

ARTICLE



Deep learning-based histotype diagnosis of ovarian carcinoma whole-slide pathology images

Hossein Farahani^{1,2,7}, Jeffrey Boschman^{1,7}, David Farnell^{2,3}, Amirali Darbandsari⁴, Allen Zhang^{2,3}, Pouya Ahmadvand¹, Steven J. M. Jones^{1,5}, David Huntsman^{2,5}, Martin Köbel⁶, C. Blake Gilks^{2,3,7}, Naveena Singh^{1,2,3,7}, and Ali Bashashati^{1,2,7}

© The Author(s), under exclusive licence to United States & Canadian Academy of Pathology 2022

Ovarian carcinoma has the highest mortality of all female reproductive cancers and current treatment has become histotype-specific. Pathologists diagnose five common histotypes by microscopic examination, however, histotype determination is not straightforward, with only moderate interobserver agreement between general pathologists (Cohen's kappa 0.54–0.67). We hypothesized that machine learning (ML)-based image classification models may be able to recognize ovarian carcinoma histotype sufficiently well that they could aid pathologists in diagnosis. We trained four different artificial intelligence (AI) algorithms based on deep convolutional neural networks to automatically classify hematoxylin and eosin-stained whole slide images. Performance was assessed through cross-validation on the training set (948 slides corresponding to 485 patients), and on an independent test set of 60 patients from another institution. The best-performing model achieved a diagnostic concordance of 81.38% (Cohen's kappa of 0.7378) in our training set, and 80.97% concordance (Cohen's kappa 0.7547) on the external dataset. Eight cases misclassified by ML in the external set were reviewed by two subspecialty pathologists blinded to the diagnoses, molecular and immunophenotype data, and ML-based predictions. Interestingly, in 4 of 8 cases from the external dataset, the expert review pathologists rendered diagnoses, based on blind review of the whole section slides classified by AI, that were in agreement with AI rather than the integrated reference diagnosis. The performance characteristics of our classifiers indicate potential for improved diagnostic performance if used as an adjunct to conventional histopathology.

Modern Pathology (2022) 35:1983–1990; <https://doi.org/10.1038/s41379-022-01146-z>

INTRODUCTION

Ovarian carcinoma is the deadliest cancer of the female reproductive system, with an estimated 13,770 deaths in the U.S. in 2021¹. It is also a heterogeneous disease with five common histotypes: high-grade serous carcinoma (HGSC) accounts for 70% of cases (and 90% of advanced-stage disease and mortality), clear cell ovarian carcinoma (CCOC) accounts for 12%, endometrioid (ENOC) for 11%, low-grade serous (LGSC) for 4%, and mucinous carcinoma (MUC) for 3%. These five common histotypes have distinct cellular morphologies and etiologies, as well as molecular, genetic, and clinical attributes^{2,3}. There has been a move towards histotype-based treatment⁴, for example with the introduction of PARP inhibitor therapy for patients with HGSC⁵, and this trend increases the importance of accurate histotype diagnosis in practice⁶.

Ovarian carcinoma histotype classification by pathologists is associated with challenges in diagnostic reproducibility and interobserver disagreement^{4,7,8}. Initial diagnosis is performed through histological assessment of hematoxylin & eosin (H&E)-stained sections, but studies have shown that for pathologists without current, gynecologic pathology-specific training, the interobserver agreement is only moderate (0.54–0.67 Cohen's kappa)^{9,10,11}.

Furthermore, the number of pathologists trained has not kept up with the increasing volume and complexity of cancer diagnoses. There has been a 4.5% increase per year in the amount of histopathology requests to laboratories in the United Kingdom, without a simultaneous increase in the pathology workforce. In fact, there is a projected pathologist shortage with >75% of full-time pathologists being 45 years or older and a relative decrease in the number of new pathology residents compared to other specialties¹².

A potential way to enhance pathologist efficiency and diagnostic accuracy is to use machine learning as a diagnostic adjunct. In the past decade, machine learning models have demonstrated their potential for a wide range of applications in digital pathology, including cancer diagnoses^{13–16}. Wang et al.¹⁷ introduced a two-stage transfer learning model for ovarian cancer classification that exceeded the Cohen's kappa of general pathologists. However, they only trained and tested their model with 305 H&E slides from a single dataset. Differences between scanners, tissue processing, and staining procedures between laboratories contribute to substantial variation between datasets^{18–20}; to truly serve as a diagnostic adjunct for practicing pathologists, a generalizable algorithm that can classify ovarian cancer cases from different centers is needed.

¹School of Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada. ²Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC, Canada. ³Vancouver General Hospital, Vancouver, BC, Canada. ⁴Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada. ⁵British Columbia Cancer Research Center, Vancouver, BC, Canada. ⁶Department of Pathology and Laboratory Medicine, University of Calgary, Calgary, AB, Canada. ⁷These authors contributed equally: Hossein Farahani, Jeffrey Boschman, C. Blake Gilks, Naveena Singh, Ali Bashashati. ✉email: blake.gilks@vch.ca; naveena.singh@vch.ca; abashash@bccrc.ca

In this paper, aiming to develop a generalizable model for automatic ovarian carcinoma histotype classification, we used four different deep learning-based algorithms^{17,21,22} on a large dataset of 948 H&E-stained whole slide images (WSIs) and evaluated the performance on a held-out set of 60 cases from a different institution. To the best of our knowledge, our model achieved the highest performance for ovarian carcinoma histotype diagnosis based on an out-of-distribution test set to date and approached the performance level of expert gynecologic pathologists.

MATERIALS AND METHODS

Data acquisition and processing

Two datasets from separate centers containing WSIs of the five common histotypes of ovarian carcinoma were used in this study. To prove the generalizability of our models, we used the first dataset for training and the second dataset (from a different hospital) was used for testing only. A visual summary of the procedure is shown in Fig. 1.

