REVIEW ARTICLE OPEN

Check for updates

Artificial intelligence in ovarian cancer histopathology: a systematic review

Jack Breen[1]✉, Katie Allen[2], Kieran Zucker[3], Pratik Adusumilli[2,4], Andrew Scarsbrook[2,4], Geoff Hall[3], Nicolas M. Orsi[2,5] and Nishant Ravikumar[1,5]

This study evaluates the quality of published research using artificial intelligence (AI) for ovarian cancer diagnosis or prognosis using histopathology data. A systematic search of PubMed, Scopus, Web of Science, Cochrane CENTRAL, and WHO-ICTRP was conducted up to May 19, 2023. Inclusion criteria required that AI was used for prognostic or diagnostic inferences in human ovarian cancer histopathology images. Risk of bias was assessed using PROBAST. Information about each model was tabulated and summary statistics were reported. The study was registered on PROSPERO (CRD42022334730) and PRISMA 2020 reporting guidelines were followed. Searches identified 1573 records, of which 45 were eligible for inclusion. These studies contained 80 models of interest, including 37 diagnostic models, 22 prognostic models, and 21 other diagnostically relevant models. Common tasks included treatment response prediction (11/80), malignancy status classification (10/80), stain quantification (9/80), and histological subtyping (7/80). Models were developed using 1–1375 histopathology slides from 1–776 ovarian cancer patients. A high or unclear risk of bias was found in all studies, most frequently due to limited analysis and incomplete reporting regarding participant recruitment. Limited research has been conducted on the application of AI to histopathology images for diagnostic or prognostic purposes in ovarian cancer, and none of the models have been demonstrated to be ready for real-world implementation. Key aspects to accelerate clinical translation include transparent and comprehensive reporting of data provenance and modelling approaches, and improved quantitative evaluation using cross-validation and external validations. This work was funded by the Engineering and Physical Sciences Research Council.

npj Precision Oncology (2023)7:83 ; https://doi.org/10.1038/s41698-023-00432-6

INTRODUCTION

Ovarian cancer is the eighth most common malignancy in women worldwide[1]. It is notoriously difficult to detect and diagnose, with ineffective screening[2] and non-specific symptoms similar to those caused by menopause[3]. Encompassing primary malignant tumours of the ovaries, fallopian tubes, and peritoneum, the disease has often started to spread within the abdomen at the time of diagnosis (FIGO[4] Stage 3). This typical late stage at diagnosis makes ovarian cancer a particularly deadly disease, with the 314,000 new cases diagnosed each year translating to 207,000 deaths per year globally[1].

Most ovarian cancers are carcinomas (cancers of epithelial origin) which predominantly fall into five histological subtypes: high-grade serous, low-grade serous, clear cell, endometrioid, and mucinous. Non-epithelial ovarian cancers are much less common and include germ cell, sex cord-stromal, and mesenchymal tumours. Ovarian cancer subtypes differ morphologically and prognostically and have varying treatment options[5]. High-grade serous carcinoma is the most common form of ovarian cancer, accounting for approximately 70% of all cases[6].

Histopathology, the examination of tissue specimens at the cellular level, is the gold standard for ovarian cancer diagnosis. Pathologists typically interpret tissue stained with haematoxylin and eosin (H&E), though interpretation can be a subjective, time-consuming process, with some tasks having a high level of inter-observer variation[7–9]. In the assessment of difficult cases, general pathologists may seek assistance from subspecialty gynaecological pathology experts, and/or use ancillary tests, such as immunohistochemistry (IHC). Referrals and ancillary testing can be essential to the accuracy of the diagnostic process but come at the cost of making it longer and more expensive. Worldwide, pathologists are in much greater demand than supply, with significant disparities in the number of pathologists between countries[10], and with better-supplied countries still unable to meet demand[11].

Traditionally, pathologists have analysed glass slides using a light microscope. However, the implementation of a digital workflow, where pathologists review scanned whole slide images (WSIs) using a computer, is becoming more common. While digital pathology uptake has likely been driven by efficiency benefits[12], it has created an opportunity for the development of automated tools to assist pathologists. These tools often aim to improve the accuracy, efficiency, objectivity, and consistency of diagnosis. Such tools could help to alleviate the global workforce shortage of pathologists, increasing diagnostic throughput and reducing the demand for referrals and ancillary tests. This is an increasingly active area of research[13] and, for some malignancies, these systems are starting to achieve clinical utility[14].

In this study, we systematically reviewed all literature in which artificial intelligence (AI) techniques (comprising both traditional machine learning (ML) and deep learning methods) were applied to digital pathology images for the diagnosis or prognosis of ovarian cancer. This included research that focused on a single diagnostic factor such as histological subtype and studies that performed computer-aided diagnostic tasks such as tumour

[1]Centre for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), School of Computing, University of Leeds, Leeds, UK. [2]Leeds Institute of Medical Research at St James's, School of Medicine, University of Leeds, Leeds, UK. [3]Leeds Cancer Centre, St James's University Hospital, Leeds, UK. [4]Department of Radiology, St James's University Hospital, Leeds, UK. [5]These authors jointly supervised this work: Nicolas M. Orsi, Nishant Ravikumar. ✉email: scjjb@leeds.ac.uk

