Mod Pathol 37 (2024) 100417

MODERN PATHOLOGY

USCAP UNITED STATES AND CANADIAN ACADEMY OF PATHOLOGY
Creating a Better Pathologist

Journal homepage: https://modernpathology.org/

Research Article

Endometrial Pipelle Biopsy Computer-Aided Diagnosis: A Feasibility Study

Sanne Vermorgen[a], Thijs Gelton[a], Peter Bult[a], Heidi V.N. Kusters-Vandevelde[b], Jitka Hausnerová[c], Koen Van de Vijver[d], Ben Davidson[e,f], Ingunn Marie Stefansson[g,h], Loes F.S. Kooreman[i], Adelina Qerimi[j], Jutta Huvila[k], Blake Gilks[l], Maryam Shahi[m], Saskia Zomer[b], Carla Bartosch[n], Johanna M.A. Pijnenborg[o], Johan Bulten[a], Francesco Ciompi[a], Michiel Simons[a,*]

[a] *Department of Pathology, Radboudumc, Nijmegen, the Netherlands;* [b] *Department of Pathology, Canisius-Wilhelmina Hospital, Nijmegen, the Netherlands;* [c] *Department of Pathology, University Hospital Brno, Brno, Czech Republic;* [d] *Department of Pathology, UZ Gent, Gent, Belgium;* [e] *Department of Pathology, Oslo University Hospital, Norwegian Radium Hospital, Oslo, Norway;* [f] *University of Oslo, Faculty of Medicine, Institute of Clinical Medicine, Oslo, Norway;* [g] *Centre for Cancer Biomarkers CCBIO, Department of Clinical Medicine, Section for Pathology, University of Bergen, Bergen, Norway;* [h] *Department of Pathology, Haukeland University Hospital Bergen, Bergen, Norway;* [i] *Department of Pathology, GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, the Netherlands;* [j] *Department of Pathology, ViraTherapeutics GmbH, Innsbruck, Austria;* [k] *Department of Pathology, University of Turku, Turku University Hospital, Turku, Finland;* [l] *Department of Pathology, University of British Columbia, Vancouver, Canada;* [m] *Department of Pathology, Mayo Clinic, Rochester, Minnesota;* [n] *Department of Pathology, Portuguese Oncology Institute Lisbon, Lisbon, Portugal;* [o] *Department of Gynecology, Radboudumc, Nijmegen, the Netherlands*

ARTICLE INFO

Article history:
Received 8 August 2023
Revised 2 December 2023
Accepted 19 December 2023
Available online 27 December 2023

Keywords:
classification
digital pathology
endometrial cancer
interobserver variability

ABSTRACT

Endometrial biopsies are important in the diagnostic workup of women who present with abnormal uterine bleeding or hereditary risk of endometrial cancer. In general, approximately 10% of all endometrial biopsies demonstrate endometrial (pre)malignancy that requires specific treatment. As the diagnostic evaluation of mostly benign cases results in a substantial workload for pathologists, artificial intelligence (AI)-assisted preselection of biopsies could optimize the workflow. This study aimed to assess the feasibility of AI-assisted diagnosis for endometrial biopsies (endometrial Pipelle biopsy computer-aided diagnosis), trained on daily-practice whole-slide images instead of highly selected images. Endometrial biopsies were classified into 6 clinically relevant categories defined as follows: nonrepresentative, normal, nonneoplastic, hyperplasia without atypia, hyperplasia with atypia, and malignant. The agreement among 15 pathologists, within these classifications, was evaluated in 91 endometrial biopsies. Next, an algorithm (trained on a total of 2819 endometrial biopsies) rated the same 91 cases, and we compared its performance using the pathologist's classification as the reference standard. The interrater reliability among pathologists was moderate with a mean Cohen's kappa of 0.51, whereas for a binary classification into benign vs (pre)malignant, the agreement was substantial with a mean Cohen's kappa of 0.66. The AI algorithm performed slightly worse for the 6 categories with a moderate Cohen's kappa of 0.43 but was comparable for the binary classification with a substantial Cohen's kappa of 0.65. AI-assisted diagnosis of endometrial biopsies was demonstrated to be feasible in discriminating between benign and (pre)malignant endometrial tissues, even when trained on unselected cases. Endometrial premalignancies remain challenging for both pathologists and AI algorithms. Future steps to improve reliability of the diagnosis are

These authors contributed equally: Sanne Vermorgen and Thijs Gelton.
These authors are first authors: Francesco Ciompi and Michiel Simons.

* Corresponding author.
E-mail address: Michiel.Simons@radboudumc.nl (M. Simons).

https://doi.org/10.1016/j.modpat.2023.100417

needed to achieve a more refined AI-assisted diagnostic solution for endometrial biopsies that covers both premalignant and malignant diagnoses.

Introduction

With a global age-standardized incidence rate of 8.7 per 100,000 women and a global age-standardized mortality rate of 1.8 per 100,000 women, endometrial carcinomas (ECs) present a significant disease burden on society.[1] Most patients present with postmenopausal bleeding and will be referred for endometrial aspiration biopsy (Pipelle) after clinical evaluation to rule out a (pre)malignancy.[2]

The endometrioid histologic subtype represents 80% of all ECs.[3] According to the fifth edition of the World Health Organization (WHO) classification,[4] 2 lesions of this subtype with treatment necessity can be distinguished: hyperplasia without atypia (H) and hyperplasia with atypia (AH)/endometrial intraepithelial neoplasia (EIN), with the latter diagnosed based on both architectural and cytomorphologic alterations.[5] However, there is substantial interrater variability for these diagnoses, leading to poor reproducibility.[6-14] Moreover, Pipelle biopsies often result in a very low yield of endometrial tissues, frequently fragmented, which could lead to sampling errors that complicate histopathologic evaluations.

The introduction of digital pathology has made it possible to use advanced image analysis on digitized whole-slide images (WSI) using artificial intelligence (AI) algorithms. In the field of computer-aided diagnosis for gynecologic oncology, only a handful of studies, with traditional learning and more recently with deep learning models, have focused on diagnosis of (pre) malignant endometrial tissues.[15-17]

Since the majority of endometrial Pipelle biopsies show benign endometrium, the largest gain in terms of a pathologist's workload reduction can be found in automating the detection of normal endometrial tissues. Deep learning algorithms that can separate benign tissue from suspicious tissue, which needs to be assessed by a pathologist to make the final diagnosis, could assist pathologists in efficiently managing the diagnostic evaluation of endometrial Pipelle biopsy samples.