The first dataset (referred to henceforth as Internal Training Dataset) retrieved from the OVCARE archives consisted of 948 WSIs (scanned at 40× objective magnification on an IntelliSite Ultra-Fast Scanner (Philips, Amsterdam, Netherlands)) of 485 patients. The breakdown of the histotypes as shown in Table 1 is as follows: 410 HGSC slides (200 patients), 167 CCOC slides (95 patients), 237 ENOC slides (114 patients), 69 LGSC slides (34 patients), and 65 MUC slides (42 patients). The reference diagnosis for each patient was defined by combination of expert pathology review and molecular assays, typically IHC but also including sequencing in a subset of cases, to give an “integrated” diagnosis²³. Using a combination of annotations from board-certified pathologists (for 416 slides) and pseudo annotations, a maximum of 150 patches per tumor and maximum of 20,000 patches per histotype were extracted from the tumor areas of all the slides at multiple sizes and magnifications (see Table 1). For example, to get 512 × 512 pixel patches at 20× magnification, patches of size 1024 × 1024 pixels at 40× magnification were down-

sampled using the Lanczos filter²⁴. The pseudo annotations were created using a stroma-tumor binary classifier trained on patches from the 416 slides with annotations (see Supplementary Information: Creating Pseudo Annotations). This stroma-tumor classifier has a mean area under receiver operating characteristic (ROC) curve (AUC) of 0.9441 (see Supplementary Table S1 and Supplementary Fig. S1) and was also shown to reliably filter non-malignant samples, such as benign ovarian tissue or benign fallopian tube cases (see Supplementary Tables S2, S3). We chose to limit the training dataset to 20,000 patches per histotype to create a partially balanced dataset because the least represented histotype (MUC) had approximately 10,000 patches extracted. Then each set of patches was grouped by patient origin into a 3-fold cross validation scheme for training (66%), validation (17%), and testing (17%).

The External Test Dataset comprised 60 WSIs (scanned at 40× magnification from an Aperio CSO scanner (Leica Biosystems, Buffalo Grove, IL, United States)) of 60 tumors (from 60 patients) from the University of Calgary. The slides consisted of 31 HGSC, 10 CCOC, 10 ENOC, 4 LGSC, and 5 MUC. The reference diagnoses for these cases was made by one of the authors (MK) and included histologic examination in addition to an 8-marker IHC panel (COSPv3) that predicts ovarian carcinoma histotype with 93% diagnostic concordance²⁵. All WSIs were annotated by a pathologist, and 150 patches of size 1024 × 1024 pixels at 40× magnification were tiled from the tumor regions of the slides belonging to each tumor and down sampled to 512 × 512 pixels at 20× magnification, similar to the Internal Training Dataset. This dataset was used for testing purposes only.

Color normalization

Both datasets were then color normalized using the strategy described by Boschman et al.²⁶. A representative reference image from the Internal Training Dataset was chosen (Supplementary Table S4), and then each patch was randomly normalized by either the Reinhard²⁷, Vahadane²⁸, or Macenko²⁹ methods. The rationale is that if there is not a singular color normalization method that is ideal for all datasets or tasks, using a combination of them to normalize the images should make the images diverse enough to train a generalizable model, but similar enough to overcome the domain shift from having different colors from different datasets.