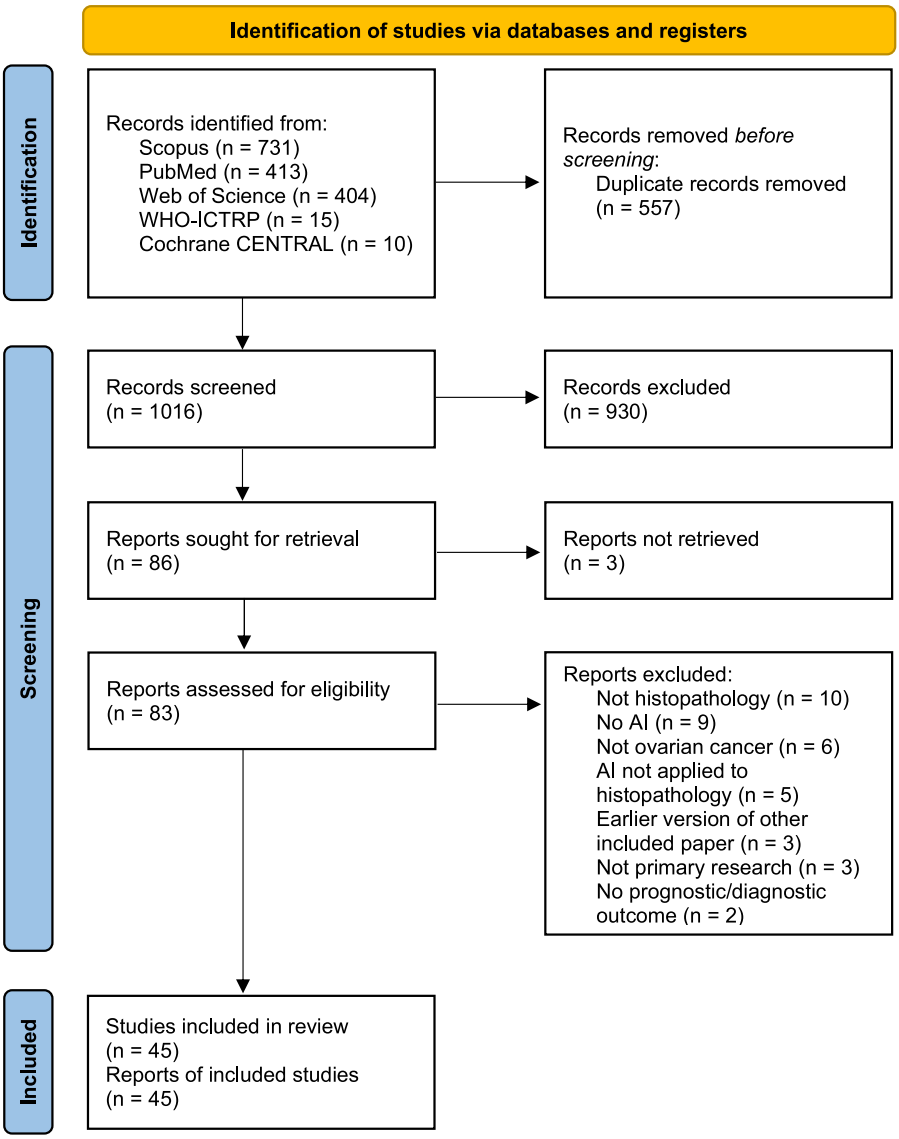

Fig. 1 PRISMA 2020 flowchart. PRISMA 2020 flowchart of the study identification and selection process for the systematic review. Records were screened on titles and abstracts alone, and reports were assessed based on the full-text content. *CENTRAL* Central Register of Controlled Trials. *WHO-ICTRP* World Health Organisation International Clinical Trial Registry Platform.

segmentation. The review characterises the state of the field, describing which diagnostic and prognostic tasks have been addressed, and assessing factors relevant to the clinical utility of these methods, such as the risks of bias. Despite ovarian cancer being a particularly difficult disease to detect and diagnose, and the shortage of available pathologists, AI models have not yet been implemented in clinical practice for this disease. This review aims to provide insights and recommendations based on published literature to improve the clinical utility of future research, including reducing risks of bias, improving reproducibility, and increasing generalisability.

RESULTS

As shown in Fig. 1, the literature searches returned a total of 1573 records, of which 557 were duplicates. Nine hundred and thirty records were excluded during the screening of titles and abstracts, and 41 were excluded based on full paper screening, including 3 records for which full articles could not be obtained. The remaining 45 studies were included in the review, of which 11 were conference papers and 34 were journal papers. All accepted studies were originally identified through searches of research databases, with no records from trial registries meeting the inclusion criteria. While the searches returned literature from as early as 1949, all of the research which met the inclusion criteria was published since 2010, with over 70% of the included literature published since 2020. Study characteristics are shown in Table 1. The 45 accepted articles contained 80 models of interest, details of which are shown in Table 2.

Risk of bias assessment

The results of the PROBAST assessments are shown in Table 3. While some studies contained multiple models of interest, none of these contained models with different risk of bias scores for any section of the PROBAST assessment, so one risk of bias analysis is presented per paper. All models showed either a high overall risk of bias (37/45) or an unclear overall risk of bias (8/45). Every high-risk model had a high-risk score in the analysis section (37/45), with several also being at high risk for participants (6/45), predictors (11/45), or outcomes (13/45). Less than half of the

Table 1. Characteristics of the 45 studies included in this systematic review.

