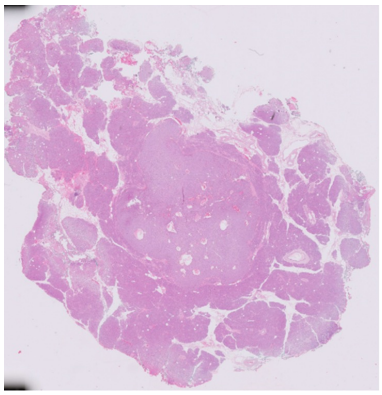
Digital pathology

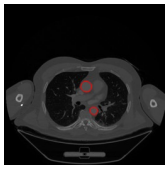
- Now become an integral part of pathology workflow
  - Rapid progress of whole slide imaging technology
  - Advances in high-speed networking
  - Increasing resources for data processing and storage
- Offers very promising solutions for slide acquisition, management, remote access, and image viewing
- Allows using *"smart"* computational tools for quantitative analysis of the slides

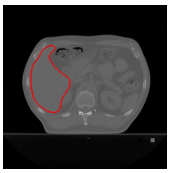


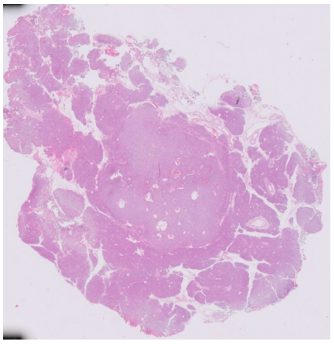
1

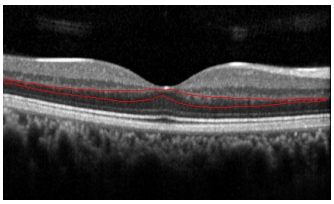
Deep learning in digital pathology

- Deep learning facilitates implementing promising computational tools
  - Segmentation models that localize regions of interest (cells, glands, tissue compartments, etc.)
  - Classification models that characterize the regions of interest
  - Correlation tools that associate slide characteristics with clinical facts and diagnoses
- However, whole slide images are much more complex to interpret than medical images in other fields
- As a result, digital transformation of pathology is happening at a much slower pace







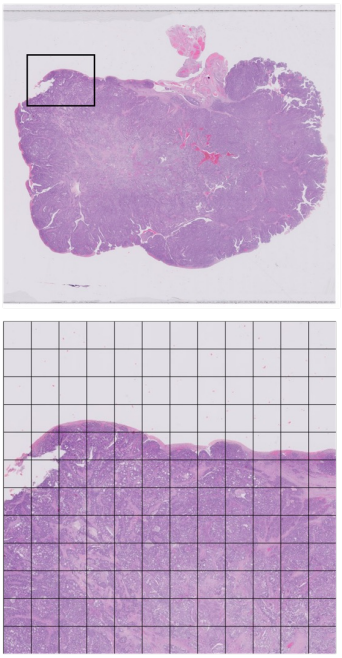


2

Challenges in digital pathology

Interpreting a whole slide at the pathologist level requires unique challenges to overcome

1. A whole slide is converted into a huge collection of high-resolution image patches
   - This huge collection should be stored and processed
   - Its processing requires high processor and memory resources
   - Standard classification and segmentation networks can process only one image patch at a time (at the time of training and testing), not the whole slide



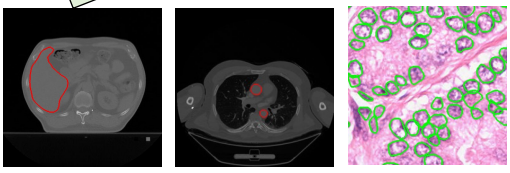
3

Challenges in digital pathology

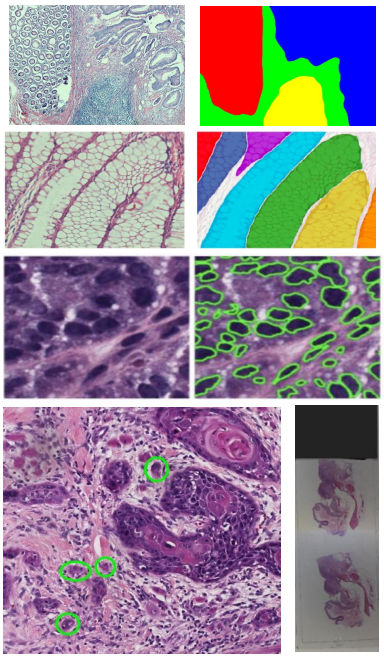
Interpreting a whole slide at the pathologist level re
unique challenges to overcome