The purpose of this study was to investigate the feasibility of an algorithm, trained on noisy straight-from-clinical-practice data, which can reliably separate suspicious WSI from normal WSI. In order to pinpoint potential limitations, we studied the interrater variability among expert pathologists worldwide in diagnosing endometrial Pipelle biopsies using the current WHO 2020 classification system and used their diagnostic opinion as a baseline to evaluate the performance of our AI algorithm.

Material and Methods

Data set Endometrial Biopsies

All consecutive Pipelle endometrial biopsies (n = 2910) analyzed between October 2013 and April 2021 were retrieved from the pathology archive of the Radboud University Medical Center (Radboudumc), Nijmegen, the Netherlands, using a PALGA (the nationwide network and registry of histopathology and cytopathology in the Netherlands) database search for the following: "tissue type = endometrium" and "method of retrieval = Pipelle." Endometrial biopsies obtained by dilatation and curettage, hysteroscopy, and polypectomy were excluded because these are generally second-line procedures when Pipelle endometrial biopsy resulted in insufficient tissue or inconclusive diagnosis. Hematoxylin and eosin–stained slides were scanned using a 3DHistech P1000 scanner at 0.25 μm/pixel. All digital data, ie, digitized slides and pathology reports, were coded before being used in this study. A total of 91 WSI, randomly chosen with a predefined category-weighted key (7 nonrepresentative (NR), 16 normal (NL), 17 nonneoplastic (NN), 16 H, 29 AH, and 6 malignant (M), see classification section; classification based on the original diagnosis), were stored separately to be used in the interrater variability study and evaluation of the algorithm. The remaining 2819 cases were used to develop the AI algorithm.

Classification

The pathology diagnosis from the original report was translated by a single coder into 6 categories according to the WHO 2020: NR, NL, NN, H, AH, and M, as summarized in Table 1. In order to preserve the resemblance of daily clinical practice, cases with an ambiguous or uncertain classification in the original pathology report were not removed. To avoid false-negative classifications, a low threshold was used to label a case as H or AH (eg, labeling a case as H, although the report only mentions a suspicion of H, or labeling a case as AH, although atypia, could still be reactive). This approach was based on the necessity of high sensitivity to avoid premalignant cases being missed by automated screening. Although this strategy could potentially lead to detection of more false positives, those cases would still be reviewed by a pathologist.

The NN category was added in order to refine NL cases from cases that are benign but show morphologic changes that might confound an algorithm during training.

Table 1
Translation of the histopathologic endometrial diagnosis into categories

Category label	Category code	Category description
Nonrepresentative	NR	Insufficient tissue for conclusive diagnosis
Normal	NL	Cyclical or atrophic endometrium without any signs of pathology or treatment effect
Nonneoplastic	NN	Any nonneoplastic change (eg, treatment effect, endometrial polyp, infection, etc), which does not belong in any of the other categories
Hyperplasia without atypia	H	(possible) Hyperplasia, no mention of atypia
Hyperplasia with atypia	AH	(possible) Hyperplasia and (possible) atypia
Malignant	M	Any malignancy

Since the ultimate purpose of the algorithm is separating benign cases from (pre)malignant cases that should be evaluated by a pathologist, we grouped the 6 categories into 2 classes as follows: benign (NR, NL, and NN) and (pre)malignant (H, AH, and M).

Interrater Variability Study

To assess the interrater variability in the evaluation of endometrial Pipelle biopsies among pathologists, expert gynecologic pathologists were contacted through the European Network for Individualized Treatment of Endometrial Cancer. A total of 15 pathologists (endometrial Pipelle biopsy computer-aided diagnosis consortium partners) participated in a reader study on the web-based grand-challenge.org platform (Diagnostic Image Analysis Group, Radboudumc), which implements a web-based viewer of WSI and a user interface with functionalities to label and manually annotate WSI.

The 91 WSI of the evaluation set were classified by the participating pathologists in 1 of the 6 categories as described above. Patient age was made available to the readers. The pathologist's certainty about the correctness of this classification was also assessed using a 4-step rating scale: very unsure, unsure, sure, or very sure. The option to leave a comment or make an annotation on the image, when they found a particular region of the WSI particularly informative for their classification score, was provided. Years of experience in gynecologic pathology were also recorded.

Statistical Analysis

Excel (Microsoft 365), SPSS (IBM, version 25), and R studio (R version 1.1.3, the R Foundation for Statistical Computing) were used for data analysis. To measure interrater reliability, we used unweighted Cohen's kappa for pairwise comparisons between the readers to compare each reader against the majority vote of the other 14 readers and to compare the performance of the AI with the performance of the readers. We also calculated a Fleiss' kappa for individual categories.

Next, the unweighted Cohen's kappa for the binary split between benign (NR, NL, and NN) versus (pre)malignant (H, AH, and M) was computed.

Interpretation of the kappa values was based on the Altman[18] guidelines (adapted from the Landis Koch[19] guidelines). Briefly, a kappa value between 0 and 0.2 is interpreted as "bad" agreement, 0.21 and 0.4 as "fair" agreement, 0.41 and 0.6 as "moderate" agreement, 0.61 and 0.8 as "substantial" agreement, and 0.81 and 1 as "(almost) perfect" agreement.

Artificial Intelligence System Development

The algorithm, which was developed during this study, takes an entire WSI as input and learns to predict a single label belonging to the set of categories: NR, NL, NN, H, AH, and M.