Publication	Ovarian cancer data source	Models of interest	Outcome type	Model outcomes	Published code
Dong 2010[49]	Unclear	1	Other	Stain segmentation	None
Dong 2010[50]	Unclear	1	Other	Stain segmentation	None
Signolle 2010[51]	Unclear	1	Other	Tumour segmentation	None
Janowczyk 2011[52]	Unclear	1	Diagnosis	Malignancy	None
Janowczyk 2012[53]	Unclear	1	Other	Stain segmentation	None
Kothari 2012[18]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Malignancy	None
Poruthoor 2013[21]	TCGA-OV (Multi-city, USA)	2	Diagnosis, prognosis	Grade; overall survival	None
BenTaieb 2015[29]	Transcanadian Study (Multi-city, Canada)	1	Diagnosis	Histological subtype	None
BenTaieb 2016[30]	Transcanadian Study (Multi-city, Canada)	1	Diagnosis	Histological subtype	Inaccessible
BenTaieb 2017[48]	Unclear	1	Diagnosis	Histological subtype	Inaccessible
Lorsakul 2017[66]	Unclear	1	Other	Cell type	None
Du 2018[40]	Unique (Oklahoma, USA)	1	Other	Tissue type	None
Heindl 2018[57]	TCGA-OV (Multi-city, USA)	1	Other	Cell type	https://yuanlab.org/file/Ov3sweave2.pdf
Kalra 2020[15]	TCGA-OV (Multi-city, USA)	4	Diagnosis	Primary cancer type	None
Levine 2020[26]	OVCARE (Vancouver, Canada)	1	Diagnosis	Histological subtype	https://github.com/AIMLab-UBC/pathGAN
Yaar 2020[22]	TCGA-OV (Multi-city, USA)	1	Prognosis	Treatment response	https://github.com/asfandasfo/LUPI
Yu 2020[19]	TCGA-OV (Multi-city, USA)	4	Diagnosis, prognosis	Malignancy, grade, transcriptomic subtype; treatment response	https://github.com/khyu/ovarian_ca/
Gentles 2021[55]	Unique (Newcastle, UK)	6	Other	Stain quantity/intensity	None
Ghoniem 2021[23]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Stage	None
Jiang 2021[31]	Mayo Clinic (Rochester, USA)	1	Diagnosis	Malignancy	https://github.com/smujiang/CellularComposition
Laury 2021[56]	Unique (Helsinki, Finland)	1	Prognosis	Progression-free survival	None
Paijens 2021[37]	Unique (Groningen & Zwolle, The Netherlands)	1	Other	Tissue type	None
Shin 2021[38]	TCGA-OV (Multi-city, USA) + Unique (Ajou, Korea)	1	Diagnosis	Malignancy	https://github.com/ABMI/HistopathologyStyleTransfer
Zeng 2021[24]	TCGA-OV (Multi-city, USA) + Unique (Shanghai, China)	5	Diagnosis, prognosis	Genetic mutation, transcriptomic subtype, microsatellite instability; overall survival	None
Boehm 2022[17]	TCGA-OV (Multi-city, USA) + MSKCC (New York, USA)	3	Diagnosis, prognosis	Malignancy; overall survival, progression-free survival	https://github.com/kmboehm/onco-fusion
Boschman 2022[27]	OVCARE (Vancouver, Canada)	1	Diagnosis	Histological subtype	None
Elie 2022[61]	Unique (Caen, France)	3	Other	Stain quantity/intensity	None
Farahani 2022[28]	OVCARE (Vancouver, Canada) + Unique (Calgary, Canada)	2	Diagnosis	Malignancy, histological subtype	https://github.com/AIMLab-UBC/ModernPath2022
Hu 2022[41]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Epithelial–mesenchymal transition	https://github.com/superhy/LCSB-MIL
Jiang 2022[32]	Mayo Clinic (Rochester, USA)	4	Diagnosis, other	Tumour–stroma reaction; tumour segmentation	https://github.com/smujiang/TumorStromaReaction
Kasture 2022[46]	TCGA-OV[a] (Multi-city, USA)	1	Diagnosis	Histological subtype	https://github.com/kokilakasture/OvarianCancerPrediction
Kowalski 2022[47]	Unclear	1	Other	Tumour segmentation	None
Lazard 2022[42]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Homologous recombination deficiency status	https://github.com/trislaz/wsi_mil
Liu 2022[20]	TCGA-OV (Multi-city, USA)	1	Prognosis	Overall survival	https://github.com/RanSuLab/EOCprognosis
Mayer 2022[39]	TCGA-OV (Multi-city, USA) + Unique (Frankfurt, Germany)	1	Diagnosis	Malignancy	None

Table 1 continued

Publication	Ovarian cancer data source	Models of interest	Outcome type	Model outcomes	Published code
Nero 2022[44]	Unique (Rome, Italy)	2	Diagnosis, prognosis	Genetic mutation; relapse	None
Salguero 2022[67]	TCGA-OV (Multi-city, USA)	1	Diagnosis	Malignancy	None
Wang 2022[33]	Tri-Service (Taipei, Taiwan)	4	Prognosis	Treatment response	None
Wang 2022[34]	Tri-Service (Taipei, Taiwan)	1	Prognosis	Treatment response	None
Yokomizo 2022[43]	Unique (Tokyo, Japan)	3	Prognosis	Overall survival, progression-free survival, relapse	Inaccessible
Ho 2023[36]	MSKCC (New York, USA)	2	Diagnosis, other	Genetic mutation; tumour segmentation	https://github.com/MSKCC-Computational-Pathology/DMMN-ovary
Meng 2023[16]	Unique (Beijing, China)	1	Diagnosis	Malignancy	https://github.com/dreambamboo/STT-BOX-public
Ramasamy 2023[54]	TCGA-OV[a] (Multi-city, USA)	2	Diagnosis, other	Primary cancer type; tumour segmentation	None
Wang 2023[35]	Tri-Service (Taipei, Taiwan)	4	Prognosis	Treatment response	https://github.com/cwwang1979/OvaryTreatment_AnginPKM2VEGF
Wu 2023[45]	TCGA-OV (Multi-city, USA)	1	Prognosis	Overall survival	None

Details are shown for individual models in Table 2. Six data sources are used in multiple studies—The Cancer Genome Atlas (TCGA-OV)[25], the British Columbia Ovarian Cancer Research Program (OVCARE), The Transcanadian Study[72], and three individual centres (Mayo Clinic, Tri-Service, and Memorial Sloan Kettering Cancer Center (MSKCC)). Code is labelled as inaccessible where it could not be found despite a link being provided in the publication.
[a]Indicates papers where significant discrepancies were found regarding the data source, as described in the "Discussion".

studies achieved a low risk of bias in any domain (21/45), with most low risks being found in the outcomes (16/45) and predictors (9/45) sections. Nearly all of the papers had an unclear risk of bias in at least one domain, most commonly the participants (36/45) and predictors (25/45) domains. Qualitative summaries are presented in Fig. 2.

Data synthesis results

Data in included literature. The number of participants in internal datasets varied by orders of magnitude, with each study including 1–776 ovarian cancer patients, and one study including over 10,000 total patients across a range of 32 malignancies[15]. Most research only used data from the five most common subtypes of ovarian carcinoma, though one recent study included the use of sex cord-stromal tumours[16]. Only one study explicitly included any prospective data collection, and this was only for a small subset which was not used for external validation[17].

As shown in Fig. 3, the number of pathology slides used was often much greater than the number of patients included, with three studies using over 1000 slides from ovarian cancer patients[18–20]. In most of the studies, model development samples were WSIs containing resected or biopsied tissue (34/45), with others using individual tissue microarray (TMA) core images (5/45) or pre-cropped digital pathology images (3/45). Most studies used H&E-stained tissue (33/45) and others used a variety of IHC stains (11/45), with no two papers reporting the use of the same IHC stains. Some studies included multi-modal approaches, using genomics [17,21–24], proteomics[21,24], transcriptomics[24], and radiomics[17] data alongside histopathological data.