3. This process involves different
   - The model should know
   - Patch-level, cel
   - Differ pretations
   - different problem, and mostly
   i ized solution

All these challenges offer unique opportunities to work on a wide variety of scientifically important research problems



Some of these problems may need to handle multiple instances/objects



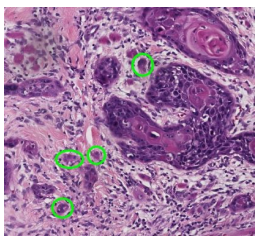
4

Challenges in digital pathology

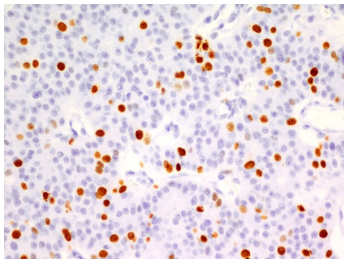
… on the other hand, the most important challenge is

4. To train and validate deep learning … limited annotated dataset
   - Slide-level labels are … *pervised* setting
   - Hard to obt… …ns for a *strongly sup…* …l, ROI level, or pixel-level)
     - …bel is assigned to every patch in a WSI, models … from noisy training samples
     - ➢ …ere might be inconsistencies in the strong labels
     - ➢ Some pixels/ROIs are hard-to-annotate due to the intrinsic characteristics of images, which causes inaccurate labels
   - Data sharing between different institutions may not be allowed

Deep learning models should be designed and trained considering these limitations



Marking all positives may not be possible or requires too much effort (false negatives may exist)



Cell boundary annotations vs approximate cell locations vs eye-balling count

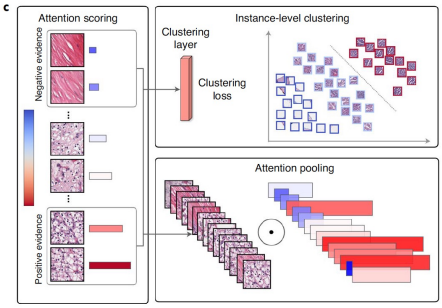
5

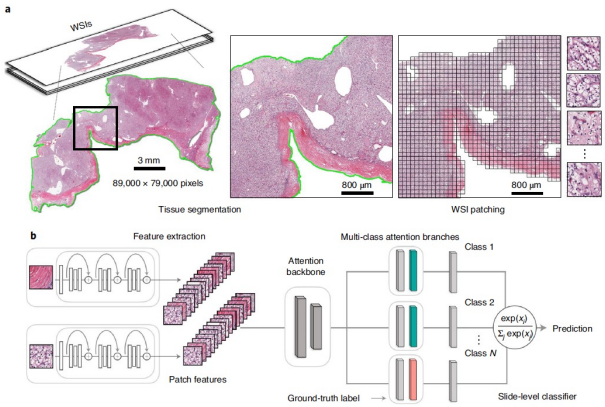
nature biomedical engineering
ARTICLES
https://doi.org/10.1038/s41551-020-00682-w

Data-efficient and weakly supervised computational pathology on whole-slide images

Ming Y. Lu[1,2,3], Drew F. K. Williamson[1,5], Tiffany Y. Chen[1,5], Richard J. Chen[1,4], Matteo Barbieri[1,2] and Faisal Mahmood[1,2,3]

Deep-learning methods for computational pathology require either manual annotation of gigapixel whole-slide images (WSIs) or large datasets of WSIs with slide-level labels and typically suffer from poor domain adaptation and interpretability. Here we report an interpretable weakly supervised deep-learning method for data-efficient WSI processing and learning that only requires slide-level labels. The method, which we named clustering-constrained-attention multiple-instance learning (CLAM), uses attention-based learning to identify subregions of high diagnostic value to accurately classify whole slides and instance-level clustering over the identified representative regions to constrain and refine the feature space. By applying CLAM to the subtyping of renal cell carcinoma and non-small-cell lung cancer as well as the detection of lymph node metastasis, we show that it can be used to localize well-known morphological features on WSIs without the need for spatial labels, that it overperforms standard weakly supervised classification algorithms and that it is adaptable to independent test cohorts, smartphone microscopy and varying tissue content.