The approach is based on the work of Lu et al[20] in 2021, namely the clustering-constrained attention multiple-instance learning (CLAM) method, which implements a form of multiple-instance learning (MIL). In brief, MIL is a machine learning technique that treats WSI as a collection of instances, ie, smaller image patches. In a binary problem where WSI should be classified as containing either completely benign tissues or some M cells, MIL assumes that if there is at least one tile categorized as M, then the entire image is diagnosed as M, otherwise as benign. Based on this assumption, only a single label (ie, benign or M) per slide is required to train a MIL-based algorithm. This approach belongs to the category of weakly supervised learning because there is no need for a pathologist to manually annotate at the cell-level (pixel by pixel) label, and slide-level labels can be used instead. This allows efficient usage of large-scale data sets in algorithm training without the need for manual annotations, instead relying on readily available slide-level information on diagnosis. In this work, we chose to use CLAM compared with traditional MIL-based approaches because it allows extending the learning procedure to a multiclass setting, and it was also shown to be more data-efficient.[20] To account for the imbalance in class labels during training, the data were resampled using the class label's inverse probability. This means that if there is only 1 AH sample and 10 N samples, then the chance of sampling a N sample is 10 times smaller than an AH sample.

The algorithm consists of 2 main steps, schematically shown in Figure 1. The first step is dividing the tissue in the slide into smaller, equally sized tiles, which are then compressed into a smaller vector of features by using a convolutional neural network. In this way, each WSI is first converted into a collection of feature vectors to be used in downstream parts of the algorithm. The second step consists of the learning process. In this phase, the algorithm learns to correlate visual patterns in the tiles to the multiple considered categories and uses an "attention mechanism" to learn what tiles in each WSI it should look at to make a correct prediction. Hyperparameters, adjustable settings that determine the behavior and performance of a neural network during the training process, can be found in the Supplementary Methods.

At the test time, the algorithm processes each tile separately and produces a slide-level prediction by combining the processed features with the output of the attention module. By linking the output of this attention mechanism, in step 3, with tiles in the WSI, an "attention map" is obtained, which has been shown to provide visual cues on the relevance of morphologic patterns in the algorithm's predictions.[20,21] This "attention map" offers insights into which parts of the WSI were judged by the algorithm to be most relevant for its decision making process, thus making the algorithm more interpretable. We made the algorithm publicly available for research purposes on the grand-challenge platform. See Data Availability for access to the algorithm.

Evaluation and Computing Workload Reduction

Aside from measuring the algorithm's performance, we computed the percentage of workload that would be reduced for a pathologist when hypothetically deploying the algorithm to bypass the pathologist when WSI are rated as benign. This computation was based on the distribution of the development data set. Reducing workload can only be performed, in terms of clinical usability, when 100% sensitivity is reached (ie, zero false negatives: no (pre)malignancies missed). The false positive rate (FPR) at that point determines the amount of workload that can be reduced. To exemplify, approximately 81% of the data set was benign and 19% (pre)malignant. If the algorithm can reach 100% sensitivity at 10% FPR, then the workload is reduced by 72.9% (81%-81% * 0.10). In other words, the algorithm will detect all (pre)malignancies and classify 8.1% of the benign cases as (pre)malignant, which will be sent to the pathologist for inspection alongside the real (pre)malignant cases. To calculate workload reduction, we looked at the classification of the 91 test set cases rated by the panel of expert pathologists in the reader study. We

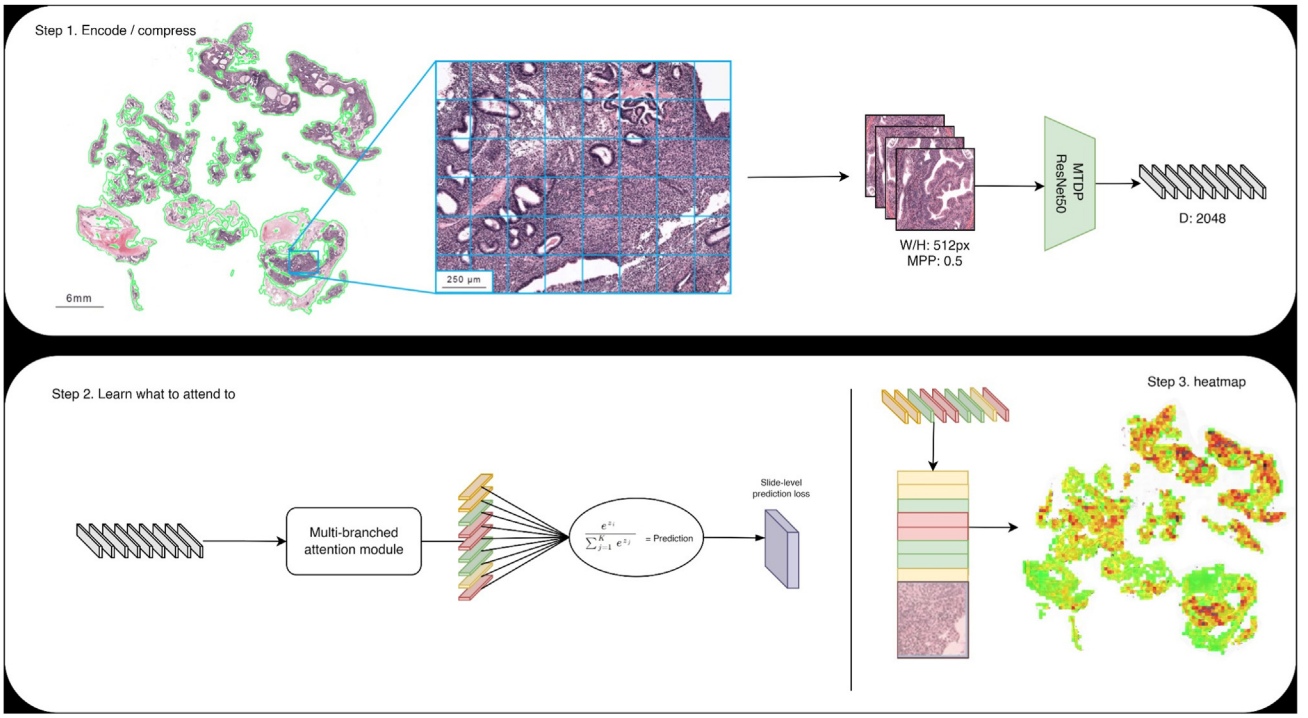

Figure 1.
Schematic overview of the CLAM algorithm. The first step is to sample patches from the tissue and encode/compress this into a small representation of 2048 numerical values (1). Then (2), the algorithm learns what it should attend to in order to aggregate the final attention score into a slide-level diagnosis. To gain a better feature representation, the attention scores are also used as pseudolabels for self-supervision during the clustering of the vector representations. The slide-level (70%) and clustering loss (30%) are combined into a total loss, which gets back propagated through the network, such that the network will improve on future examples. When the network is fully trained, it can produce heat maps as seen in (3), based on the attention scores. Adapted from the study of Lu et al.[20] CLAM, clustering-constrained attention multiple-instance learning.

evaluated 3 situations: (1) 100% sensitive detection of all cases rated by majority vote as (pre)malignant; (2) 100% sensitive detection of all cases rated by at least one reader as (pre)malignant; and (3) 100% sensitive detection of all cases rated by 2 or more readers as (pre)malignant. Configuration selection was made based on the amount of workload reduced by the algorithm on the validation set.

