, Anna Knuuttila3, Xiaofei Zhou2, Darshan Kumar3, Peter Mesenbrink2, Serenella Eppenberger-Castori1, Luigi M Terracciano1, Marcos C. Pedrosa2

Affiliations: *Co-first Authorship; #Contact Author; 1Institute of Pathology and Medical Genetics, University Hospital Basel, University of Basel, Basel, Switzerland; 2Novartis Pharma AG, Basel, Switzerland; 3 Aiforia Technologies Plc, Helsinki, Finland

Video Abstract
tinyurl.com/AI4AIH

Introduction

The current diagnosis of Autoimmune Hepatitis (AIH) occurs using a combination of clinical-pathological criteria that includes histology, which is one of the most challenging diagnoses in liver pathology. A precise digital tool that facilitates AIH diagnosis would be helpful in the daily practice for both general pathologists and specialized hepato-pathologists to overcome diagnostic challenges. We aim to develop a deep learning model, Artificial Intelligence for Hepatitis [AI(H)] that classifies different regions of liver biopsies compared with hepato-pathologist reading, to provide granular, quantifiable and rapid analysis of histological features of AIH.

Background

AI(H) is an early machine-learning prototype which evaluates liver biopsies and accurately detects the disease-related features. It is a potential diagnostic tool composed of a set of newly developed customized convolutional neural networks (CNN) to predict various components of autoimmune hepatitis histology, including liver microanatomy structures, fibrosis, necroinflammation features, and immune cells.

Summary of key findings

AI(H) is a potential diagnostic tool for AIH histology. It demonstrates comparable results with potentially greater consistency to hepato-pathologist for several specific diagnostic tasks on AIH biopsies and performs the assignments over 100-fold faster than a human. Hepato-pathologist can improve their accuracy and consistency with the aid of this tool. Further planned development of AI(H) includes implementing other machine learning methods to predict AIH stage and grades based on the features extracted from CNN models.

Methods

One-hundred-thirteen pretreatment liver biopsies with confirmed AIH diagnosis from the biobank of the University Hospital Basel were selected and split into training (80%) and test (20%) datasets and later analyzed in the Aiforia platform (Aiforia Technologies Plc, Helsinki, Finland) using several convolutional neural network (CNN) models. CNNs are machine learning models developed from the structure of the visual cortex and have been used in many visual tasks. The liver microstructure detection model was trained to segment liver tissue into portal area, lobular area and central vein compartments, while the necroinflammation model was trained for focal necrosis, interface hepatitis and confluent necrosis. The immune cell classification model can detect, classify and quantify lymphocytes, plasma cells, macrophages, eosinophils and neutrophils. The AI models are evaluated on independent test dataset slides against hepato-pathologist's manual annotations to indicate how well they will generalize on future datasets.

Poster presented at the Liver Meeting® 2021 at the 72nd Annual Meeting of the American Association for the Study of Liver Diseases (AASLD) in Anaheim, CA, United States on 12-15 Nov 2021.

Contact: Caner Ercan (Caner.Ercan@usb.ch), University Hospital Basel, Basel, Switzerland

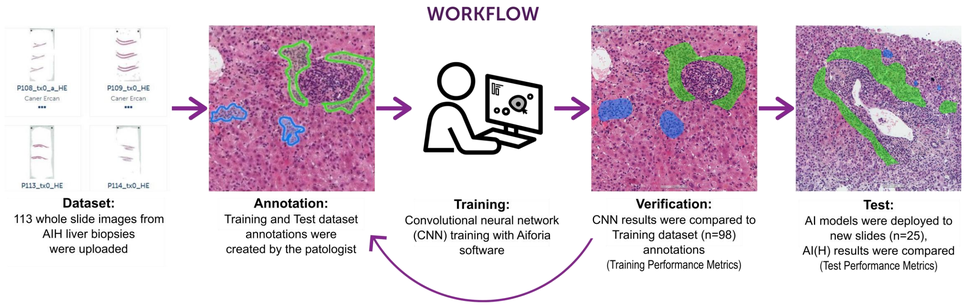


Table 1: AI(H) performance on inflammation-focused tasks, hematoxylin and eosin (H&E) stained slides