Smooth SVM loss. For the instance-level clustering task, we chose to use the smooth top-1 SVM loss[44], which is based on the well-established multi-class SVM loss[45]. In a general N-class classification problem, neural network models output a

the rest of the model using the slide-level labels as no ground-truth attention is available. The total loss for a given slide $\mathcal{L}_{\text{total}}$ is the sum of both the slide-level classification loss $\mathcal{L}_{\text{slide}}$ and the instance-level clustering loss $\mathcal{L}_{\text{patch}}$ with optional scaling via scalar c_1 and c_2:

$$\mathcal{L}_{\text{total}} = c_1\mathcal{L}_{\text{slide}} + c_2\mathcal{L}_{\text{patch}} \quad (6)$$

6

MS-CLAM: Mixed Supervision for the classification and localization of tumors in Whole Slide Images

Paul Tourniaire[a,*], Marius Ilie[b,c,d], Paul Hofman[b,c,d], Nicholas Ayache[a,d], Hervé Delingette[a,d]

ABSTRACT

Given the size of digitized Whole Slide Images (WSIs), it is generally laborious and time-consuming for pathologists to exhaustively delineate objects within them, especially with datasets containing hundreds of slides to annotate. Most of the time, only slide-level labels are available, giving rise to the development of weakly-supervised models. However, it is often difficult to obtain from such models accurate object localization, e.g., patches with tumor cells in a tumor detection task, as they are mainly designed for slide-level classification. Using the attention-based deep Multiple Instance Learning (MIL) model as our base weakly-supervised model, we propose to use mixed supervision – i.e., the use of both slide-level and patch-level labels – to improve t the classification and the localization performances of the original model, using on limited amount of patch-level labeled slides. In addition, we propose an attention term to regularize the attention between key instances, and a paired batch method to ate balanced batches for the model. First, we show that the changes made to the m already improve its performance and interpretability in the weakly-supervised sett Furthermore, when using only between 12 and 62% of the total available patch-l annotations, we can reach performance close to fully-supervised models on the tu classification datasets DigestPath2019 and Camelyon16.

- To improve the performance of the tile-level classifier, it is trained on both true tile labels (when available) and pseudo-labels generated using the tiles' attention scores. This allows the classifier to leverage more training samples with accurate labels. To correct the potential class imbalance between tumorous and non-tumorous tiles, we also propose a paired batch method that uses both kinds of slides at the same time at each training step.

forming opposite tasks would be redundant. Second, for each slide with available tile-level labels, the instances are sampled and assigned their true label instead of the pseudo-generated one (see Figure 1, top). This not only allows us to train the tile-level classifier without potentially erroneous labels, it also helps sampling more tiles within the slides, since for all of them the label is accessible. To distinguish between the cases where tile labels are known or not, we use two different hyperparameters to sample the instances: B_+ when labels are available, and B_- when they are not ($B_+ > B_-$). For the case where B_+ might be greater than the actual number of tumorous tiles N_{tum} in the slide, we set $B_+ = N_{tum}$. Third, since in normal slides all tiles are normal (thanks to the MIL assumption), we use a different tile-sampling strategy, as the original method assigned wrong labels to the tiles with low attention scores. In our case, in normal slides, sampled tiles are only assigned the same label as the slide. Moreover, if we sample B tumorous tiles in tu-

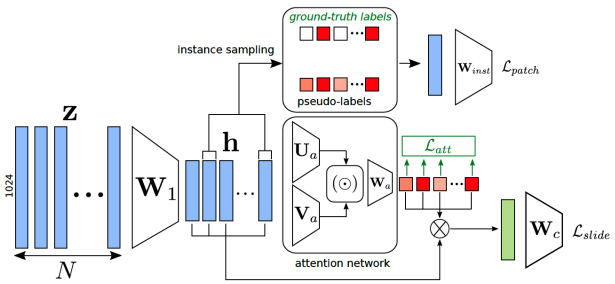


Fig. 1. Overview of the MS-CLAM model. Regarding the attention scores, a faint (resp. bright) color represents a low (resp. high) score. For the ground-truth colors, red means the instance is positive (with respect to the bag label), while no color means the instance is negative. The light-green rectangle represents the attention-weighted average of the feature vectors. Our contributions to the original CLAM architecture are printed in dark green.

7

Hybrid Supervision Learning for Pathology Whole Slide Image Classification

Jiahui Li[1], Wen Chen[1], Xiaodi Huang[1], Shuang Yang[1], Zhiqiang Hu[1], Qi Duan[1], Dimitris N. Metaxas[2], Hongsheng Li[3,4], and Shaoting Zhang[1,4⋆]

[1] SenseTime Research, Shanghai, China
{lijiahui,chenwen,huangxiaodi,yangshuang1,huzhiqiang,duanqi,zhangshaoti
[2] Rutgers University, New Jersey, USA
dnm@cs.rutgers.edu
[3] The Chinese University of Hong Kong, Hong Kong, China
hsli@ee.cuhk.edu.hk
[4] Centre for Perceptual and Interactive Intelligence (CPII) Ltd, Hong Kong, Chir