The most commonly used data source was The Cancer Genome Atlas (TCGA) (18/45), a project from which over 30,000 digital pathology images from 33 malignancies are publicly available. The ovarian cancer subset, TCGA-OV[25], contains 1481 WSIs from 590 cases of ovarian serous carcinoma (mostly, but not exclusively, high-grade), with corresponding genomic, transcriptomic, and clinical data. This includes slides from eight data centres in the United States, with most slides containing frozen tissue sections (1374/1481) rather than formalin-fixed, paraffin-embedded (FFPE) sections. Other recurring data sources were the University of British Columbia Ovarian Cancer Research Program (OVCARE) repository[26–28], the Transcanadian study[29,30], and clinical records at the Mayo Clinic[31,32], Tri-Service General Hospital[33–35], and Memorial Sloan Kettering Cancer Center[17,36]. All other researchers either used a unique data source (12/45) or did not report the provenance of their data (8/45). TCGA-OV, OVCARE, and the Transcanadian study are all multi-centre datasets. Aside from these, few studies reported the use of multi-centre data[17,24,28,37–39]. Only two studies reported the use of multiple slide scanners, with every slide scanned on one of two available scanners[27,28]. The countries from which data were sourced included Canada, China, Finland, France, Germany, Italy, Japan, the Netherlands, South Korea, Taiwan, the United Kingdom, and the United States of America.

Methods in included literature. There was a total of 80 models of interest in the 45 included papers, with each paper containing 1–6 such models. There were 37 diagnostic models, 22 prognostic models, and 21 other models predicting diagnostically relevant information. Diagnostic model outcomes included the classification of malignancy status (10/37), histological subtype (7/37), primary cancer type (5/37), genetic mutation status (4/37), tumour-stroma reaction level (3/37), grade (2/37), transcriptomic subtype (2/37), stage (1/37), microsatellite instability status (1/37), epithelial-mesenchymal transition status (1/37), and homologous recombination deficiency status (1/37). Prognostic models included the prediction of treatment response (11/23), overall survival (6/23), progression-free survival (3/23), and recurrence

Table 2. Characteristics of the 80 models of interest from the 45 papers included in this systematic review, grouped by model outcome.

Diagnosis outcome	Publication	Internal participants	Internal pathology images	Other data[a]	Stain type	Original image size	Patch size (pixels)	Magnification(s)	Feature extraction	Histopathological features	Final model	Prediction precision	Classes	Validation type	External validation data	Metric	Internal results	Internal variability (measure)	External results	External variability (measure)
Malignancy status	Janowczyk 2011[52]	6	11		IHC	1400x1400	Unclear	40x	Hand-crafted	Texture, cellular morphology	Probabilistic Boosting Tree	Patch	2 - Tumour, stroma	Monte Carlo cross-validation (5 reps)		AUC	0.8341			
	Kothari 2012[18]	571	1301		H&E	WSI	512x512	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	SVM	Patch	2 - Tumour, non-tumour	Single train/test split		Accuracy	90%			
	Yu 2020[23]	587	1375		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	2 - Malignant, benign	Monte Carlo cross-validation (3 reps)		AUC	0.975	±0.001 (unclear)		
	Jiang 2021[41]	30	≥30		H&E	WSI	512x512	40x	Hand-crafted	Colour, cellular and nuclear morphology	SVM	Patch	2 - HGSC, Serous borderline tumour	Unclear		Accuracy	90.64%			
																AUC	0.96			
	Shin 2021[38]	142	≥142		H&E	WSI	256x256	Unclear	Learned	CNN features (Inception V3)	CNN (Inception V3)	Patch	2 - Cancer, non-cancer	External validation	32 WSIs from different centre	Accuracy	98.3%		80.8%	
																AUC	0.998	0.995-0.999 (95% CI)	0.916	0.899–0.930 (95% CI)
	Boehm 2022[17]	283	≥283		H&E	WSI	128x128	Unclear	Learned	CNN features (ResNet18)	CNN (ResNet18)	Patch	4 - Tumour, stroma, fat, necrosis	4-fold cross-validation		Accuracy	88%			
	Farahani 2022[28]	≤416	416		H&E	WSI	512x512	20x	Learned	CNN features (ResNet18)	CNN (ResNet18)	Patch	2 - Tumour, stroma	3-fold cross-validation		Balanced accuracy	96.99%			
																AUC	0.9441			
	Mayer 2022[39]	≤101	101		H&E	WSI	512x512	Unclear	Learned	CNN features (ResNet18)	CNN Ensemble (ResNet18)	Patch	2 - Cancer, non-cancer	Monte Carlo cross-validation (10 reps) & external validation	41 WSIs from different centre	Accuracy per patient	56.3%-93.2%	Unclear plot (IQR & range)	Unclear plot	Unclear plot (IQR & Range)
	Salguero 2022[67]	18	≥18		H&E	WSI	100x100	40x	Hand-crafted	Colour, texture, cellular morphology	SVM	Patch	2 - Cancer, non-cancer	Single train/test split		Accuracy	73%			
	Meng 2023[16]	80	94		H&E	WSI	512x512	Unclear	Learned	CNN features (ResNet50)	CNN (novel STT-BOX)	WSI	2 - Malignant, benign	3-fold cross-validation (non-ovarian) & external validation (ovarian)	50 WSIs from 30 patients	AUC per subtype	0.9815-0.9953		0.8883	
Histological subtype	BenTaieb 2015[29]	80	80		H&E	WSI	Unclear	20x, 90x	Learned	CNN features (deconvolution network)	SVM	WSI	5 - HGSC, LGSC, CCC, MC, EC	Monte Carlo cross-validation (3 reps)		Accuracy	91.0%	±1.0% (unclear)		
																AUC	0.86			
	BenTaieb 2016[20]	80	80		H&E	WSI	500x500	20x, 40x	Hand-crafted	Colour, texture, cellular morphology, cytology	SVM	WSI	5 - HGSC, LGSC, CCC, MC, EC	Leave-one-patient-out cross-validation (5 reps)		Accuracy	95.0%	±1.5% (one SD)		
	BenTaieb 2017[48]	133	133		H&E	WSI	500x500	4x, 10x, 20x, 40x	Learned	CNN features (novel K-means)	SVM	WSI	5 - HGSC, LGSC, CCC, MC, EC	Single train/test split		Accuracy	90%			
	Levine 2020[26]	≤406	406		H&E	WSI	256x256	40x	Learned	CNN features (VGG19)	CNN (VGG19)	Patch	5 - HGSC, LGSC, CCC, MC, EC	Monte Carlo cross-validation (10 reps)		Accuracy	70.87%	±6.35% (one SD)		
																Balanced accuracy	75.15%	±10.44% (one SD)		
																AUC	0.9177	±2.30% (one SD)		
	Boschman 2022[27]	160	308		H&E	WSI	256x256	20x	Learned	CNN features (ResNet18)	CNN (ResNet18)	WSI	5 - HGSC, LGSC, CCC, MC, EC	External validation	60 WSIs from different centre	AUC	0.97	Unclear plot (unclear)	0.94	Unclear plot (unclear)
	Farahani 2022[28]	485	948		H&E	WSI	512x512	20x	Learned	CNN features (VGG19)	CNN (VGG19)	WSI	5 - HGSC, LGSC, CCC, MC, EC	3-fold cross-validation & external validation	60 WSIs from different centre	Balanced accuracy	81.38%		80.97%	
																AUC	0.9475		0.9469	
	Kasture 2022[46]	≤500	500		H&E	227x227	NA	20x	Learned	CNN features (novel KK-Net)	CNN (novel KK-Net)	Patch	5 - Serous, MC, CCC, EC, Non-cancer	10-fold cross-validation		Accuracy	91%			
																AUC	0.95			
Primary cancer type	Kalra 2020[25]	933	1039		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	4 - Ovarian, uterine carcinosarcoma, uterine endometrial, cervical (FFPE slides)	Leave-one-patient-out cross-validation		Accuracy (Ovarian)	66.98%			
	Kalra 2020[25]	1450	2216		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	4 - Ovarian, uterine carcinosarcoma, uterine endometrial, cervical (frozen slides)	Leave-one-patient-out cross-validation		Accuracy (Ovarian)	98.98%			
	Kalra 2020[25]	9484	11,561		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	13 - Gynaecological, brain, pulmonary, prostate/testis, breast, ... (FFPE slides)	Leave-one-patient-out cross-validation		Accuracy (Gynaecological)	68.86%			
	Kalra 2020[25]	10,571	14,887		H&E	WSI	1000x1000	20x	Learned	NNs features (unclear architectures)	Yottixel Search	WSI	13 - Gynaecological, brain, pulmonary, prostate/testis, breast, ... (frozen slides)	Leave-one-patient-out cross-validation		Accuracy (Gynaecological)	66.89%			