Results

Data set and Reader Study Participants

In total, 2910 WSI were created from archived glass slides: NR ($n = 516$), NL ($n = 1376$), NN ($n = 473$), H ($n = 258$), AH ($n = 136$), and M ($n = 151$). From these WSI, 91 cases (7 NR, 16 NL, 17 NN, 16 H, 29 AH, and 6 M) were included in the test set for the interrater variability study and to evaluate the algorithm.

Fifteen expert pathologists (endometrial Pipelle biopsy computer-aided diagnosis consortium partners), practicing gynecologic pathology in 9 different countries (the Netherlands, Norway, Belgium, Austria, Portugal, Czech Republic, Finland, Canada, and the United States of America) participated in the interrater variability study. Their experience in gynecologic pathology ranged between 3 and 34 years. No statistical differences in category and certainty score were found when stratifying for experience level (for category $P = .85$; for certainty $P = .99$, Kruskal-Wallis test).

One case was rated by only 13 pathologists, and 3 cases, by 14 pathologists. The 87 other cases were rated by all 15 pathologists.

Two of the readers, MS and JB, also participated in the design of the interrater variability study, although they were blinded to the cases selected for the study. Their Cohen's kappa did not significantly differ from that of other readers (*data not shown*).

Interrater Variability Among Pathologists

The overall (average) scores for each of the interrater reliability measures are displayed for the binary and six-category classification in Supplementary Tables S1 and S2, both unweighted and quadratically weighted. For the original 6 categories, the mean Cohen's kappa was 0.51 (moderate agreement) with individual Cohen's kappas, comparing each pathologist's classification with the majority vote of all other pathologists, ranging between 0.3 and 0.61. Pairwise Cohen's kappas were also calculated for each pair of pathologists and ranged between 0.18 and 0.57 (Table 2).

Fleiss' kappa was calculated for the separate categories, and we found large differences between the categories, with a kappa value of 0.85 for M (almost perfect agreement), and a low kappa value for the "hyperplasia" categories (H: 0.20 and AH: 0.17; bad agreement). A closer inspection of the data set showed not a single image exists where more than 8 pathologists agreed about the label H or AH. The classes most often used by other nonagreeing pathologists in cases where at least one pathologist chose one of the "hyperplasia" categories were as follows: NN (56.4% for H and 44.6% for AH) and NL (31.33% for H and 26.48% for AH).

For a binary split (benign: NR, NL, and NN vs (pre)malignant: H, AH, and M), the kappa values increase slightly to a mean Cohen's kappa of 0.66 and individual Cohen kappas ranging between 0.38 and 0.95 (Table 3).

Table 2
Unweighted Cohen's kappa with 6 categories

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AI
MV	0.55[a]	0.41[a]	0.53[a]	0.61[a]	0.60[a]	0.58[a]	0.45[a]	0.53[a]	0.46[a]	0.63[a]	0.30[a]	0.57[a]	0.53[a]	0.51[a]	0.45[a]	*0.43*[b]
1		0.22	0.42	0.51	0.52	0.42	0.34	0.53	0.34	0.35	0.18	0.48	0.39	0.54	0.43	*0.34*[b]
2			0.25	0.44	0.37	0.36	0.26	0.28	0.21	0.44	0.18	0.33	0.33	0.31	0.26	*0.22*[b]
3				0.43	0.45	0.42	0.38	0.46	0.46	0.51	0.37	0.39	0.54	0.55	0.36	*0.47*[b]
4					0.48	0.53	0.33	0.43	0.39	0.57	0.22	0.45	0.50	0.53	0.47	*0.36*[b]
5						0.45	0.43	0.52	0.33	0.53	0.20	0.52	0.47	0.49	0.45	*0.34*[b]
6							0.30	0.42	0.42	0.54	0.32	0.42	0.51	0.41	0.41	*0.37*[b]
7								0.34	0.30	0.45	0.35	0.40	0.37	0.34	0.25	*0.31*[b]
8									0.48	0.45	0.25	0.39	0.40	0.53	0.34	*0.38*[b]
9										0.41	0.36	0.40	0.40	0.40	0.41	*0.52*[b]
10											0.36	0.46	0.51	0.40	0.37	*0.41*[b]
11												0.30	0.38	0.23	0.21	*0.31*[b]
12													0.37	0.48	0.40	*0.37*[b]
13														0.44	0.38	*0.39*[b]
14															0.48	*0.47*[b]
15																*0.42*[b]

MV, majority vote.
Pairwise comparison between readers.
[a] Comparison with majority vote.
[b] Comparison with AI algorithm.

The mode of the category score (all participants included) was compared with the labels as translated from the original report. The readers tended to favor the categories NL and NN, whereas pathologists frequently chose H and AH than what was translated from the original report (see Supplementary Fig. S1). Figure 2 shows some examples of cases with high category agreement among the readers.

Comments and Annotations

In total, 383 comments were left, and 205 annotations were made. Comments were most often used to express uncertainty about the classification, requests for immunohistochemistry, or clinical information ($n = 189$) to clarify the chosen category ($n = 135$) and report the suboptimal quality of WSI ($n = 59$).

Overall, pathologists used the "comments" option more often when they were unsure about the category, especially when diagnosing H or AH (Supplementary Tables S3 and S4).