Abstract. Weak supervision learning on classification labels has demonstrated high performance in various tasks, while a few pixel-level fine annotations are also affordable. Naturally a question comes to us that whether the combination of pixel-level (e.g., segmentation) and image level (e.g., classification) annotation can introduce further improvement. However in computational pathology this is a difficult task for this reason: High resolution of whole slide images makes it difficult to do end-to-end classification model training, which is challenging to research of weak or hybrid supervision learning in the past. To handle this problem, we

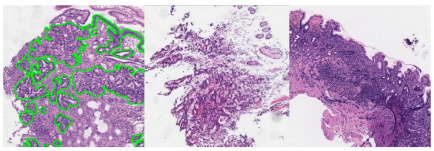


Fig. 1. Illustration of hybrid supervision data, containing 2 types of label. The first expensive and rare type is pixel-level fine-grained labels, contoured by green lines. The second type is image-level labeled images. The rest two images are image-level positive and negative.

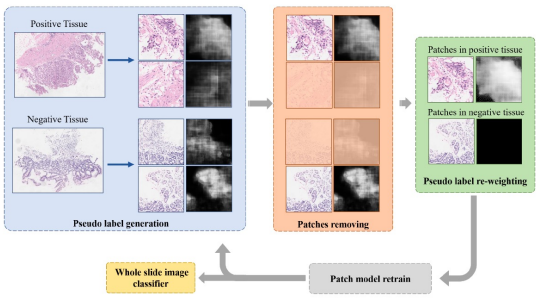


Fig. 2. The overall pipeline of our hybrid supervision learning for whole slide image. The image-level label guided pixel-level pseudo label generation is iterated in this manner. The gray scale map is the predicted probability map of patch model.

8

while *the model do not converge* **do**

Stage 1: patch segmentation ;

E-step: $\hat{y_p} \leftarrow P(y_p|\mathrm{x}, \theta_1)$;

Remove those patch x whose maximum pixel-level $\hat{y_p}$ in this patch is less than T ;

if $y_i == 1$ **then**

$\hat{y_{p+}} \leftarrow \hat{y_p} \times \mathrm{V}$ (pseudo labels);

$\hat{y_{p+}} \leftarrow 1.0$ if $\hat{y_{p+}} > 1.0$ (clip within 1.0, to remain true positive);

$\hat{y_{p+}} \leftarrow 0.0$ if $\hat{y_{p+}} < 0.01$ (clip to 0.0 if lower than 0.01, to remain true negative which was slightly scaled up by V);

else

$\hat{y_{p-}} \leftarrow 0$ (hard negative labels);

M-step: Retrain patch segmentation model θ_1 at proper sampling ratio of R in each training batch, and by pixel-level soft label cross entropy loss $= -\frac{1}{N} \sum_N \sum_{y=y_p \cup \hat{y_{p+}} \cup \hat{y_{p-}}}$

$y \times logP(y_p|x, \theta_1) + (1.0 - y) \times log(1.0 - P(y_p|x, \theta_1))$;

Stage 2: whole slide image classification ;

Select top K patches for each whole slide image according to pixel-level maximum $P(y_p|x, \theta_1)$;

$P(y_i|I, \theta_2) = \frac{1}{K} \sum_K P(y_i|x, \theta_2)$

Train classification model θ_2 by Loss $=$

$-\frac{1}{N} \sum_N y_i \times logP(y_i|I, \theta_2) + (1.0 - y_i) \times log(1.0 - P(y_i|I, \theta_2))$

Convergence criteria: For each round of stage1, perform stage2 training. Select the round whose stage2 training loss is the lowest, which means the top K patches selected in that round by stage1 is the optimal to fit image-level annotations.

To implement our intuition, shown in Fig. 2, we separate large size of whole slide images into patches, then develop an *Expectation-Maximization (EM)*-like method to make full use of three types of annotation: image-level labels y_i, pixel-level fine-grained labels y_p and pixel-level pseudo labels $\hat{y_p}$. In the E-step, pixel-level pseudo labels $\hat{y_p}$ are firstly created from segmentation confidence map of patches from both positive and negative images. We remove all the patches

In the M-step, patch segmentation model is then trained on a sampling ratio of pixel-level fine-grained labels y_p, pseudo labeled positive patches $\hat{y_{p+}}$ and hard negative patches $\hat{y_{p-}}$. For sampling ratio of y_p, $\hat{y_{p+}}$ and $\hat{y_{p-}}$ in each training batch, hard negative patches $\hat{y_{p-}}$ shall be much more than pseudo labeled positive patches $\hat{y_{p+}}$ so that if one pattern is both labeled as negative in hard negative

possible to be eventually discriminated as positive by models. Such procedure is iterated for several rounds and we evaluate sensitivity and specificity for each to decide when to stop. Loss function in patch segmentation stage is pixel-level soft-label cross entropy loss to deal with both soft pseudo and fine-grained labels.