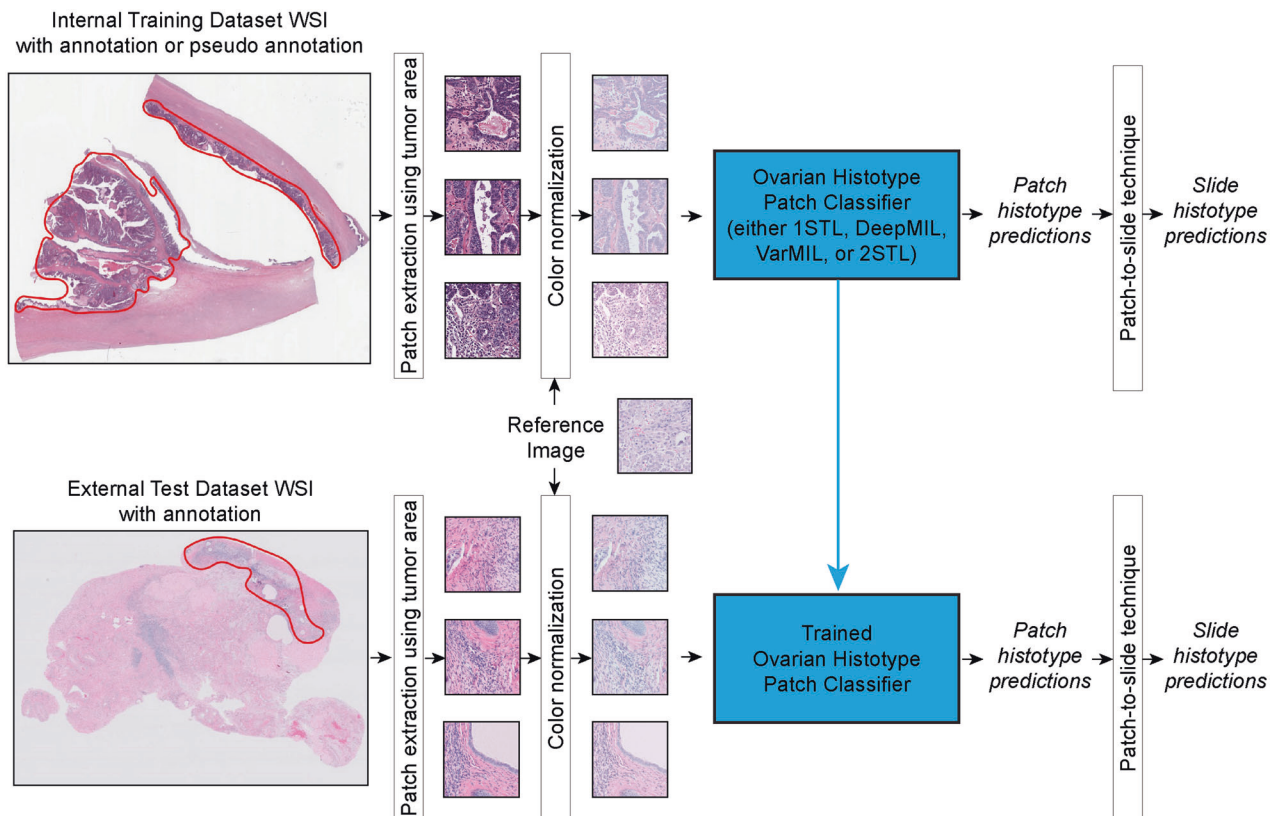


Fig. 1 Training and testing pipeline. Overall pipeline showing how the deep learning classifiers were trained and tested on color-normalized patches from the Internal Training Dataset, and then the trained models were tested on the External Test Dataset.

Table 1. Overview of datasets.

Dataset	Unit	Histotype				
		HGSC	CCOC	ENOC	LGSC	MUC
Internal Training Dataset (max 150 patches per patient, max 20,000 patches per histotype)	Patients	200	95	114	34	42
	Slides	410	167	237	69	65
	Annotations	185	67	74	50	50
	Pseudo annotations	225	100	163	19	15
	256 × 256 pixel patches at 10× magnification	19,998	19,998	19,998	10,109	9476
	512 × 512 pixel patches at 20× magnification	19,998	19,998	19,998	10,109	9618
External Test Dataset (max 150 patches per patient)	Patients	31	10	10	4	5
	Slides	31	10	10	4	5
	Annotations	31	10	10	4	5
	512 × 512 pixel patches at 20× magnification	4650	1500	1500	600	750

Training the deep learning-based histotype classifier

We compared four deep learning-based models (details outlined below) for ovarian carcinoma histotype classification. The models were initialized with ImageNet pre-trained weights, and then each were fine-tuned with the color-normalized patches from the Internal Training Dataset, using the ImageNet mean and standard deviation to normalize the RGB pixel values of each patch between -1 and 1 .

After training, the performance of the models were compared using the testing set of the respective Internal Training Dataset cross-validation split and the out-of-distribution External Test Dataset. In cases where there was a discrepancy between the reference diagnosis and the histotype diagnosed by artificial intelligence, two of the study pathologists (NS and CBG) reviewed the WSIs blinded to either the reference diagnosis or the AI diagnosis, and without access to the immunostaining results of the COSPv3 8-marker panel.

The GPU hardware used was either a Quadro RTX 500 (Nvidia, Santa Clara, CA, United States) or a Tesla V100-SXM2-32GB (Nvidia) based on availability.

One-stage transfer learning. For the first model, we used a one-stage transfer learning (1STL) algorithm. We implemented the PyTorch³⁰ VGG19³¹ model with a modified last layer for five-histotype classification. VGG19 is a popular convolutional neural network that uses smaller, typically 3×3 , filters in order to create a deeper network. Training was done with the patches of size 512×512 pixels at $20\times$ magnification using a batch size of 8 and AMSGrad optimization³² with 0.0002 learning rate. For each experiment, the model was fine-tuned for seven epochs, with the state having the best validation patch-level overall accuracy saved and used for testing.

Test slide-level results were calculated using majority vote from patch-level results (i.e., argmax on the counts of the different patch histotype predictions for each slide).

DeepMIL. The second architecture was DeepMIL²², a model that combines permutation-invariant multiple instance learning (MIL) with an attention-based neural network. MIL is a type of supervised learning where the labelled data (i.e., a WSI with a diagnosis) is broken up into a “bag of instances” which are considered to be weakly labelled (i.e., patches from the WSI, but each patch is not individually labelled). DeepMIL computes bag-level (WSI-level) features from attention-weighted patch instance feature vectors, and then classifies them with a fully connected layer. In our DeepMIL implementation, the patch feature vectors were extracted using the 1STL model trained on the same cross-validation split of 512×512 patches at $20\times$ magnification, using patch-level balanced accuracy as the metric for saving the best model state. Then DeepMIL was trained for 300 epochs with an initial learning rate of 0.0001; the learning rate was decreased by half if the validation loss did not decrease for 15 consecutive epochs, and training was stopped if the validation loss did not decrease for 30 epochs. The metric used for saving the best model state for testing was the slide-level overall accuracy. The test set slide-level results were calculated directly using this model.

VarMIL. Thirdly, we implemented VarMIL²¹, a model based on DeepMIL. One limitation of DeepMIL is that the bag-level latent features based on the attention-weighted instance vectors do not consider tile interactions or high-level features of the bag (WSI). VarMIL extends the architecture with an additional attention-weighted variance module to represent the tissue heterogeneity of the different tiles in a WSI. Just as our implementation of DeepMIL, we used trained 1STL models for patch-level feature extraction. We then trained VarMIL with the 512×512 patches at $20\times$ magnification and the same learning rate decay patience strategy as with DeepMIL. Slide-level overall accuracy was used to save the best state for testing. This model also calculated slide-level results directly.

Two-stage transfer learning. The final model we compared was a two-stage deep transfer learning (2STL) algorithm introduced for ovarian cancer histotype classification¹⁷. The first stage trains with patches of size 256×256 at $10\times$ magnification, while the second stage training and testing uses patches of size 512×512 at $20\times$ magnification. The rationale is that training with WSI patches of multiple sizes and magnifications gains context of the tissue at different perspectives. Each mini-batch is manipulated so that the same number of patches of each histotype is used during training. We used a batch size of 16 for the first stage and a batch size of 8 for the second stage. Each stage was trained with learning rate 0.0002 for 10 epochs, with the state with the highest patch-level overall balanced accuracy saved for the next stage of testing. The 3-fold cross validation scheme was modified by alternatively swapping the validation and test set for each training set, yielding a 6-fold scheme. The slide-level results were calculated by training a random forest classifier on the 6 patch-level cross-validation splits.