(2/23). The other models performed tasks that could be used to assist pathologists in analysing pathology images, including measuring the quantity/intensity of staining, generating segmentation masks, and classifying tissue/cell types.

A variety of models were used, with the most common types being convolutional neural network (CNN) (41/80), support vector machine (SVM) (10/80), and random forest (6/80). CNN architectures included GoogLeNet[40], VGG16[19,32], VGG19[26,28], InceptionV3[33–35,38], ResNet18[17,27,28,39,41,42], ResNet34[43], ResNet50[16,44,45], ResNet182[36], and MaskRCNN[32]. Novel CNNs typically used multiple standardised blocks involving convolutional, normalisation, activation, and/or pooling layers[22,46,47], with two studies also including attention modules[20,35]. One study generated their novel architecture by using a topology optimisation approach on a standard VGG16[23].

Most researchers split their original images into patches to be separately processed, with patch sizes ranging from 60×60 to 2048×2048 pixels, the most common being 512×512 pixels (19/56) and 256×256 pixels (12/56). A range of feature extraction techniques were employed, including both hand-crafted/pre-defined features (23/80) and features that were automatically learned by the model (51/80). Hand-crafted features included a plethora of textural, chromatic, and cellular and nuclear morphological features. Hand-

Table 2 continued

	Ramasamy 2023[54]	≤776	776		Unclear	WSI	Unclear	Unclear	Learned	CNN features (novel architecture)	CNN (novel)	WSI	2 - ovarian cancer, non-ovarian cancer	5-fold cross-validation		Accuracy	99.2%			
Genetic mutation status	Zeng 2021[24]	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	2 - BRCA1 Mutated, not mutated	Single train/test split		AUC	0.952			
	Zeng 2021[24]	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	2 - BRCA2 Mutated, not mutated	Single train/test split		AUC	0.912			
	Nero 2022[46]	664	664		H&E	WSI	256x256	Unclear	Learned	CNN features (ResNet50)	CNN (CLAM)	WSI	2 - BRCA1/2 Mutated, wild-type	Single train/test split		AUC	0.59			
	Ho 2023[26]	609	609		H&E	WSI	224x224	5x	Learned	CNN features (ResNet182)	CNN (ResNet182)	WSI	2 - BRCA1/2 Mutated, wild-type	Single train/test split		AUC	0.43			
Tumour-stroma reaction	Jiang 2022[32]	≤306	≤306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (VGG16)	Patch	3 - Low, intermediate, high (fibrosis score)	Single train/test split		Sensitivity per class	0.91-0.93			
	Jiang 2022[32]	≤306	≤306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (VGG16)	Patch	3 - Low, intermediate, high (cellularity score)	Single train/test split		Sensitivity per class	0.79-0.95			
	Jiang 2022[32]	≤306	≤306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (VGG16)	Patch	3 - Low, intermediate, high (orientation score)	Single train/test split		Sensitivity per class	0.74-0.95			
Grade	Poruthoor 2013[21]	387	≥387		H&E	WSI	512x512	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	SVM	WSI	2 - Grade 1-2, Grade 3-4	Monte Carlo cross-validation (15 reps)		Accuracy	88%	Unclear plot (one SD)		
Transcriptomic subtype	Yu 2020[19]	570	≤1358		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	2 - Low-to-moderate, high	Monte Carlo cross-validation (3 reps)		AUC	0.812	±0.088 (unclear)		
	Yu 2020[19]	553	≤1341		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	4 - Proliferative, differentiated, immunoreactive, mesenchymal	5-fold cross-validation		p-value	<0.0001			
	Zeng 2021[24]	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	4 - Proliferative, differentiated, immunoreactive, mesenchymal	Single train/test split		AUC per class	0.918-0.961			
Stage	Ghoniem 2021[23]	587	587	G	H&E	WSI	224x224	Unclear	Learned	CNN features (altered VGG16)	CNN (altered VGG16)	WSI	5 - I, II, III, IV, Not available	5-fold cross-validation (20 reps)		Accuracy	98.87%			
Microsatellite instability	Zeng 2021[24]	229	≥229		H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	3 - High instability, stable, NA	Single train/test split		AUC (High instability)	0.919			
																AUC (Stable)	0.924			
Epithelial-mesenchymal transition status	Hu 2022[41]	≤70	70		H&E	WSI	256x256	40x	Learned	CNN features (ResNet18)	CNN (novel adInter-MIL)	WSI	2 - High, low	Monte Carlo cross-validation (10 reps)		Balanced accuracy	85.45%	±0.48% (variance)		
																AUC	0.7455	±0.0043 (variance)		
HRD status	Lazard 2022[42]	≤90	90		H&E	WSI	224x224	20x	Learned	CNN features (ResNet18)	NN	WSI	2 - Homologous Recombination Deficient, Proficient	Unclear		AUC	0.73			