The 10 cases with the least amount of agreement were inspected more closely. For 9 of them, the disagreement was not limited to the adjacent categories belonging to the same benign or premalignant binary division (NR, NN, and NL vs H, AH, and M), see Supplementary Fig. S2. Often, this was just 1 or 2 pathologists, who indicated their level of certainty as "unsure" or "very unsure." One of the cases with low agreement was a case that might have been suboptimally scanned. The other 9 images did not show scanning or staining quality issues. The case with the most disagreement among pathologists is displayed in Figure 3. Five pathologists made an annotation on the WSI, which identified 2 regions of interest.

Table 3
Cohen's kappa with 2 categories

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AI
MV	0.77 [a]	0.69 [a]	0.47 [a]	0.81 [a]	0.62 [a]	0.90 [a]	0.47 [a]	0.53 [a]	0.38 [a]	0.73 [a]	0.66 [a]	0.95 [a]	0.78 [a]	0.70 [a]	0.51 [a]	*0.65* [b]
1		0.58	0.65	0.87	0.52	0.82	0.39	0.59	0.43	0.62	0.56	0.81	0.71	0.88	0.65	*0.66* [b]
2			0.44	0.62	0.73	0.65	0.47	0.46	0.36	0.73	0.66	0.73	0.73	0.53	0.40	*0.46* [b]
3				0.59	0.35	0.51	0.39	0.42	0.35	0.41	0.51	0.47	0.47	0.64	0.36	*0.50* [b]
4					0.65	0.77	0.41	0.62	0.40	0.75	0.59	0.85	0.65	0.84	0.61	*0.69* [b]
5						0.59	0.38	0.49	0.33	0.66	0.59	0.66	0.66	0.53	0.42	*0.41* [b]
6							0.44	0.50	0.40	0.69	0.62	0.90	0.79	0.71	0.57	*0.50* [b]
7								0.31	0.19	0.63	0.81	0.51	0.51	0.32	0.20	*0.31* [b]
8									0.45	0.55	0.44	0.56	0.41	0.71	0.40	*0.70* [b]
9										0.38	0.29	0.38	0.33	0.47	0.40	*0.52* [b]
10											0.82	0.78	0.78	0.61	0.42	*0.55* [b]
11												0.71	0.71	0.47	0.31	*0.44* [b]
12													0.78	0.70	0.49	*0.56* [b]
13														0.61	0.42	*0.41* [b]
14															0.63	*0.71* [b]
15																*0.57* [b]

MV, majority vote.
Pairwise comparison between readers.
[a] Comparison with majority vote.
[b] Comparison with AI algorithm.

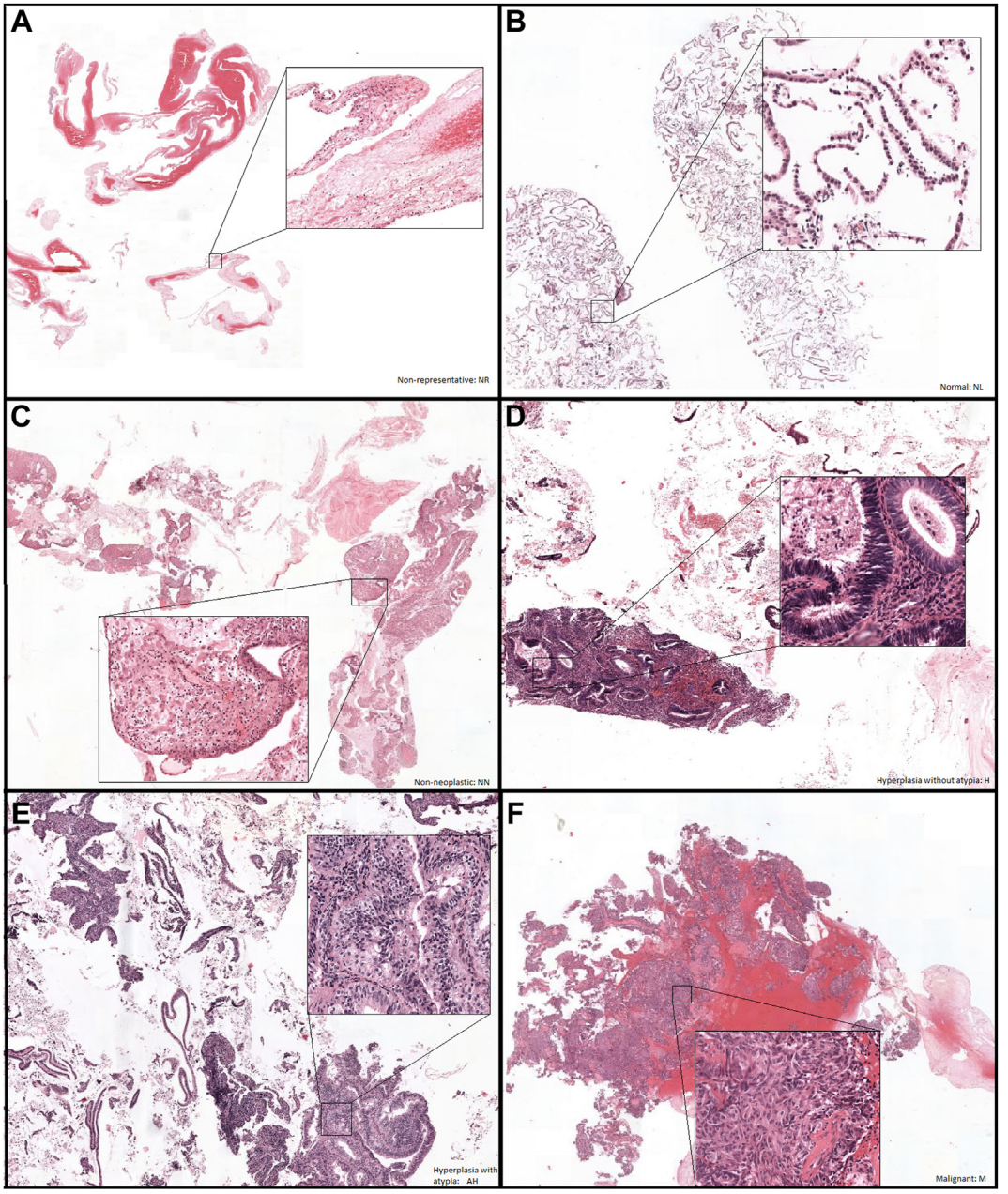

Figure 2.
Example case of each class. (A) Nonrepresentative (NR); (B) normal (NL); (C) nonneoplastic (NN); (D) hyperplasia without atypia (H); (E) hyperplasia with atypia (AH); and (F) malignant (M).