During whole slide image classification stage, for each super size whole slide image, top K patches with maximum pixel-level probability are the input to image classifier θ_2. Average probability is the final image-level confidence to calculate loss with image-level labels. Stage2 training also decides the convergence criteria. We perform stage2 training for each round of stage1 and select that round whose stage2 training loss is the lowest. That means the top K patches selected in this round by stage1 is the optimal to fit image-level annotations in stage2.

9

Segmentation with mixed supervision: Confidence maximization helps knowledge distillation

Bingyuan Liu [a,*], Christian Desrosiers [a], Ismail Ben Ayed [a,b], Jose Dolz [a,b,**]

[a] *ÉTS Montréal, Canada*

[b] *Centre de recherche du Centre hospitalier de l'Université de Montréal (CRCHUM), Canada*

ARTICLE INFO

Keywords:
CNN
Image segmentation
Mixed-supervision
Semi supervision

ABSTRACT

Despite achieving promising results in a breadth of medical image segmentation tasks, deep neural networks (DNNs) require large training datasets with pixel-wise annotations. Obtaining these curated datasets is a cumbersome process which limits the applicability of DNNs in scenarios where annotated images are scarce. Mixed supervision is an appealing alternative for mitigating this obstacle. In this setting, only a small fraction of the data contains complete pixel-wise annotations and other images have a weaker form of supervision, e.g., only a handful of pixels are labeled. In this work, we propose a dual-branch architecture, where the upper branch (teacher) receives strong annotations, while the bottom one (student) is driven by limited supervision and guided by the upper branch. Combined with a standard cross-entropy loss over the labeled pixels, our novel formulation integrates two important terms: (i) a Shannon entropy loss defined over the less-supervised images, which encourages confident student predictions in the bottom branch; and (ii) a Kullback–Leibler (KL) divergence term, which transfers the knowledge (i.e., predictions) of the strongly supervised branch to the less-supervised branch and guides the entropy (student-confidence) term to avoid trivial solutions. We show that the synergy between the entropy and KL divergence yields substantial improvements in performance. We also discuss an interesting link between Shannon-entropy minimization and standard pseudo-mask generation, and argue that the former should be preferred over the latter for leveraging information from unlabeled pixels. We evaluate the effectiveness of the proposed formulation through a series of quantitative and qualitative experiments using two publicly available datasets. Results demonstrate that our method significantly outperforms other strategies for semantic segmentation within a mixed-supervision framework, as well as recent semi-supervised approaches. Moreover, in line with recent observations in classification, we show that the branch trained with reduced supervision and guided by the top branch largely outperforms the latter. Our code is publicly available: https://github.com/by-liu/ConfKD.

10

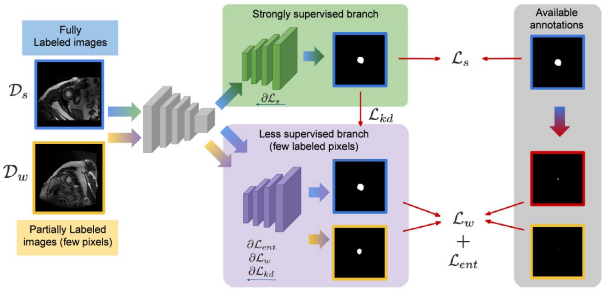


Fig. 1. Overview of the proposed method. **Training:** Both fully and partial labeled images are fed to the network. The top branch generates predictions for fully labeled images, whereas the bottom branch generates the outputs for partially labeled images. Furthermore, the bottom branch also generates segmentations for the fully labeled images, which are guided by the KL term between the two branches. **Inference:** Once the model is trained, we can remove the strongly supervised branch (*Top*), and get the final segmentation result from the bottom stream. The gradients for each loss term are highlighted in the figure. Note that, similarly to UNet-like architectures, the proposed model has skip connections between the encoder and the decoders.

3.4. Distilling strong knowledge

In addition to the specific supervision available at each branch, we transfer the knowledge from the teacher (top branch) to the student (bottom branch). This is done by forcing the softmax distributions from the bottom branch to mimic the probability predictions generated by the top branch for the fully labeled images in D_s. This knowledge-distillation regularizer takes the form of a Kullback–Leibler divergence (D_{KL}) between both distributions:

$$\mathcal{L}_{kd} = \sum_{i=1}^{m} \sum_{l \in \Omega_i} D_{\mathrm{KL}} \left((\mathbf{p}_l^i)_{\mathrm{top}} \,\|\, (\mathbf{p}_l^i)_{\mathrm{bottom}} \right) \tag{3}$$

where $D_{\mathrm{KL}}(\mathbf{p} \,\|\, \mathbf{q}) = \mathbf{p}^T \log \frac{\mathbf{p}}{\mathbf{q}}$, with T denoting the transpose operator.