Prognosis outcome	Publication	Internal participants	Internal pathology images	Other data[a]	Stain type	Original image size	Patch size (pixels)	Magnification(s)	Feature extraction	Histopathological features	Final model	Prediction precision	Classes	Validation type	External validation data	Metric	Internal results	Internal variability (measure)	External results	External variability (measure)
Treatment response	Yaar 2020[22]	220	≥220	G	H&E	WSI	512x512	20x	Learned	CNN features (novel)	CNN	WSI	2 - Chemo-resistant, chemo-sensitive	5-fold cross-validation		AUC	0.79	±0.07 (one SD)		
	Yu 2020[19]	277	≤1065		H&E	WSI	Unclear	Unclear	Learned	CNN features (VGG16)	CNN (VGG16)	WSI	2 - Early relapse, late relapse	5-fold cross-validation		p-value	0.003			
	Wang 2022[25]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (AIM2 stain)	5-fold cross-validation		AUC	0.91	±0.05 (unclear)		
	Wang 2022[25]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (C3 stain)	5-fold cross-validation		AUC	0.78	±0.12 (unclear)		
	Wang 2022[25]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (C5 stain)	5-fold cross-validation		AUC	0.66	±0.07 (unclear)		
	Wang 2022[25]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (modified Inception V3)	CNN (modified Inception V3)	TMA	2 - Effective, invalid (NLRP3 stain)	5-fold cross-validation		AUC	0.55	±0.08 (unclear)		
	Wang 2022[26]	78	288		H&E	WSI	512x512	Unclear (multiple)	Learned	CNN features (Inception V3)	CNN (Inception V3)	WSI	2 - Effective, invalid	5-fold cross-validation & external validation	175 TMAs from 71 patients	Accuracy	88.2%	±6% (unclear)	77.5%	
	Wang 2023[35]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN (InceptionV3)	TMA	2 - Effective, invalid (PKM2 stain)	5-fold cross-validation		AUC	0.99	±0.01 (unclear)		
	Wang 2023[35]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN (InceptionV3)	TMA	2 - Effective, invalid (Ang-2 stain)	5-fold cross-validation		AUC	1.00	±0.01 (unclear)		
	Wang 2023[35]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN (InceptionV3)	TMA	2 - Effective, invalid (VEGF stain)	5-fold cross-validation		AUC	0.89	±0.08 (unclear)		
	Wang 2023[35]	≤180	180		IHC	TMA	512x512	20x	Learned	CNN features (novel)	CNN Ensemble (InceptionV3)	TMA	2 - Effective, invalid (PKM2+Ang-2 stain)	5-fold cross-validation		AUC	1.00	±0.00 (unclear)		
Overall survival	Poruthoor 2013[21]	382	≥382	G,P	H&E	WSI	512x512	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	SVM	WSI	2 - <5 years, ≥5 years	Monte Carlo cross-validation (15 reps)		Accuracy	55%	Unclear plot (one SD)		

crafted features were commonly used as inputs to classical ML methods, such as SVM and random forest models. Learned features were typically extracted using a CNN, which was often also used for classification.

Despite the common use of patches, most models made predictions at the WSI level (29/80), TMA core level (18/80), or patient level (6/80), requiring aggregation of patch-level information. Two distinct aggregation approaches were used, one aggregating before modelling and one aggregating after modelling. The former approach requires the generation of slide-level features before modelling, the latter requires the aggregation of patch-level model outputs to make slide-level predictions. Slide-level features were generated using summation[16], averaging[21,24,36], attention-based weighted averaging[20,41,42,44,45], concatenation[15,30], as well as more complex embedding approaches using Fisher vector encoding[29] and k-means clustering[48]. Patch-level model outputs were aggregated to generate slide-level predictions by taking the maximum[22,35], median[43], or average[23], using voting strategies[27,34], or using a random forest classifier[28]. These approaches are all examples of *multiple instance learning* (MIL), though few models of interest were reported using this terminology[22,41,42,44].