RESULTS

To find the model architecture with the best performance, we compared four deep learning networks for ovarian carcinoma histotype classification: a one-stage transfer learning algorithm (1STL), DeepMIL, VarMIL, and a two-stage transfer learning algorithm (2STL). Each model was trained and tested on patches from the large Internal Training Dataset for three cross-validation splits. The three trained models from cross-validation splits were then tested on the External Test Dataset from a different center.

Due to the distribution shift that exists between H&E datasets from different centers, even of the same tissue type, we focus on the results on the External Test Dataset. In order to effectively supplement pathologists with a machine learning-based ovarian carcinoma classifier, the model must be generalizable enough to work on WSIs from various locations. Our criterion for choosing the best model was the highest mean slide-level diagnostic concordance on the External Test Dataset.

Our results show that the highest performing model was 1STL (Table 2), which achieved a mean slide-level diagnostic

Table 2. Comparison of mean slide-level diagnostic concordance of four different deep learning architectures.

	1STL	DeepMIL	VarMIL	2STL
Internal Training Dataset	81.38 ± 0.02%	79.55 ± 0.05%	80.65 ± 0.04%	85.27 ± 0.02%
External Test Dataset	80.97 ± 0.03%	72.09 ± 0.05%	72.52 ± 0.03%	64.08 ± 0.09%

The bolded value highlights the best performing model as measured on the External Test Dataset.

Table 3. Detailed results of applying the 1STL model with color normalization and partially balanced dataset.

		HGSC Concordance	CCOC Concordance	ENOC Concordance	LGSC Concordance	MUC Concordance	Cohen's Kappa	F1 Score	AUC	Balanced Concordance
Internal Training Dataset	Model 1	82.02%	88.89%	81.63%	87.50%	54.55%	0.7393	0.7838	0.9404	78.92%
	Model 2	84.72%	81.25%	72.73%	100.00%	78.57%	0.7448	0.8188	0.9547	83.45%
	Model 3	82.95%	94.12%	67.31%	86.67%	77.78%	0.7294	0.7775	0.9473	81.76%
	Mean	83.23%	88.09%	73.89%	91.39%	70.30%	0.7378	0.7934	0.9475	81.38%
External Test Dataset	Model 1	93.55%	100.00%	50.00%	100.00%	80.00%	0.7985	0.8377	0.9599	84.71%
	Model 2	77.42%	100.00%	40.00%	100.00%	80.00%	0.6669	0.7124	0.9223	79.48%
	Model 3	93.55%	100.00%	70.00%	50.00%	80.00%	0.7988	0.8022	0.9586	78.71%
	Mean	88.17%	100.00%	53.33%	83.33%	80.00%	0.7547	0.7841	0.9469	80.97%
	Ensemble Model	93.55%	100.00%	50.00%	75.00%	80.00%	0.7722	0.8085	0.9592	79.71%

For the Internal Training Dataset, Models 1–3 refer to the models trained and tested on the three cross-validation splits of the Internal Training Dataset. For the External Test Dataset, Models 1–3 refer to the three models trained based on the cross-validation splits on the Internal Dataset and tested with the External Test Dataset.

Bolded values refer to either the mean or ensemble results across Models 1–3.

Table 4. Patient-level results of 1STL with color normalization and partially balanced dataset.

		HGSC Concordance	CCOC Concordance	ENOC Concordance	LGSC Concordance	MUC Concordance	Cohen's Kappa	Balanced Concordance
Internal Training Dataset	Model 1	86.66%	86.67%	78.94%	100.00%	57.14%	0.7607	81.88%
	Model 2	90.00%	100.00%	77.78%	100.00%	85.71%	0.8494	90.70%
	Model 3	87.50%	100.00%	78.95%	83.33%	85.71%	0.8300	87.10%
	Mean	88.05%	95.56%	78.56%	94.44%	76.19%	0.8134	86.56%

Models 1–3 refer to the models trained and tested on the three cross-validation splits of the Internal Training Dataset.

Bolded values refer to the mean results across Models 1–3.

concordance of 80.97 ± 0.03 % on the External Test Dataset. This model was trained with color normalized patches from the Internal Training Dataset that were partially balanced across the histotypes. We performed additional experiments for all the models, including no color normalization and a larger, unbalanced set of patches (see Supplementary Table S5), and found that this strategy yielded the best results on the External Test Dataset (see Supplementary Tables S6–S17 and Supplementary Figs. S3–S6). For completeness, we also calculated slide-level results for 1STL with random forest classifiers and for 2STL with majority vote, but these results did not affect our conclusion (see Supplementary Tables S8, S9, S14, S15). As well, we tested this best-performing model on patches from the External Test Dataset using the pseudo annotation classifier rather than pathologist annotations and found the exact same ensemble classifier results (see Supplementary Table S18).

Table 3 shows the performance of the 1STL model across the three cross-validation splits for the Internal Training Dataset. Furthermore, it shows the results for the External Test Dataset, where the three models trained based on the cross-validation splits (as separate raters) were applied to the External Test Dataset. In addition, we formed an ensemble classifier in which the three models (i.e., raters) in a majority voting strategy predicted the histotype. Using this strategy, we achieved a Cohen's kappa value of 0.77 in predicting histotypes which was better than the mean kappa value of 0.75 (Table 3).

Given that in clinical practice, multiple slides per tumor are examined to make a diagnosis, we asked whether our deep learning model would perform better when provided with multiple slides for a given tumor. Because we only had one slide per tumor in the External Dataset, we were only able to test this hypothesis in our Internal Training Dataset. Table 4 shows that the mean case-level concordance based on examination of multiple slides was higher than the slide-level results (86.56% (Table 4) versus 81.38% (Table 3)) when we used a majority voting strategy in which histotype was assigned based on the histotype diagnosis of the majority of slides.