3.2. Supervised learning

The top-branch is trained under the fully-supervised paradigm, where a set of training images containing pixel-level annotations for all the pixels is given, i.e., D_s. The problem amounts to minimizing with respect to the network parameters a standard full-supervision loss, which typically takes the form of a cross-entropy:

$$\mathcal{L}_s = -\sum_{i=1}^{m} \sum_{l \in \Omega_i} (\mathbf{y}_l^i)^T \log (\mathbf{p}_l^i)_{\mathrm{top}} \tag{1}$$

where vector $\mathbf{y}_l^i = \left(y_l^{i,1}, \ldots, y_l^{i,C}\right) \in \{0,1\}^C$ describes the ground-truth annotation for pixel $l \in \Omega_i$. Here, notation $(\cdot)_{\mathrm{top}}$ refers to the softmax outputs of the *top* branch of the network. Note that all the losses are normalized by the cardinality of the training dataset, which has been omitted to simplify the notation.

3.3. Not so-supervised branch

We consider the scenario where only the labels for a handful of pixels are known, i.e., scribbles or points. Particularly, we use the dataset D_w whose pixel-level labels are partially provided. Furthermore, for each image on the labeled training set, D_s, we generate partially supervised labels (more details in the experiments' section), which are added to augment the dataset D_w. Then, for the partially-labeled set of pixels, denoted as $\Omega_i^{\mathrm{partial}}$ for each image $i \in \{1, \ldots n\}$, we can resort to the following partial-supervision loss, which takes the form of a cross-entropy on the fraction of labeled pixels:

$$\mathcal{L}_w = -\sum_{i=m+1}^{n} \sum_{l \in \Omega_i^{\mathrm{partial}}} (\mathbf{y}_l^i)^T \log (\mathbf{p}_l^i)_{\mathrm{bottom}} \tag{2}$$

where notation $(\cdot)_{\mathrm{bottom}}$ refers to the softmax outputs of the *bottom* branch of the network.

11

nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-020-0186-1

Check for updates

Secure, privacy-preserving and federated machine learning in medical imaging

Georgios A. Kaissis[1,2,3], Marcus R. Makowski[1], Daniel Rückert[2] and Rickmer F. Braren[1]✉

The broad application of artificial intelligence techniques in medicine is currently hindered by limited dataset availability for algorithm training and validation, due to the absence of standardized electronic medical records, and strict legal and ethical requirements to protect patient privacy. In medical imaging, harmonized data exchange formats such as Digital Imaging and Communication in Medicine and electronic data storage are the standard, partially addressing the first issue, but the requirements for privacy preservation are equally strict. To prevent patient privacy compromise while promoting scientific research on large datasets that aims to improve patient care, the implementation of technical solutions to simultaneously address the demands for data protection and utilization is mandatory. Here we present an overview of current and next-generation methods for federated, secure and privacy-preserving artificial intelligence with a focus on medical imaging applications, alongside potential attack vectors and future prospects in medical imaging and beyond.

Decentralized data and federated machine learning

The concept of federated machine learning began gathering significant attention around the year 2015[46]. It belongs to a class of decentralized/distributed systems that rely on the principle of remote execution—that is, distributing copies of a machine learning algorithm to the sites or devices where the data is kept (nodes), performing training iterations locally, and returning the results of the computation (for example, updated neural network weights) to a central repository to update the main algorithm. Its main benefit is the ability of the data to remain with its owner (retention of sovereignty), while still enabling the training of algorithms on the data. The federation topology is flexible (model sharing among the nodes and aggregation at a later time (peer to peer/gossip strategy[47]) or full decentralization, combined, for example, with contribution tracking/audit trails using blockchains[48]). Continuous online availability is not required since training can be performed offline and results returned later. Thus, federated learning approaches have arguably become the most widely used next-generation privacy preservation technique, both in industry[49] and medical AI applications[50].