Table 2 continued

	Zeng 2021[24]	229	≥229	G,P,T	H&E	WSI	1000x1000	Unclear	Hand-crafted	Texture, cellular and nuclear morphology	Random Forest	Patient	2 - High risk, low risk	External validation	TMAs from 92 patients	Hazard ratio p-value	18.23 <0.0001	10.34-34.38 (95% CI)	0.0097	
	Boehm 2022[17]	444	≥283	R	H&E	WSI	128x128	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	Cox model	WSI	Risk score	Single train/test split		C-index	0.61	0.594-0.625 (95% CI)		
	Liu 2022[20]	583	1296		H&E	WSI	512x512	20x	Learned	CNN features (novel DeepConvAttentionSurv)	CNN (novel DCAS)	Patient	Risk score	Single train/test split		C-index	0.98	±0.0085 (unclear)		
	Yokomizo 2022[43]	110	≥110		H&E	WSI	255x255	Unclear	Learned	CNN features (ResNet34)	CNN (ResNet34)	TMA	2 - Short, long	Monte Carlo cross-validation (10 reps)		AUC	"~0.95"			
	Wu 2023[45]	90	90		H&E	WSI	256x256	Unclear	Learned	CNN features (ResNet50)	CNN (CLAM)	TMA	Risk score	5-fold cross-validation		C-index	0.5789	0.5096-0.6053 (CV range)		
Progression-free survival	Laury 2021[56]	52	227		H&E	WSI	Unclear	Unclear	Learned	CNN features (unclear architecture)	NN	WSI	2 - <6 months, >18 months	Single train/test split		Accuracy	82%			
	Boehm 2022[17]	422	≥261	G,R	H&E	WSI	128x128	Unclear	Hand-crafted	Colour, texture, cellular and nuclear morphology	Cox model	WSI	2 - High, low	Single train/test split		p-value	0.040			
	Yokomizo 2022[43]	110	≥110		H&E	WSI	255x255	Unclear	Learned	CNN features (ResNet34)	CNN (ResNet34)	TMA	2 - Short, long	Monte Carlo cross-validation (10 reps)		AUC	0.98			
Relapse	Nero 2022[44]	656	656		H&E	WSI	256x256	Unclear	Learned	CNN features (ResNet50)	CNN (CLAM)	WSI	2 - Relapse/progression, no relapse/progression	Single train/test split		AUC	0.714			
	Yokomizo 2022[43]	110	≥110		H&E	WSI	255x255	Unclear	Learned	CNN features (ResNet34)	CNN (ResNet34)	TMA	2 - Recurrent, non-recurrent	Monte Carlo cross-validation (10 reps)		AUC	0.98			
Other outcome	**Publication**	**Internal participants**	**Internal pathology images**	**Other data[a]**	**Stain type**	**Original image size**	**Patch size (pixels)**	**Magnification(s)**	**Feature extraction**	**Histopathological features**	**Final model**	**Prediction precision**	**Classes**	**Validation type**	**External validation data**	**Metric**	**Internal results**	**Internal variability (measure)**	**External results**	**External variability (measure)**
Stain quantity/intensity	Gentles 2021[55]	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	ATM stain H-score (0-18)	Single test set		R^2	0.1833			
	Gentles 2021[55]	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	ATR stain H-score (0-18)	Single test set		R^2	0.4330			
	Gentles 2021[55]	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	DNAPKcs stain H-score (0-18)	Single test set		R^2	0.6296			
	Gentles 2021[55]	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	Ku70 stain H-score (0-18)	Single test set		R^2	0.5891			
	Gentles 2021[55]	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	PAR stain H-score (0-18)	Single test set		R^2	0.3978			
	Gentles 2021[55]	33	≥66		IHC	TMA	NA	20x	Unclear	Unclear	Genie Classifier	TMA	RPA stain H-score (0-18)	Single test set		R^2	0.4453			
	Elie 2022[41]	25	25		IHC	WSI	Unclear	20x	Hand-crafted	Colour, texture	Gaussian Mixture Model	Patch	3 - Mcl-1 high, medium, low	None		Accuracy per patient	96.94%-99.51%			
	Elie 2022[41]	25	25		IHC	WSI	Unclear	20x	Hand-crafted	Colour, texture	Gaussian Mixture Model	Patch	3 - Bim high, medium, low	None		Accuracy per patient	92.77%-95.75%			
	Elie 2022[41]	25	25		IHC	WSI	Unclear	20x	Hand-crafted	Colour, texture	Gaussian Mixture Model	Patch	3 - P-ERK high, medium, low	None		Accuracy per patient	89.08%-100%			
Tumour segmentation	Signolle 2010[61]	Unclear	Unclear		IHC	WSI	2048x2048	20x	Hand-crafted	Texture	Hidden Markov Tree	Pixel	5 - Cancer, inflammatory stroma, loose connective tissue, cellular stroma, background	Single train/test split		Accuracy	71.50%	±12.83 (one SD)		
	Jiang 2022[62]	306	306		H&E	WSI	256x256	Unclear	Learned	CNN features (Mask-RCNN)	CNN (Mask-RCNN)	Pixel	2 - Tumour, stroma	Single train/test split		Dice coefficient	93.5%	Unclear plot (unclear)		
	Kowalski 2022[47]	≤26	26		H&E	1698x1242	100x200	Unclear	Learned	CNN features (novel architecture)	CNN (novel)	Pixel	2 - Cancer, healthy	Single train/test split		Accuracy	82%			
	Ho 2023[36]	39	39		H&E	WSI	256x256	5x, 10x, 20x	Learned	CNN features (novel architecture)	CNN (novel DMMN)	Pixel	2 - Cancer, non-cancer	Single train/test split		Intersection over union	0.74			
	Ramasamy 2023[54]	≤776	776		Unclear	WSI	Unclear	Unclear	Learned	CNN features (novel architecture)	CNN (novel)	Pixel	2 - Tumour, non-tumour	5-fold cross-validation		Dice coefficient	92%			
Stain segmentation	Dong 2010[49]	1	1		IHC	Unclear	NA	Unclear	Hand-crafted	Colour	ISODATA clustering	Pixel	2 - Positive, Negative	None		Qualitative	"Satisfactory"			
	Dong 2010[50]	1	1		IHC	Unclear	NA	Unclear	Hand-crafted	Colour	OTSU thresholding	Pixel	2 - Positive, Negative	None		Qualitative	"Satisfactory"			
	Janowczyk 2012[53]	100	≥500		IHC	TMA	NA	20x	Hand-crafted	Colour	HNCuts (novel)	Pixel	2 - Positive, Negative	Single test set		Sensitivity Specificity	59.24% 99.01%	±7.36% (variance) ±0.56% (variance)		
Cell type	Lorsakul 2017[66]	≤45	45		IHC	WSI	Unclear	20x	Hand-crafted	Nuclear morphology	Random Forest	Cell	5 - Cancer, carcinoma-associated fibroblast, non-tumour, background, artifact	5-fold cross-validation		Accuracy	91.7%			
	Heindl 2018[57]	514	514		H&E	WSI	2000x2000	Unclear	Hand-crafted	Texture, cellular morphology	SVM	Cell	3 - Cancer, stroma, lymphocyte	Single train/test split		Balanced accuracy per class	80.64%-85.05%			
Tissue type	Du 2018[40]	≤154	154		H&E	Unclear	60x60	Unclear	Learned	CNN features (GoogLeNet)	SVM	Superpixel	2 - Epithelium, stroma	Single train/test split		Accuracy AUC	91.8% 0.974			
	Paijens 2021[37]	268	268		IHC	TMA	Unclear	Unclear	Learned	NN features (unclear architecture)	NN	Pixel	2 - Epithelium, stroma	None		None	NA			

SVM support vector machine, *CNN* convolutional neural network, *AUC* area under the receiver operating characteristic (ROC) curve, *HGSC* high-grade serous carcinoma, *LGSC* low-grade serous carcinoma, *CCC* clear cell carcinoma, *MC* mucinous carcinoma, *EC* endometrioid carcinoma, *H&E* haematoxylin and eosin, *IHC* immunohistochemistry, *TMA* individual cores from tissue microarrays, *WSI* whole slide images of biopsy or resection specimens.
[a]Other data types are Genomics (G), Proteomics (P), Radiomics (R), and Transcriptomics (T).