Figure 2 shows the confusion matrix associated with the Ensemble Model (Table 3). We can see that the models generally struggled with classifying the ENOC slides of the External Test Dataset; even the ensemble classifier (Fig. 2) misdiagnosed half of the ENOC external test cases as HGSC or MUC. The 8 discrepant cases were independently reviewed by 2 of the authors (CBG, NS) blinded to the reference and AI diagnoses. Looking specifically at these 8 cases that were misclassified (Table 5 and Fig. 3), there are a variety of scenarios that could account for the misclassification. Cases A and D showed transitional pattern, an architecture that can be seen in either HGSC or ENOC, and IHC may be needed, as in these cases, for correct histotype diagnosis. Case B is a rare tumor in which the histotype is not clear, even after performing IHC, and arguably would best be classified as carcinoma NOS, as was done in the original cancer registry entry. Cases C, G, and H

		Actual				
		HGSC	CCOC	ENOC	LGSC	MUC
Predicted	HGSC	29	0	3	0	0
	CCOC	1	10	0	0	0
	ENOC	1	0	5	0	1
	LGSC	0	0	0	4	0
	MUC	0	0	2	0	4

Fig. 2 Confusion matrix of the overall histotype prediction of 1STL (by Ensemble Model) with color normalization and partially balanced dataset for the External Test Dataset.

are examples of the differential diagnosis between MUC and ENOC; when there is depletion of intracellular mucin, or slides showing the borderline areas are not available for review, this differential diagnosis is indeed challenging and IHC may be needed. Of these 3 cases, Case C additionally demonstrates the difficulty in distinction between CCOC and ENOC; this case was morphologically considered to be CCOC on independent review, and the IHC results, principally PR-, would support this view, although the reference and cancer registry diagnoses are of ENOC. In Case F, the differential diagnosis rests between CCOC and HGSC with clear cell change, a diagnostic challenge that can be resolved with IHC. For these 7 cases, the independent pathologist review agreed with the AI diagnosis in 4 cases, highlighting how histotypes of ovarian carcinoma can exhibit morphological mimicry, showing features on H&E that mimic other histotypes with respect to architecture and cytological features, at least focally. Only case E was a clear error by AI classification, with two of three models diagnosing HGSC when it is a classic low-grade ENOC.

This phenomenon of morphological mimicry is further demonstrated in Fig. 4, which visually illustrates the best-performing 1STL model on patch-level data. The predicted class of each patch generally makes sense, even for the incorrectly predicted tiles. For example, in Fig. 4D, the chosen patch that was misclassified as MUC has a structure that could be misconstrued for mucin, the defining morphological feature of MUC.

Table 5. Comparison of 3 AI models from 3 cross-validation splits in 1STL algorithm against gynecological pathologists' diagnoses.

Case	AI predicted diagnosis				Reference diagnosis	Independent study review diagnosis	Cancer Registry diagnosis	COSPV3 histotype prediction	Immunophenotype
	Model 1	Model 2	Model 3	Overall Majority of models					
A	HGSC	HGSC	HGSC	HGSC	ENOC	HGSC with transitional differentiation	Endometrioid	ENOC	WT1-;p53wt; PR-
B	HGSC	HGSC	ENOC	HGSC	ENOC	HGSC	NOS carcinoma	HGSC	WT1-;p53abn;PR-
C	MUC	MUC	CCOC	MUC	ENOC	CCOC; IHC needed for diagnosis	Endometrioid	ENOC	WT1-;p53wt; PR-;NapsinA-
D	ENOC	ENOC	ENOC	ENOC	HGSC	HGSC with transitional differentiation	Serous	HGSC	WT1+;p53abn;PR+
E	HGSC	HGSC	ENOC	HGSC	ENOC	ENOC	Endometrioid	ENOC	WT1-;p53wt; PR+
F	CCOC	CCOC	CCOC	CCOC	HGSC	CCOC; IHC needed for diagnosis	Clear cell	HGSC	WT1-;p53abn; PR-;NapsinA-
G	ENOC	ENOC	ENOC	ENOC	MUC	ENOC	Mucinous	MUC	PR-
H	MUC	MUC	MUC	MUC	ENOC	ENOC	Endometrioid	ENOC	PR-

The eight discrepant cases were independently reviewed by 2 of the authors (CBG, NS) blinded to the reference and AI diagnoses. COSPV3 Calculator of ovarian carcinoma subtype/histotype probability version 3, p53wt wild type immunostaining pattern for p53, p53abn mutant pattern immunostaining pattern for p53.

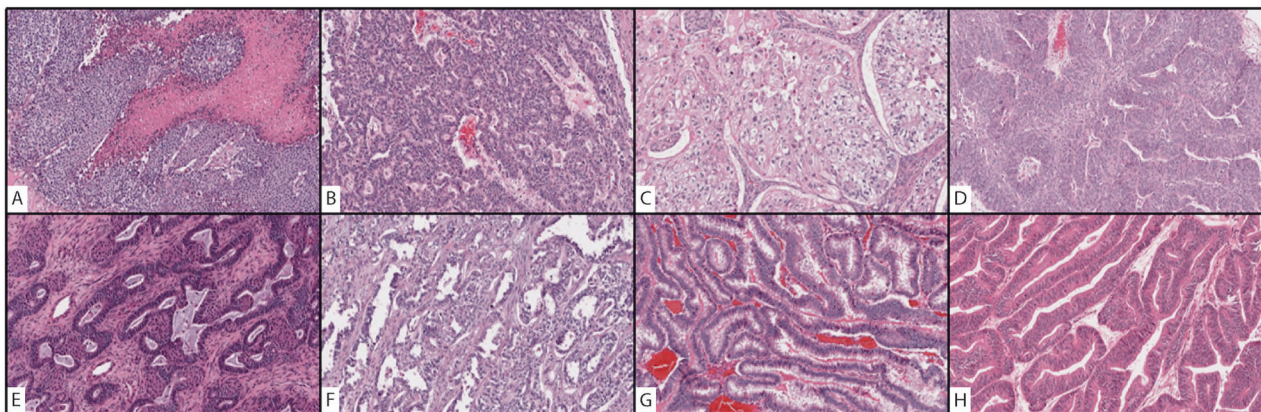


Fig. 3 Snapshots of the H&E slides for the 8 misclassified cases. Figures A–H correspond to the 8 discordant cases from the external dataset that were misclassified by AI; details of these cases are listed in Table 5.