12

Federated Learning for Computational Pathology on Gigapixel Whole Slide Images

Ming Y. Lu[a,c], Dehan Kong[a], Jana Lipkova[a,c], Richard J. Chen[a,b,c], Rajendra Singh[e], Drew F. K. Williamson[a,c], Tiffany Y. Chen[a,c], Faisal Mahmood[*a,c,d]

[a]*Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA*
[b]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA*
[c]*Cancer Program, Broad Institute of Harvard and MIT, Cambridge, MA*
[d]*Cancer Data Science, Dana-Farber Cancer Institute, Boston, MA*
[e]*Department of Pathology, Northwell Health, NY*

Abstract

Deep Learning-based computational pathology algorithms have demonstrated profound ability to excel in a wide array of tasks that range from characterization of well known morphological phenotypes to predicting non human-identifiable features from histology such as molecular alterations. However, the development of robust, adaptable and accurate deep learning-based models often rely on the collection and time-costly curation large high-quality annotated training data that should ideally come from diverse sources and patient populations to cater for the heterogeneity that exists in such datasets. Multi-centric and collaborative integration of medical data across multiple institutions can naturally help overcome this challenge and boost the model performance but is limited by privacy concerns amongst other difficulties that may arise in the complex data sharing process as models scale towards using hundreds of thousands of gigapixel whole slide images. In this paper, we introduce privacy preserving federated learning for gigapixel whole slide images in computational pathology using weakly-supervised attention multiple instance learning and differential privacy. We evaluated our approach on two different diagnostic problems using thousands of histology whole slide images with only slide-level labels. Additionally, we present a weakly-supervised learning framework for survival prediction and patient stratification from whole slide images and demonstrate its effectiveness in a federated setting. Our results show that using federated learning, we can effectively develop accurate weakly supervised deep learning models from distributed data silos without direct data sharing and its associated complexities, while also preserving differential privacy using randomized noise generation.

Herin, we present the **key contributions** of our work as follows:

1. We present the first large-scale computational pathology study to demonstrate the feasibility and effectiveness of privacy-preserving federated learning using thousands of gigapixel whole slide images from multiple institutions.
2. We account for the challenges associated with the lack of detailed annotations in most real world whole slide histopathology datasets and are the first to demonstrate how federated learning can be coupled with weakly-supervised multiple instance learning to perform both binary and multi-class classification problems (demonstrated on breast cancer and renal cell cancer histological subtyping) using only slide-level labels for supervision.
3. We extend the usage of attention-based pooling in multiple instance learning-based classification and present a weakly-supervised framework for survival prediction (demonstrated on renal cell carcinoma patients) in computational pathology using whole slide images and patient-level prognostic information, without requiring manual ROI-selection or randomly sampling a predetermined number of patches.

13

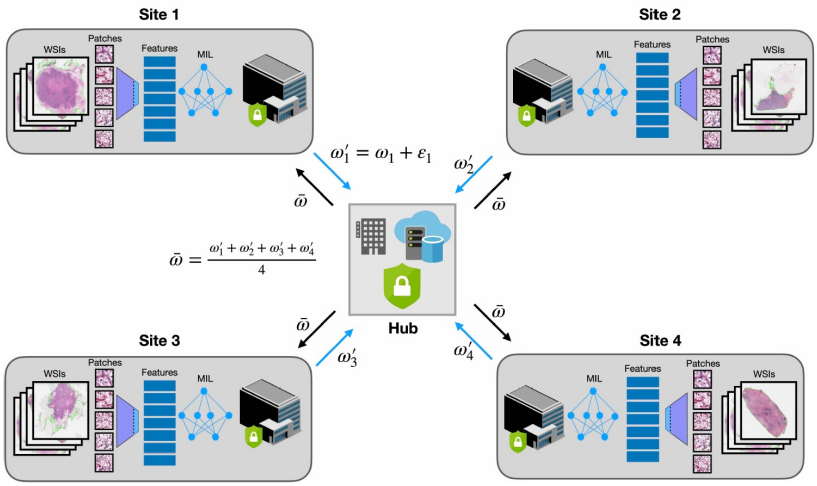


Figure 1: **Overview of the weakly supervised multiple instance learning in a federated learning framework.** At each client site, for each WSI, the tissue regions are first automatically segmented and image patches are extracted from the segmented foreground regions. Then all patches are embedded into a low-dimension feature representation using a pretrained CNN as the encoder. Each client site trains a model using weakly-supervised learning on local data (requires only the slide-level or patient-level labels) and sends the model weights each epoch to a central server. Random noise can be added to the weight parameters before communicating with the central hub for differential privacy preservation. On the central server, the global model is updated by averaging the model weights retrieved from all client sites. After the federated averaging, the updated weights of the global model is then sent to each client model for synchronization prior to starting the next federated round.