Results Algorithm

We decided to only optimize the hyperparameters for the binary classifier because the correct classification within the (pre) malignant categories is not necessary when the ultimate purpose was the separation of benign versus (pre)malignant. The configuration for the best-performing binary classifier can be found in Supplementary Table S5. The bootstrapped area under the curve (AUC) score ($N = 10{,}000$) on the validation set was 0.855 (0.818-0.855) and the test set was 0.960 (0.924-0.986).

The same hyperparameters, which were optimized for the binary classifier, were used to train a 6-class classifier to inspect the models' agreement for the binary and 6-class version with the pathologists in the reader study. Cohen's kappa for the algorithm's responses compared with the majority vote of all readers was 0.43 (moderate agreement) for the 6-category classification and 0.65 (substantial agreement) for the binary classification. The pairwise Cohen's kappa between each reader and the algorithm for the binary classification ranged between 0.31 and 0.71, which is comparable with the pairwise agreement between the individual readers (Tables 2 and 3).

The receiver operating characteristic curves show that when the majority vote is used as the reference standard, 100% sensitivity was reached with a FPR of 10% (Fig. 4), which would lead to an estimated workload reduction of 72.9%. When cases are considered (pre)malignant if at least one pathologist evaluated

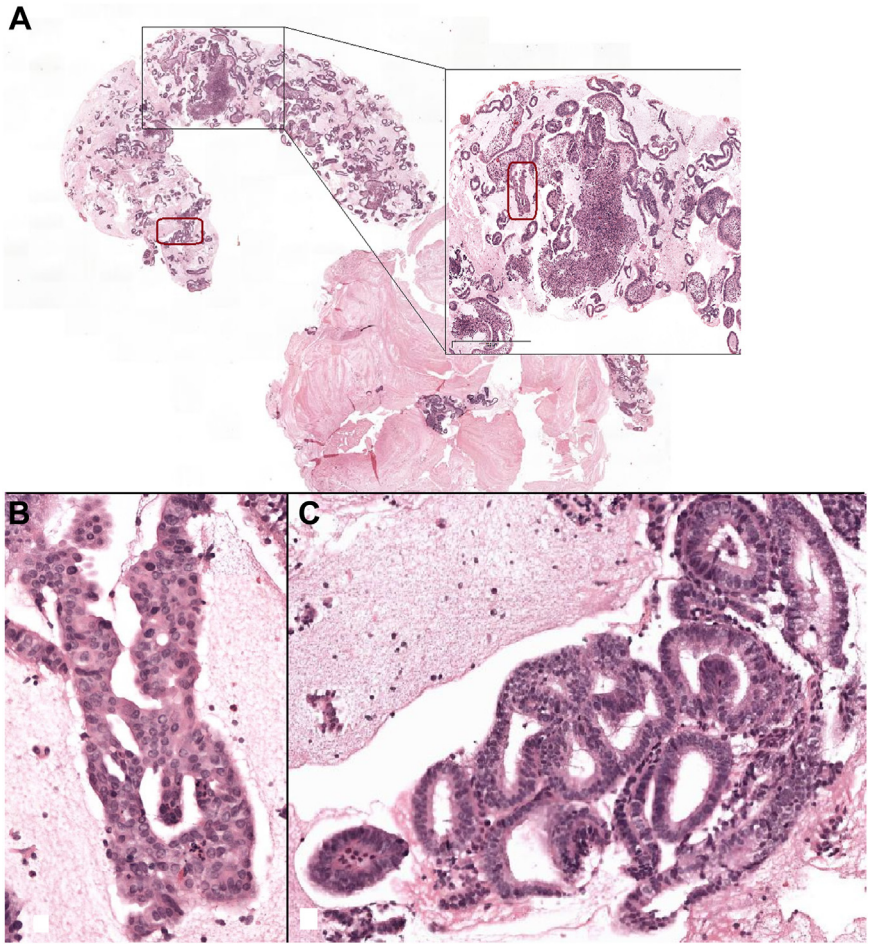

Figure 3.
The case with the most disagreement among pathologists (case 1 from Supplementary Figure S2). (A) Overview of the image with encircled in red the regions annotated by the pathologists. (B) Closer view of the right region, scored by the annotating pathologists as AH. (C) Closer view of the left region, scored by the annotating pathologists as H. No annotations were left by the pathologists who classified the image as normal (NL) or nonneoplastic (NN).

them as such, the model did not reach 100% sensitivity. In situation 3, if at least 2 pathologists had to rate a case as (pre)malignant to set the label as such, the model reached 100% sensitivity with a FPR of 37% (Fig. 4). This results in an estimated workload reduction of 51.03%.

Discussion

We demonstrated in the current study that it is feasible to develop AI solutions to categorize endometrial biopsies into benign and potentially (pre)malignant. Using a data set of unselected daily-practice cases, we managed to train an algorithm that achieved a diagnostic reproducibility comparable with an international consortium of 15 expert pathologists. Further development and implementation of an AI algorithm to filter benign endometrial tissue samples from the day-to-day diagnostic workload could lead to a substantial workload reduction, saving time and resources for the pathologists to focus on more challenging tasks.

The algorithm developed in this study resulted in likelihood for the presence of (pre)malignant tissues and an interpretable heatmap, which gives more insights to pathologists while evaluating endometrial diagnostic samples. The algorithm, available on the grand-challenge platform,[1] can give a diagnosis on WSI in approximately 8 min and could, in clinical practice, be parallelized for even faster processing.

Therefore, AI algorithms in pathology have been studied for some time with promising results in prostate and breast cancer diagnostics.[22,23] For endometrial tissue classifications, few studies with AI algorithms have been published. Downing et al[15] showed promising results using a random forest algorithm to help classify benign, premalignant, and M endometrial tissues. Papke[16] used the same random forest algorithm to highlight neoplastic glands to facilitate premalignancy classification. Sun et al[17] described a convolutional neural network for computer-assisted diagnosis of endometrial tissues, focusing on 4 classes as follows: normal tissues, endometrial polyps, hyperplasia, and malignancy.