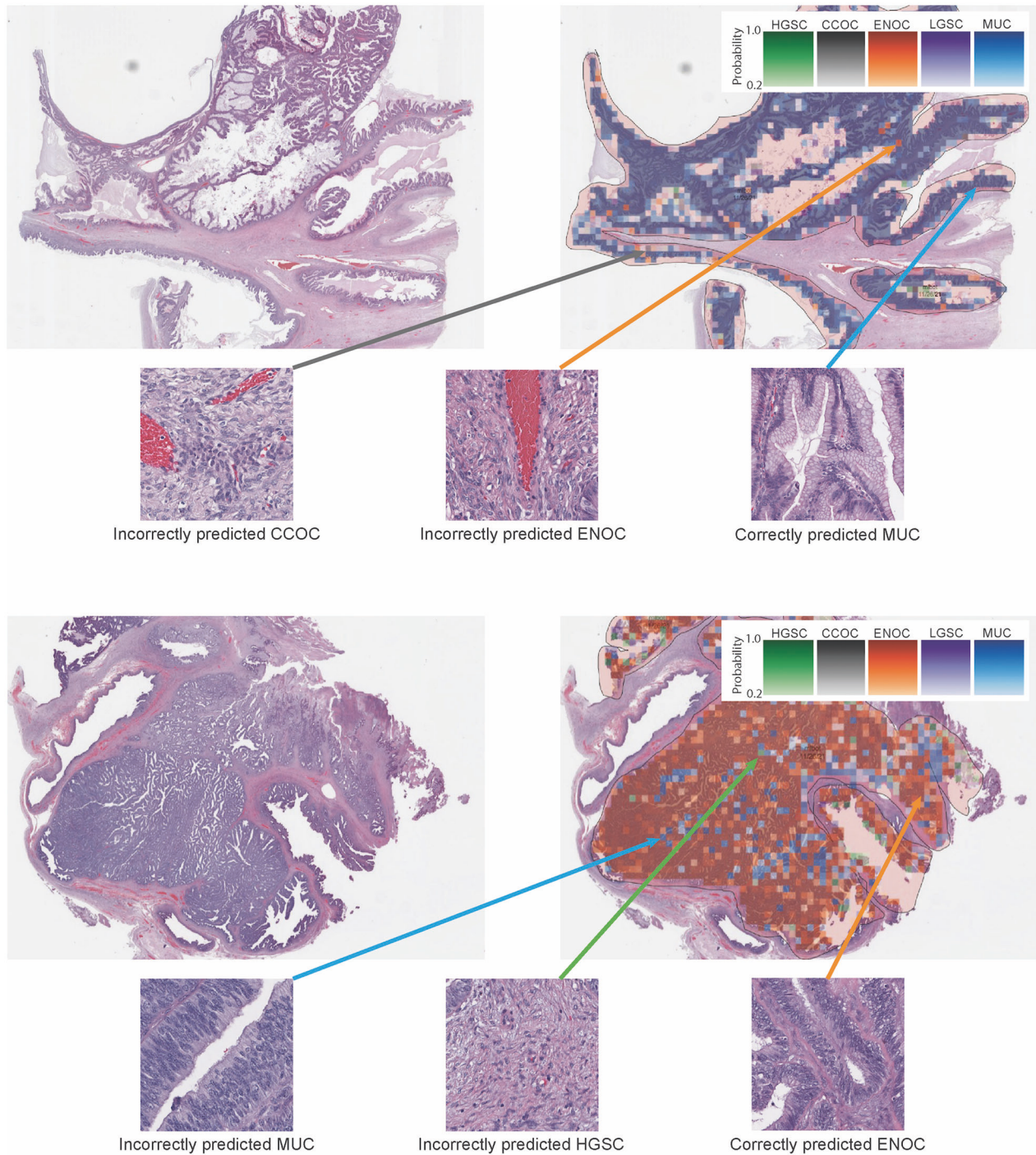


Fig. 4 Examples of predicted patches. **A** and **C** are original slides. **B** and **D** show the pathologist annotations overlaid with colors corresponding to the predicted class for each extracted patch.

We also tested the best performing 1STL model on 21 WSI samples of ovarian tumors that do not fall within the five common histotypes; although the prediction confidence was lower for these cases (see Supplementary Results and Supplementary Table S19), further work is needed to reliably detect and classify these other tumor types, however, this is outside the scope of the current investigation.

DISCUSSION

Our main objective in this work was to elucidate a generalizable (i.e., applicable to slides prepared in different laboratories)

machine learning-based strategy for improving ovarian carcinoma histotype diagnosis. We trained four different machine learning architectures with a variety of data engineering strategies and evaluated their performance on an external dataset. To the best of our knowledge, our training dataset of 948 WSIs is the largest collection of labelled ovarian carcinoma histotype images in published machine learning studies. Our chosen metric for comparing methods is average diagnostic concordance with expert integrated histotype diagnosis, based on consideration of both H&E morphology and IHC, using an external dataset from a different hospital to prove the

generalizability of our results on an independent test set different from the training set.