Model	Overall Accuracy[1]		Macro Precision[2]		Macro Sensitivity (Macro Recall)[2]	
	Test	Training	Test	Training	Test	Training
Tissue Detection	99.4%	99.9%	99.7%	98.9%	99.3%	99.5%
Microanatomy (Portal area, Lobular area, Central vein)	88.0%	97.5%	94.2%	98.3%	93.7%	96.6%
Necroinflammation (Focal necrosis, Interface hepatitis, Confluent necrosis, Pericentral necrosis, Bridging necrosis, Pan acinar necrosis)	83.9%	98.2%	49.7%	81.0%	37.2%	94.5%
Portal Inflammation	79.2%	78.5%	88.4%	99.7%	79.2%	79.9%
Immune Cells (Lymphocytes, Plasma cells, Macrophages, Eosinophils, Neutrophils)	72.4%	83.6%	86.9%	91.8%	85.2%	91.8%
Bile Duct Damage	81.7%	90.3%	91.3%	95.4%	90.3%	95.0%

Table 2: AI(H) performance on fibrosis-focused tasks, sirius red

Model	Overall Accuracy[1]		Macro Precision[2]		Macro Sensitivity (Macro Recall)[2]	
	Test	Training	Test	Training	Test	Training
Tissue Detection	99.4%	99.9%	99.8%	99.3%	99.0%	99.8%
Microanatomy (Portal area, Central vein)	94.0%	97.0%	67.0%	92.4%	65.9%	83.7%
Fibrosis (Portal fibrosis, Perivenular fibrosis, Pericellular fibrosis, Nodular fibrosis, Cirrhosis)	87.6%	97.2%	73.3%	96.2%	68.3%	95.1%

[1]Overall accuracy is a standalone metrics that measures how well machine learning models perform in multiclass classifications. It denotes the ratio of correct predications: for example, for a three category (category A, B and C) classification task, overall accuracy is calculated as the sum of correct predications on category A, B and C divided by the grand total.

[2]Precision and sensitivity (also called recall) are paired metrics (i.e., they cannot be used individually) that measure how well machine learning models perform in classification tasks. In binary classification, precision is calculated as TP/(TP+FP) and sensitivity is computed as TP/(TP+FN), where TP, FP, and FN are the number of true positives, false positives and false negatives, respectively. In multiclass classification, each category forms its own positive class and combines other categories as the negative class, thus, rendering several binary classifications. Macro precision and macro sensitivity are arithmetic mean (average) of individual binary precisions and of individual binary sensitivities, respectively.

Figure 1: AI(H) predictions on H&E stained slide

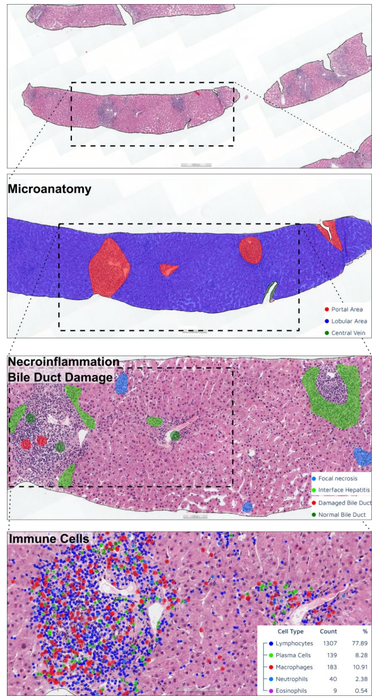
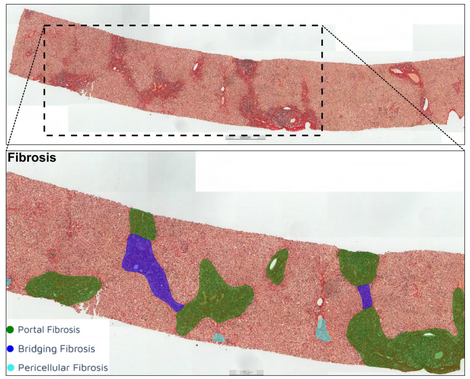


Figure 2: AI(H) predictions on sirius red stained slide



Results

The AI models are both accurate and efficient in predicting various morphological components of AIH biopsies. When evaluated on a separate test dataset, the models return high test accuracies (ratios of correct predictions): on hematoxylin and eosin stained slides, the test accuracies are 99.4%, 88.0%, 83.9%, 72.4%, 81.7% and 79.2% for tissue detection, liver microstructures, necroinflammation features, immune cell classification, bile duct damage detection, and portal inflammation feature, respectively **(Figure 1, Table 1)**; on sirius red stained slides, the test accuracies are 99.4%, 94.0%, and 87.6% for tissue detection, liver microstructures, and fibrosis detection, respectively **(Figure 2, Table 2)**. In addition, the capacity for cell counting of the model was far more efficient than a pathologist within the same time frame, with speed comparison of 121,387 cells/min vs. 97 cells/min.

