COMPUTATIONAL HISTOPATHOLOGY

No pixel-level annotations needed

A deep-learning model for cancer detection trained on a large number of scanned pathology slides and associated diagnosis labels enables model development without the need for pixel-level annotations.

Jeroen van der Laak, Francesco Ciompi and Geert Litjens

Recent developments in machine learning have resulted in algorithms that perform certain complex (yet narrow) tasks at levels that are comparable (or exceeding) those of experts. One such type of task relates to anatomic pathology: deep learning is suitable for the detection and classification of disease in scanned microscopic tissue sections (whole slide images; WSIs)[1]. The development and training of deep-learning algorithms for histopathology generally requires a large amount of WSIs that encompass the spectrum of patterns typical of different tissue components. Moreover, to construct algorithms that generalize well to new situations, it is advisable to include training data with as much technical variability as possible (such as variabilities caused by tissue processing and staining, and by slide scanning). The construction of deep-learning models that generalize well to different clinical situations is hampered by the need for strongly labelled WSIs. These consist of scanned tissue sections in which experienced human observers have visually delineated a large number of areas containing different tissue components. For instance, the 400 WSIs of sentinel lymph nodes of patients with breast cancer used in the CAMELYON16 challenge included exhaustive annotations defining tissue regions with metastases[2]. Yet having trained pathologists annotate large numbers of cases is often infeasible (Fig. 1). Hence, most published work has described promising results on the basis of datasets of limited size, which has prevented the deep-learning models from reaching clinical implementation. Reporting in *Nature Medicine*, Thomas Fuchs and colleagues now show that training deep-learning models with a multiple-instance-learning (MIL) approach (a form of weakly supervised learning) by using more than 10,000 cases of weakly labelled WSIs — that is, labels describing only the presence or absence of disease in the WSI, rather than the location and extent of disease — collected from multiple clinical centres obviates the need for strongly labelled images in certain applications[3].

Fuchs and co-authors focused on three applications: the detection of adenocarcinoma in prostate biopsies, of basal cell carcinoma in biopsies and excisions of neoplastic and non-neoplastic skin lesions, and of breast-cancer metastases in axillary lymph nodes. The authors show that, for these applications, the MIL approach achieves excellent performance (an area under the receiver operating characteristic curve (AUC) larger than 0.98 for the three applications), provided that a sufficient amount of weakly labelled images is available (earlier work on MIL for WSIs reached a performance that was clearly inferior to the use of strongly labelled images[4]). They conclude that more than 10,000 WSIs are typically required to reach performances comparable to those of deep-learning models trained with strongly labelled WSIs.

Because of the variation already present in Fuchs and co-authors' WSI dataset, they did not use data augmentation — a widely used technique for artificially increasing variability in the training data that strongly increases generalizability[5]. However, when training the deep-learning model by using a large number of WSIs from a single centre, the model did not generalize well: the AUC dropped by 6% when the model was trained on images from one single centre rather than images from many centres, and by 3% when the model trained on data produced by one type of scanner was applied to images produced by another type of scanner. The authors show that their approach generalizes better than a model trained on a small set of fully annotated slides, yet they do not address how this compares to simpler strategies (such as colour augmentation or normalization) for improving generalization. Although the inclusion of slides from multiple centres relieves this problem, it is unclear whether the model would work on WSIs from centres not included in the training dataset. In this respect, the MIL approach is not different from the training of models with strongly labelled WSIs. Also, the inclusion of data from multiple centres is not straightforward: the authors had access to a high-quality and well-structured pathology archive; yet in most pathology laboratories, structured reporting is not common, diagnostic information is mostly available at the case level rather than at the slide level, and such information may

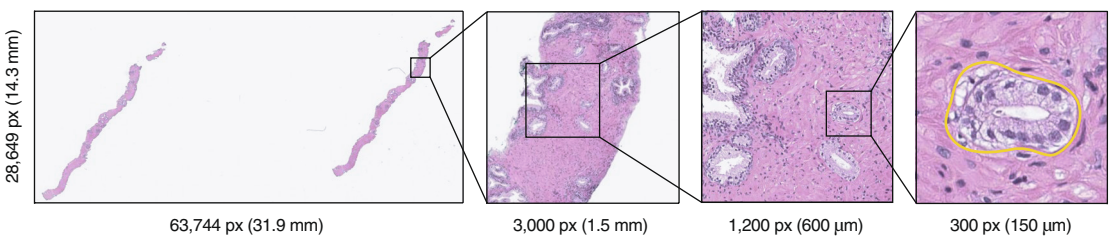

Fig. 1 | Manual pixel-level annotations are challenging to perform at scale. Diagnoses based on haematoxylin-and-eosin-stained slides (here from a biopsy of prostatic adenocarcinoma) can rely on small foci (rightmost image) of cancer (accounting for less than 1% of the tissue surface). px, pixels. Figure reproduced from ref. [3], Springer Nature America, Inc.