Our proposed models achieved a mean slide-level diagnostic concordance of $80.97 \pm 0.03\%$ using a one-stage deep transfer learning network (1STL). Based on our training set results, we expect the patient-level concordance (if given a test set with multiple slides per patient, which is typical of clinical diagnostic work) to be higher than the slide-level concordance. Our 1STL model also outperformed our implementations of DeepMIL²², VarMIL²¹, and a two-stage deep transfer learning network, which was previously used for ovarian cancer classification on a single dataset, trained and tested on 305 WSIs¹⁷. When using such a large dataset for training, we found that balancing the histotypes helps to prevent overfitting on any overrepresented classes. We also found that color normalization is essential for making a model generalizable for H&E images processed at different hospitals. We used a color normalization strategy that utilizes multiple normalization methods to create sets of images that are similar enough but variable enough to make the network robust²⁶, a promising strategy for overcoming the color inconsistencies of H&E images that has been a persistent problem for computer-aided diagnostics.

Wang et al. previously reported a high level of interobserver agreement using the two-stage transfer learning network, with better performance than general pathologists¹⁷. However, their ovarian cancer classifier performance was trained and tested on a single dataset of 305 WSIs. We have achieved an overall Cohen's kappa of 0.7722 on a test set stained and processed in a completely different location than the training set; this exceeds the inter-rater reliability by general pathologists (0.54–0.67 kappa)^{10,11} and approaches the level of expert pathologists with gynecologic pathology training (0.73–0.97 kappa)^{23,33}. Our generalizable strategy, which yields high performance on histopathology slides originating from other centers, is a further step towards the implementation of deep learning tools as a diagnostic adjunct for pathologists in diagnosing ovarian carcinoma histotype.

The “gold standard” for ovarian carcinoma histotype diagnosis is the integrated expert diagnosis, taking into account H&E morphology across all slides showing tumor, and using select molecular markers²³. There remain challenges in histotype diagnosis, however, as there can be discrepancy between H&E morphology and molecular markers, as assessed by immunohistochemistry²⁵. This is especially true for ENOC, where the differential diagnosis includes CCOC, MUC and HGSC²⁵, and where IHC data may be necessary for accurate histotype diagnosis. It is encouraging that the challenges in histotype diagnosis by AI are identical to those encountered by expert pathologists. Indeed, in 4 of 8 cases from the external dataset, the expert review pathologists rendered diagnoses, based on blind review of the WSIs classified by AI, that were in agreement with AI rather than the integrated reference diagnosis. Based on this we believe that, opportunities for improvement notwithstanding, the diagnostic algorithm presented is ready for validation studies in clinical practice, performing at a level comparable to an expert gynecological pathologist in formulating a favored histotype diagnosis based on H&E morphology. It is important to note, however, that the algorithm will misclassify cases, and these are exactly those cases that surgical pathologists must be aware of and resort to IHC in order to accurately diagnose histotype, e.g., cases with unusual features where ENOC is in the differential diagnosis. This algorithm cannot replace the function of diagnostic surgical pathologists to take into account all information in a case, beyond that present on the H&E stained slide, but can formulate a favored diagnosis with a high degree of diagnostic concordance, within seconds. We envision that such a tool could be used routinely in the setting of a fully digital surgical pathology service, as a diagnostic adjunct.

In conclusion, we demonstrate a deep learning strategy for ovarian carcinoma histotype classification based only on

histological features that is generalizable even on an externally stained test set. The performance is at a level that it could be implemented into practice, for validation. This approach holds potential as an adjunct for informing histotype diagnosis and in supporting histotype-specific ovarian cancer treatment.

DATA AVAILABILITY

The color normalization and deep learning codebase developed for this study will be made available through the following address upon publication: <https://github.com/AIMLab-UBC/>.