14

content may change prior to final publication. Citation information: DOI 10.1109/TMI.2022.3233405

FedMix: Mixed Supervised Federated Learning for Medical Image Segmentation

Jeffry Wicaksana, Zengqiang Yan, *Member, IEEE*, Dong Zhang, Xijie Huang, Huimin Wu, Xin Yang, *Member, IEEE*, and Kwang-Ting Cheng, *Fellow, IEEE*

***Abstract*— The purpose of federated learning is to enable multiple clients to jointly train a machine learning model without sharing data. However, the existing methods for training an image segmentation model have been based on an unrealistic assumption that the training set for each local client is annotated in a similar fashion and thus follows the same image supervision level. To relax this assumption, in this work, we propose a label-agnostic unified federated learning framework, named FedMix, for medical image segmentation based on mixed image labels. In FedMix, each client updates the federated model by integrating and effectively making use of all available labeled data ranging from strong pixel-level labels, weak bounding box labels, to weakest image-level class labels. Based on these local models, we further propose an adaptive weight assignment procedure across local clients, where each client learns an aggregation weight during the global model update. Compared to the existing methods, FedMix not only breaks through the constraint of a single**

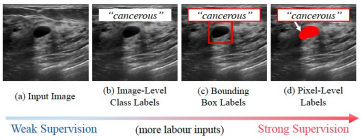


Fig. 1: Examples of different levels of medical image labels, where the image-level class label in (b) contains only the lesion category. The bounding box label in (c) contains not only the lesion category but also a coarse location. The pixel-level label in (d) contains both the lesion category and location information of each pixel, which is strong image supervision. Though strong image supervisions are more informative, they are very expensive to obtain. The utilization of some easy-to-access image supervision is beneficial in practice.

The optimization of deep learning models usually relies on a vast amount of training data [5]. For example, for a fully-supervised semantic segmentation model, the ideal scenario is that we can collect the pixel-level annotated images as much as possible from diverse sources. However, this scenario is almost infeasible due to the following two reasons: 1) the strict sharing protocol of sensitive patient information between medical institutions and 2) the exceedingly high pixel-level annotation cost. As the expert knowledge usually required for annotating medical images is much more demanding and difficult to obtain, various medical institutions have very limited strong pixel-level annotated images and most available images are unlabeled or weakly-annotated [3], [7], [21], [22], [27]. Therefore, a realistic clinical mechanism that utilizes every available supervision for cross-institutional collaboration without data sharing is highly desirable.

[18], due to the obvious reason that medical images often contain some personal information. During the training process of a standard FL model, each local client first downloads the federated model from a server and updates the model locally. Then, the locally-trained model parameters of each client are sent back to the server. Finally, all clients' model parameters are aggregated to update the global federated model. Most of the existing FL frameworks [14], [19] require that the data used for training by each local client needs to follow the same level of labels, *e.g.*, pixel-level labels (as shown

15

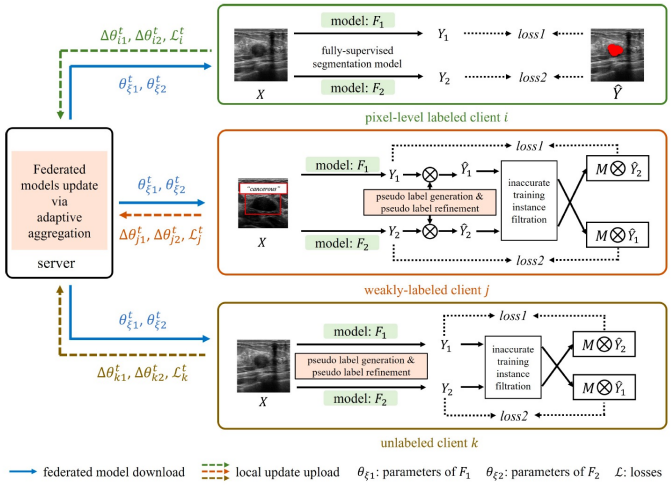


Fig. 2: Illustration of the proposed Mixed Supervised Federated Learning (FedMix) framework. The local client update utilizes every available supervision for training. Based on this, an adaptive weight aggregation procedure is used for the global federated model update. Compared to existing methods, FedMix not only breaks through the constraint of a single level of image supervision but also can dynamically adjust the aggregation weight of each local client, achieving rich yet discriminative feature representations.

1) **Pseudo Label Generation and Selection** ***(Sample & Refine).*** To utilize *every available data* and ensure reliable local model updates from clients without pixel-level labels, we design a novel unified framework using *every level of label* to amplify and filter useful signals from pseudo supervision. Specifically, FedMix utilizes consistency regularization [39] to generate pseudo labels which are then dynamically filtered and refined before being used for training.
2) **Adaptive Aggregation for Federated Model Update** ***(Aggregate).*** FedMix presents a novel adaptive aggregation operation to alleviate the training instability which may arise from naively aggregating local model updates from weakly-labeled and unlabeled clients. By taking consistency regularization and dynamic sample selection into account, the weight of each client is determined according to its data quantity and quality (inferred from training loss). In this way, more reliable clients will be assigned with higher weights, leading to better convergence.

16