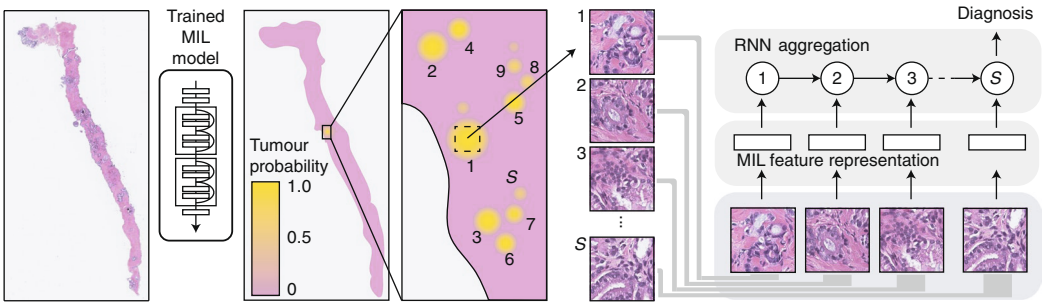

Fig. 2 | Tumour diagnosis via MIL in the absence of pixel-level annotations. The trained MIL model finds and orders the tiles according to the probability of tumour features, which are passed to a recurrent neural network (RNN) that integrates the information and provides a diagnosis. Figure reproduced with permission from ref. [3], Springer Nature America, Inc.

contain a variety of diagnoses and even errors. Furthermore, for all samples of basal cell carcinoma a pathologist had to assess free-text reports in order to determine the slide-level labels. It remains to be seen whether this relatively moderate level of data curation for a very large number of cases (especially when limited information in reports requires a pathologist to review multiple slides per case) is more feasible than producing pixel-level annotations for a much smaller number of cases.

Fuchs and colleagues explored a two-stage approach to MIL (Fig. 2). A learned 'tile-level' feature representation assigns a tumour-probability score to every image tile in the WSI. A second trained classifier then calculates the tumour probability for the entire WSI on the basis of the tile-level information. More powerful MIL approaches, especially those where WSI-level classification is directly integrated with learning tile-level feature representations, may substantially reduce the size of the required weakly labelled dataset[6]. Still, MIL approaches are not well-suited for applications in which the positive fraction (such as a tumour) is much smaller than the amount of negative data, especially if this rate is variable between samples[7]. Hence, very small tumours may be missed because of such a 'witness rate' phenomenon. This may be reflected in the fact that the MIL approach for breast-cancer lymph-node metastases scored considerably lower (AUC = 0.965) than the current best deep-learning model based on strong labels (best CAMELYON16 result, AUC = 0.996; ref. [8]).

In Fuchs and colleagues' work, the task (the detection of prostate cancer, basal cell carcinoma and lymph-node metastases) is relatively well-defined. How well the MIL approach would perform on tasks that are inherently more complex, such as the automation of Gleason grading for prostate cancer (a multiclass problem rather than a binary 'absence or presence of disease' problem) is unclear. Class boundaries are relatively poorly defined, and 'label noise' (arising from the incorrect labelling of training data) can be a big problem. Nevertheless, the authors' approach is a significant addition to the computational-pathology toolbox, as it relieves the burden of obtaining strong annotations. Largely removing the involvement of experts in algorithm training reduces costs and decreases the time required for data collection. Also, weak labels are readily available for many of the cases stored in pathology archives, and therefore collecting large amounts of such data is achievable, and the inclusion of data from a larger number of clinical centres can yield models that generalize better. Especially for relatively straightforward applications, this may result in powerful and clinically useful deep-learning models with enhanced generalizability. Moreover, the algorithm could be trained against an end point for which no clear morphological biomarker is known. For instance, a deep-learning model could be trained to predict 5-year cancer-free survival directly from the WSIs, and could thus help discover image-based histopathological biomarkers[9]. ❐

Jeroen van der Laak[1,2*], Francesco Ciompi[1] and Geert Litjens[1]

[1]Department of Pathology, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands. [2]Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden.

**e-mail: Jeroen.vanderlaak@radboudumc.nl*

Published online: 17 October 2019
https://doi.org/10.1038/s41551-019-0472-6

References

1. Litjens, G. et al. *Sci. Rep.* **6**, 26286 (2016).
2. Ehteshami Bejnordi, B. et al. *JAMA* **318**, 2199–210 (2017).
3. Campanella, G. et al. *Nat. Med.* **25**, 1301–1309 (2019).
4. Xu, Y., Zhu, J. Y., Chang, E., Lai, M. & Tu, Z. *Med. Image Anal.* **18**, 591–604 (2014).
5. Tellez, D. et al. *IEEE Trans. Med. Imaging* **37**, 2126–2136 (2018).
6. Ilse, M., Tomczak, J. M. & Welling, M. Preprint at *arXiv* https://arxiv.org/abs/1802.04712v4 (2018).
7. Carbonneau, M. A., Cheplygina, V., Granger, E. & Gagnon, G. *Pattern Recog.* **77**, 329–353 (2018).
8. Liu, Y. et al. *Arch. Pathol. Lab. Med.* **143**, 859–868 (2019).
9. Abels, E. et al. *J. Pathol.* https://doi.org/10.1002/path.5331 (2019).

Competing interests

Jeroen van der Laak is a member of the scientific advisory boards of Philips (The Netherlands) and ContextVision (Sweden), and receives research funding from Philips (The Netherlands) and from Sectra (Sweden). The remaining authors declare no competing interests.