REFERENCES

- Siegel, RL, Miller, KD, Fuchs, HE, Jemal, A. Cancer Statistics, 2021. *CA Cancer J Clin* **71**, 7–33, <https://doi.org/10.3322/CAAC.21654> (2021).
- WHO Classification of Tumours Editorial Board. Female Genital Tumours. WHO Classification of Tumours (IARC Lyon, France, 2020).
- Köbel, M, Kalloger, SE, Boyd, N, McKinney, S, Mehl, E, Palmer, C, et al. Ovarian carcinoma subtypes are different diseases: implications for biomarker studies. *PLoS Med* **5**, 1749–1760 (2008)
- Gilks, CB, Oliva, E, Soslow, RA. Poor interobserver reproducibility in the diagnosis of high-grade endometrial carcinoma. *Am J Surg Pathol* **37**, 874–881 (2013)
- Coleman, RL, Fleming, GF, Brady, MF, Swisher, EM, Steffensen, KD, Friedlander, M, et al. Veliparib with first-line chemotherapy and as maintenance therapy in ovarian cancer. *N. Engl J Med* **381**, 2403–2415 (2019)
- Bartoletti, M, Musacchio, L, Giannone, G, Tuninetti, V, Bergamini, A, Scambia, G, et al. Emerging molecular alterations leading to histology-specific targeted therapies in ovarian cancer beyond PARP inhibitors. *Cancer Treat Rev* **101**, 102298 (2021)
- Han, G, Sidhu, D, Duggan, MA, Arseneau, J, Cesari, M, Clement, PB, et al. Reproducibility of histological cell type in high-grade endometrial carcinoma. *Mod Pathol* **26**, 1594–1604 (2013)
- Clarke, BA, Gilks, CB. Endometrial carcinoma: controversies in histopathological assessment of grade and tumour cell type. *J Clin Pathol* **63**, 410–415 (2010)
- Cohen, J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* **20**, 37–46 (1960)
- Patel, C, Harmon, B, Soslow, R, Garg, K, DeLair, D, Hwang, S, et al. Interobserver agreement in the diagnosis of ovarian carcinoma types: Impact of subspecialization. *Lab Invest* **92**, 292A–292A (2012)
- Köbel, M, Kalloger, SE, Lee, S, Duggan, MA, Kelemen, LE, Prentice, L, et al. Biomarker-based ovarian carcinoma typing: A histologic investigation in the ovarian tumor tissue analysis consortium. *Cancer Epidemiol Biomark Prev* **22**, 1677–1686 (2013)
- Lujan, G, Quigley, JC, Hartman, D, Parwani, A, Roehmholdt, B, Meter, B Van, et al. Dissecting the business case for adoption and implementation of digital pathology: a white paper from the Digital Pathology Association. *J Pathol Inf* **12**, 17 (2021)
- Coudray, N, Ocampo, PS, Sakellaropoulos, T, Narula, N, Snuderl, M, Fenyö, D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* **24**, 1559–1567 (2018)
- Farnell, DA, Huntsman, D, Bashashati, A. The coming 15 years in gynaecological pathology: digitisation, artificial intelligence, and new technologies. *Histopathology* **76**, 171–177 (2020)
- Campanella, G, Hanna, MG, Geneslaw, L, Mirafior, A, Werneck Krauss Silva, V, Busam, KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* **25**, 1301–1309, (2019)
- Bejnordi, BE, Veta, M, Van Diest, PJ, Van Ginneken, B, Karssemeijer, N, Litjens, G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017)
- Wang, Y, Farnell, D, Farahani, H, Nurse, M, Tessier-cloutier, B, Jones, SJM et al. Classification of epithelial ovarian carcinoma whole-slide pathology images using deep transfer learning. *Med Imaging Deep Learn* **2020**, 3–7, <https://doi.org/10.48550/arXiv.2005.10957> (2020).
- Yagi, Y. Color standardization and optimization in Whole Slide Imaging. *Diagn Pathol* **6**, 1–12 (2011)
- Bancroft JD, Gamble M. Theory and Practice of Histological Techniques (Elsevier Health Sciences, 2008)
- Lyon, HO, De Leenheer, AP, Horobin, RW, Lambert, WE, Schulte, EKW, Van Lie-dekerke, B, et al. Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains, and chromogenic reagents. *Histochem J* **26**, 533–544 (1994)
- Schirris, Y, Gavves, E, Nederlof, I, Horlings, HM, Teuwen, J, DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage

- response defect classification directly from H&E whole-slide images, arXiv preprint (2021), <https://doi.org/10.48550/arXiv.2107.09405>
22. Ilse, M, Tomczak, JM, Welling, M Attention-based deep multiple instance learning. *35th Int Conf Mach Learn* **5**, 3376–3391 (2018)
 23. Peres, LC, Cushing-Haugen, KL, Anglesio, M, Wicklund, K, Bentley, R, Berchuck, A, et al. Histotype classification of ovarian carcinoma: A comparison of approaches. *Gynecol Oncol* **151**, 53–60, (2018)
 24. Turkowski K, Filters for common resampling tasks (Academic Press Professional, Inc., USA, 1990)
 25. Köbel, M, Luo, L, Grevers, X, Lee, S, Brooks-Wilson, A, Gilks, CB, et al. Ovarian carcinoma histotype: strengths and limitations of integrating morphology with immunohistochemical predictions. *Int J Gynecol Pathol* **38**, 353–362 (2019)
 26. Boschman, J, Farahani, H, Darbandsari, A, Ahmadvand, P, Van Spankeren, A, Farnell, D, et al. The utility of color normalization for AI-based diagnosis of hematoxylin and eosin-stained pathology images. *J Pathol* **256**, 15–24 (2021)
 27. Reinhard, E, Adhikhmin, M, Gooch, B, Shirley, P Color transfer between images. *IEEE Comput Graph Appl* **21**, 34–41 (2001)
 28. Vahadane, A, Peng, T, Sethi, A, Albarqouni, S, Wang, L, Baust, M, et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med imaging* **35**, 1962–1971 (2016)
 29. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, X Guan, et al. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107–1110 (2009)
 30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* **32**, 8024–8035 (2019)
 31. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 770–778 (2016)
 32. Reddi, SJ, Kale, S, Kumar, S, On the Convergence of Adam and Beyond. arXiv preprint (2018), <https://doi.org/10.48550/arXiv.1904.09237>
 33. Köbel, M, Kalloger, SE, Baker, PM, Ewanowich, CA, Arseneau, J, Zherebitskiy, V, et al. Diagnosis of ovarian carcinoma cell type is highly reproducible: a transcanadian study. *Am J Surg Pathol* **34**, 984–993 (2010)

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of Thomas Kryton, BFA, Media Producer for scanning and processing the External Test Dataset.

AUTHOR CONTRIBUTIONS

AB, CBG, and NS conceived the study. JB, HF, and AB designed the machine learning models and experiments. JB, AD, and PA developed necessary computational and

visualization tools. JB and AD performed experiments. JB, HF, DF, CBG, NS, and AB interpreted and analyzed the data. JB, HF, CBG, NS, and AB wrote the first draft of the manuscript. DF, AZ, SJ, MK, DH, CBG, and NS contributed to data collection, slide annotation, computational analysis, and infrastructure, or provided pathology expertise. All authors critically reviewed the manuscript for important intellectual content and approved the final manuscript.

FUNDING

This work was supported by CIHR (No. 418734), NSERC (RGPIN-2019-04896), Cascadia Data Alliance, Michael Smith Foundation for Health Research Scholar Award, OVCARE Carrarsi, and UBC ObGyn funds.

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All experiments were conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects. Anonymized archival tissue samples were retrieved from the pathology archive at the BC Cancer Ovarian Care Research Program (OVCARE), University of British Columbia, and Vancouver General Hospital and were digitized after approval by the institutional ethics board (REB# H18-03646).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41379-022-01146-z>.

Correspondence and requests for materials should be addressed to C. Blake Gilks, Naveena Singh or Ali Bashashati.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.