The aforementioned studies have shown promising results, but in a well-controlled experimental environment, without showcasing the clinical applicability of computer models. For the development of their algorithm, Sun et al[17] only included tissue samples with good agreement by 3 pathologists, which were considered good examples for the categories. Their images consisted of fair amounts of tissue without fragmentation or artifacts. Mimics of premalignancy, such as reactive atypia, were not included, which hampers routine use as an algorithm only trained with textbook examples and might not perform well in clinical practice.

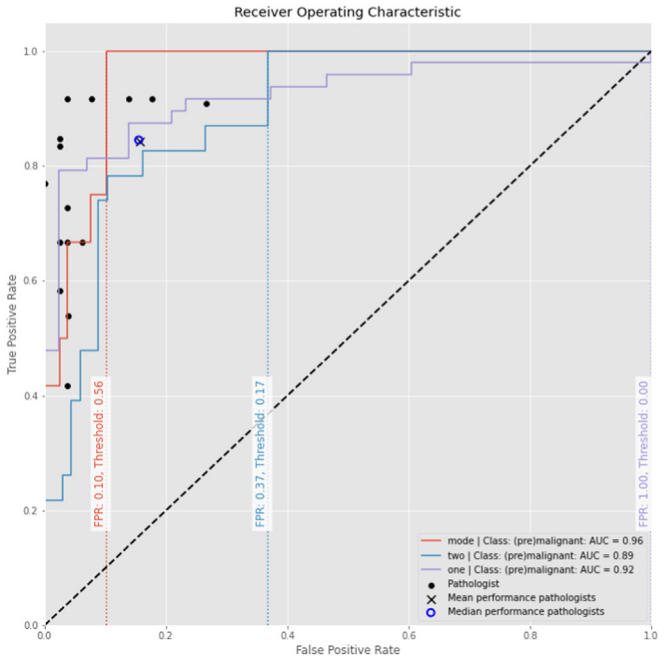

Figure 4.
Performance of AI algorithm and panel of pathologists. The red line indicates the receiver operating characteristic curve of the AI algorithm when the majority vote is the reference standard, the blue curve when 2 votes is the reference standard, the purple curve when one vote is the reference standard. The true and false positive rates of each pathologist in the reader study are displayed as a black dot. The mean and median performances of all pathologists (overlapping) are displayed as a black cross and blue circle, respectively. The dotted lines mark the FPR where the model reaches 100% sensitivity, and the corresponding thresholds are displayed vertically. AI, artificial intelligence; FPR, false positive rate.

In this study, we aimed to investigate the feasibility of training an algorithm with WSI from clinical practice without preselection and without rigorous labeling of training data sets. The inclusion of ambiguous and difficult cases in the training of the algorithm introduces noise in the training setting, but we hypothesize that this could lead to better performance during eventual implementation because training conditions would resemble clinical practice more closely. Using WSI, instead of subregions of the whole slide, and making predictions at the whole-slide level represent also an important novel contribution toward the applicability of the algorithm in clinical practice because digitized laboratories use WSI for daily clinical practice. Additionally, this enables the model to capture all contextual information of the WSI at once, which has been a challenging task for computational pathology, due to the large size of WSI.

Despite the noise in the training data set, the algorithm performed on par with expert pathologists. The assumed necessity to select and annotate cases is one of the major limiting factors to the development of AI implementations in pathology.[24] Very few curated data sets exist, and it is highly time-consuming to create such data sets. Our results with an "unclean" data set could potentially circumvent this limitation, opening up the possibility to train algorithms on vast numbers of archived material available at any pathology department in the world. This method could potentially fast-track AI development in pathology and other medical disciplines with visual data sets.

Although this study offers a proof of concept for the development of algorithms using unselected and unannotated WSI, the applicability of the current algorithm in daily clinical practice is still limited. At this time, we focused on the detection of (endometrioid) premalignancies and epithelial malignancies. Even our model with 6 classes is a simplification of the complex diagnostic landscape of endometrial Pipelles. Benign endometrial abnormalities, such as infectious conditions, hormonal effects, and polyps, which might have clinical significance, would not be flagged by the algorithm at this time. Further development of the algorithm, including enlargement of the training data set and further subdivision of diagnostic categories, both benign and uncommon (pre)malignant, is needed before true implementation can be considered.

This study used only a small test set from a single institution to evaluate its performance. Our test set cases were unselected as well, which led to low reproducibility in the premalignant categories. This is in part due to intrinsic challenges of diagnosing premalignant conditions, as described in previous studies.[6-14] In order to truly evaluate the performance of our algorithm, larger series from multiple institutions, with more unequivocal premalignant cases and the inclusion of rare alternative diagnoses such as serous endometrial intraepithelial carcinoma, in this study, classified as AH ($n = 4$), are needed. To design a proper test set, enrichment with consensus cases by a panel of expert pathologists is advisable. The open-access grand-challenge platform facilitates interested pathologists to test the algorithm on their own digital slides.

In the context of our study, we identified several important issues for consideration in future research. One such issue is the presence of excessive noise in labeling of WSI in unselected cases. To address this, we suggest involving a pathologist in the training process to revise problematic WSI. Additionally, scan quality remains a limitation for digital pathology and AI implementation, and we recommend incorporating quality control measures such as automated rescanning to mitigate this issue. Finally, the inclusion of patient-specific clinical information, such as endometrial thickness and postmenopausal status, might significantly improve algorithm performance in borderline cases. This, however, needs to be explored further.

To summarize, we used an unselected, routine-practice training data set to develop a screening algorithm for endometrial Pipelle biopsies. This algorithm succeeded in separating NR and NL endometrial Pipelle biopsies from potentially premalignant and M Pipelle biopsies, with a reliability similar to a cohort of expert pathologists. Development of algorithms using unselected data sets is feasible. Further studies with multicentric, large cohorts are needed. Our algorithm can be tested on the grand-challenge platform. See Data Availability for access to the algorithm.

Acknowledgments

The authors would like to thank the computation pathology group of Radboudumc for the use of their resources in the reader study and the development of the algorithm.

Author Contributions

S.V. and T.G. collected and coded the data, developed the algorithm, and wrote the manuscript. J.P., M.S., and J.B. were involved in conceptualizing and supervising the study and the manuscript from a clinical perspective. F.C. supervised the algorithm development and the computational parts of the manuscript. P.B. contributed significantly to the editing of the manuscript. All other authors participated in the study and read and approved the manuscript.