Separate AI models were combined to obtain deeper insight of the AIH landscape. For example, the combination of the necroinflammation model with the bile duct model allows the user to observe the different features together **(Figure 1).** Furthermore, its combination with immune cells detection model provides a unique opportunity to obtain precise spatial information of the immune inflammation microenvironment of hepatitis (not shown here).

Limitations

- The sample size is relatively small for test dataset. The training and test datasets, biopsies and their digital slides are from the same institution.
- Both test and training annotations were made by a single hepato-pathologist.
- AIH liver biopsies with highly disrupted architecture or slides enriched with necroinflammation have a decline on the accuracy of the AI model predictions. While the AI(H) model is mostly accurate, there were mislabeled areas where fibrotic staging was high.

Conclusions

- The AI(H) tool shows highly accurate predictive abilities on several diagnostic tasks for AIH histology. Tissue detection was reached with an accuracy of 99.4 % on both hematoxylin and eosin as well as sirius red stained slides. The AI(H) test accuracies for liver microstructures was 88.0% and 94% depending on the slide staining and other characteristics were predicted with a test accuracy of 72.4% to 87.6%. On sirius red stained slides, the tool has test accuracies of 94.0%, and 87.6% for liver microstructures, and fibrosis detection, respectively.
- The AI(H) method is consistent in predicting AIH components, such as portal lymphoplasmacytic infiltrate, interface hepatitis, lobular activity, and fibrosis, and is highly efficient in classifying and counting cells/tissues over 100-fold faster than a human. Different from fully "black box" models, our AI(H) model can provide intermediate results interpretable by pathologists. Additionally, classical advantages of computational methods (e.g., reliability, consistency and speed) make AI(H) a promising computational pathology tool to facilitate histological recognition.
- Future studies with larger datasets from different biobanks or cohorts are needed to further refine the model and validate findings; as well as expand the number of hepato-pathologist readers to evaluate potential performance differences. We plan on developing a model that combines the various AIH components to report Ishak grading/staging, as well as the AIH pathological diagnosis scores. Since AI(H) is capable of recognizing the elementary lesions, it can be easily translated for evaluating the other chronic hepatitis biopsies, such as viral hepatitis.
- The AI(H) algorithm is automated in predicting AIH histology components and can save pathologists a significant amount of time if utilized as an early step in an integrated workflow: once the biopsy slides are scanned, AI(H) can be first applied to identify the biopsies more likely associated with AIH. This allows pathologists to focus more on the biopsies that require special attention.
- We look forward to seeing a near future where pathologists further incorporate artificial intelligence tools in their daily routine and improve the histologic diagnosis and care of AIH patients.

References

1. LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324..
2. Fukushima, Kunihiko, and Sei Miyake. "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition." *Competition and cooperation in neural nets*. Springer, Berlin, Heidelberg, 1982. 267-285.
3. McCulloch, Warren S., and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity." *The bulletin of mathematical biophysics* 5.4 (1943): 115-133..
4. LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1.4 (1989): 541-551..
5. Jones, David, et al. "Unmet needs and new models for future trials in autoimmune hepatitis." *The Lancet Gastroenterology & Hepatology* 3.5 (2018): 363-370.
6. Manns, Michael P., et al. "Diagnosis and management of autoimmune hepatitis." *Hepatology* 51.6 (2010): 2193-2213.

Acknowledgements

We would like to thank Nora Holmberg, Kristiina Tarkiainen, Olivia Tapaninen, Christine Relander, and Sami Blom from Aiforia Technologies Plc for their assistance.

Disclosures

MP, KK, XZ, PM are employees and shareholders of Novartis Pharmaceuticals. DK, AK are from Aiforia Technologies Plc. Study was partially sponsored by Novartis.