Data Availability

The algorithm is available through the link: https://grand-challenge.org/algorithms/endometrial-carcinoma-classification. The data set used for the reader study is available from the corresponding author on request

Funding

This study was funded by internal funds from Radboudumc Nijmegen and Radboud University.

Declaration of Competing Interest

None reported.

Ethics Approval and Consent to Participate

The need for ethical approval was waived by the Commissie Mensgebonden Onderzoek of Radboudumc. The study was performed in accordance with the Declaration of Helsinki.

Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.modpat.2023.100417.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249. https://doi.org/10.3322/caac.21660
2. Timmermans A, Opmeer BC, Khan KS, et al. Endometrial thickness measurement for detecting endometrial cancer in women with postmenopausal bleeding: a systematic review and meta-analysis. *Obstet Gynecol*. 2010;116(1):160–167. https://doi.org/10.1097/AOG.0b013e3181e3e7e8
3. Boll D, Karim-Kos HE, Verhoeven RHA, et al. Increased incidence and improved survival in endometrioid endometrial cancer diagnosed since 1989 in the Netherlands: a population based study. *Eur J Obstet Gynecol Reprod Biol*. 2013;166(2):209–214. https://doi.org/10.1016/j.ejogrb.2012.10.028
4. Masood M, Singh N. Endometrial carcinoma: changes to classification (WHO 2020). *Diagn Histopathol*. 2021;27(12):493–499. https://doi.org/10.1016/j.mpdhp.2021.09.003
5. Scully RE, Bonfiglio TA, Kurman RJ, Silverberg SG, Wilkinson EJ. *Histological Typing of Female Genital Tract Tumours*. Springer Berlin; 1994. Accessed August 7, 2022. http://public.eblib.com/choice/PublicFullRecord.aspx?p=6555049
6. Zaino RJ, Kauderer J, Trimble CL, et al. Reproducibility of the diagnosis of atypical endometrial hyperplasia: a gynecologic oncology group study. *Cancer*. 2006;106(4):804–811. https://doi.org/10.1002/cncr.21649
7. Izadi-Mood N, Khaniki M, Irvanloo G, Ahmadi SA, Hayeri H, Meysamie A. Determining the inter- and intraobserver reproducibility of the diagnosis of endometrial hyperplasia subgroups and well-differentiated endometrioid carcinoma in endometrial curettage specimens. *Arch Iran Med*. 2009;12:377–382.
8. Hecht JL, Ince TA, Baak JPA, Baker HE, Ogden MW, Mutter GL. Prediction of endometrial carcinoma by subjective endometrial intraepithelial neoplasia diagnosis. *Mod Pathol*. 2005;18(3):324–330. https://doi.org/10.1038/modpathol.3800328
9. Sherman ME, Ronnett BM, Ioffe OB, et al. Reproducibility of biopsy diagnoses of endometrial hyperplasia: evidence supporting a simplified classification. *Int J Gynecol Pathol*. 2008;27(3):318–325. https://doi.org/10.1097/PGP.0b013e3181659167
10. Usubutun A, Mutter GL, Saglam A, et al. Reproducibility of endometrial intraepithelial neoplasia diagnosis is good, but influenced by the diagnostic style of pathologists. *Mod Pathol*. 2012;25(6):877–884. https://doi.org/10.1038/modpathol.2011.220
11. Baak JP, Mutter GL, Robboy S, et al. The molecular genetics and morphometry-based endometrial intraepithelial neoplasia classification system predicts disease progression in endometrial hyperplasia more accurately than the 1994 World Health Organization classification system. *Cancer*. 2005;103(11):2304–2312. https://doi.org/10.1002/cncr.21058
12. Salman MC, Usubutun A, Boynukalin K, Yuce K. Comparison of WHO and endometrial intraepithelial neoplasia classifications in predicting the presence of coexistent malignancy in endometrial hyperplasia. *J Gynecol Oncol*. 2010;21(2):97–101. https://doi.org/10.3802/jgo.2010.21.2.97
13. Travaglino A, Raffone A, Saccone G, et al. Endometrial hyperplasia and the risk of coexistent cancer: WHO versus EIN criteria. *Histopathology*. 2019;74(5):676–687. https://doi.org/10.1111/his.13776
14. Ordi J, Bergeron C, Hardisson D, et al. Reproducibility of current classifications of endometrial endometrioid glandular proliferations: further evidence supporting a simplified classification. *Histopathology*. 2014;64(2):284–292. https://doi.org/10.1111/his.12249
15. Downing MJ, Papke DJ, Tyekucheva S, Mutter GL. A new classification of benign, premalignant, and malignant endometrial tissues using machine learning applied to 1413 candidate variables. *Int J Gynecol Pathol*. 2020;39(4):333–343. https://doi.org/10.1097/PGP.0000000000000615
16. Papke DJ, Lohmann S, Downing M, Hufnagl P, Mutter GL. Computational augmentation of neoplastic endometrial glands in digital pathology displays. *J Pathol*. 2021;253(3):258–267. https://doi.org/10.1002/path.5586
17. Sun H, Zeng X, Xu T, Peng G, Ma Y. Computer-aided diagnosis in histopathological images of the endometrium using a convolutional neural network and attention mechanisms. *IEEE J Biomed Health Inform*. 2020;24(6):1664–1676. https://doi.org/10.1109/JBHI.2019.2944977
18. Altman DG. *Practical Statistics for Medical Research*. Chapman & Hall/CRC; 1999.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. https://doi.org/10.2307/2529310
20. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2021;5(6):555–570. https://doi.org/10.1038/s41551-020-00682-w
21. Chen RJ, Chen C, Li Y, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*. ppp. 16144–16155 Accessed September 14, 2022. https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Scaling_Vision_Transformers_to_Gigapixel_Images_via_Hierarchical_Self-Supervised_Learning_CVPR_2022_paper.html
22. Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*. 2020;21(2):233–241. https://doi.org/10.1016/S1470-2045(19)30739-9
23. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*. 2017;318(22):2199–2210. https://doi.org/10.1001/jama.2017.14585
24. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005