Most studies included segmentation at some stage, with many of these analysing tumour/stain segmentation as a model outcome[32,36,37,47,49–54]. Some other studies used segmentation to determine regions of interest for further modelling, either simply separating tissue from background[15,18,44,45], or using tumour segmentation to select the most relevant tissue regions[33–35,55,56]. One study also used segmentation to detect individual cells for classification[57]. Some studies also used segmentation in determining hand-crafted features relating to the quantity and morphology of different tissues, cells, and nuclei[17,18,21,24,30,31].

While attention-based approaches have been applied to other malignancies for several years[58,59], they were only seen in the most recent ovarian cancer studies[20,28,33–35,41,42,44,45], and none of the

Table 3. PROBAST risk of bias assessment results for the 45 papers included in this review.

Publication	Participants	Predictors	Outcome	Analysis	Overall
Dong 2010[49]	High	High	High	High	High
Dong 2010[50]	High	High	High	High	High
Signolle 2010[51]	Unclear	Unclear	High	High	High
Janowczyk 2011[52]	Unclear	Unclear	Low	High	High
Janowczyk 2012[53]	Unclear	High	Unclear	High	High
Kothari 2012[18]	Unclear	Low	Low	Unclear	Unclear
Poruthoor 2013[21]	Unclear	High	High	High	High
BenTaieb 2015[29]	Unclear	Unclear	Low	High	High
BenTaieb 2016[30]	Unclear	High	Unclear	High	High
BenTaieb 2017[48]	Unclear	Unclear	Low	High	High
Lorsakul 2017[66]	Unclear	Unclear	High	High	High
Du 2018[40]	Unclear	Unclear	Unclear	Unclear	Unclear
Heindl 2018[57]	Unclear	Low	Low	High	High
Kalra 2020[15]	Unclear	Low	Low	High	High
Levine 2020[26]	Unclear	Low	Low	Unclear	Unclear
Yaar 2020[22]	Unclear	Unclear	Low	High	High
Yu 2020[19]	Unclear	Low	Low	High	High
Gentles 2021[55]	High	Unclear	High	High	High
Ghoniem 2021[23]	Unclear	Unclear	Unclear	High	High
Jiang 2021[31]	High	High	Unclear	High	High
Laury 2021[56]	Low	High	High	High	High
Paijens 2021[37]	Low	High	Unclear	High	High
Shin 2021[38]	Unclear	Unclear	Unclear	High	High
Zeng 2021[24]	Unclear	Unclear	Low	High	High
Boehm 2022[17]	Unclear	High	Unclear	High	High
Boschman 2022[27]	Unclear	Low	Low	High	High
Elie 2022[61]	Unclear	Low	High	High	High
Farahani 2022[28]	Unclear	Unclear	Low	Unclear	Unclear
Hu 2022[41]	Unclear	Unclear	Unclear	Unclear	Unclear
Jiang 2022[32]	Unclear	Unclear	High	High	High
Kasture 2022[46]	High	High	High	High	High
Kowalski 2022[47]	Unclear	Unclear	Unclear	High	High
Lazard 2022[42]	Unclear	Unclear	Unclear	Unclear	Unclear
Liu 2022[20]	Unclear	Unclear	Unclear	Unclear	Unclear
Mayer 2022[39]	Unclear	Unclear	High	High	High
Nero 2022[44]	Unclear	Low	High	High	High
Salguero 2022[67]	Unclear	Unclear	Low	High	High
Wang 2022[33]	Unclear	Unclear	Unclear	High	High
Wang 2022[34]	Unclear	Unclear	Low	High	High
Yokomizo 2022[43]	Low	Low	Unclear	Unclear	Unclear
Ho 2023[36]	Unclear	Unclear	Unclear	High	High
Meng 2023[16]	Unclear	Unclear	Low	High	High
Ramasamy 2023[54]	High	High	High	High	High
Wang 2023[35]	Unclear	Unclear	Unclear	High	High
Wu 2023[45]	Unclear	Unclear	Low	High	High

This is presented as one row for each paper because every paper that contained multiple models of interest was found to have the same risk of bias for every model.

methods included self-attention, an increasingly popular method for other malignancies[60]. Most models were deterministic, though hidden Markov trees[51], probabilistic boosting trees[52], and Gaussian mixture models[61] were also used. Aside from the common use of low-resolution images to detect and remove non-tissue areas, images were typically analysed at a single resolution, with only six papers including multi-magnification techniques in their models of interest. Four of these combined features from different resolutions for modelling[29,30,36,48], and the other two used different magnifications for selecting informative tissue regions and for modelling[33,34]. Out of the papers for which it could be determined, the most common modelling magnifications were ×20 (35/41) and ×40 (7/41). Few models integrated histopathology data with other modalities (6/80). Multi-modal approaches included the concatenation of separately extracted uni-modal features before modelling[21,23,24], the amalgamation of uni-modal predictions from separate models[17], and a teacher–student approach where multiple modalities were used in model training but only histopathology data was used for prediction[22].

Analysis in included literature. Analyses were limited, with less than half of the model outcomes being evaluated with cross-validation (39/80) and with very few externally validated using independent ovarian cancer data (7/80), despite small internal cohort sizes. Cross-validation methods included *k*-fold (22/39) with 3–10 folds, Monte Carlo (12/39) with 3–15 repeats, and leave-one-patient-out cross-validations (5/39). Some other papers included cross-validation on the training set to select hyperparameters but used only a small unseen test set from the same data source for evaluation. Externally validated models were all trained with WSIs, with validations either performed on TMA cores (2/7) or WSIs from independent data sources (5/7), with two of these explicitly using different scanners to digitise internal and external data[27,28]. Some reported methods were externally validated with data from non-ovarian malignancies, but none of these included ovarian cancer data in any capacity, so were not included in the review. However, there was one method which trained with only gastrointestinal tumour data and externally validated with ovarian tumour data[16].

Most classification models were evaluated using accuracy, balanced accuracy, and/or area under the receiver operating characteristic curve (AUC), with one exception where only a p-value was reported measuring the association between histological features and transcriptomic subtypes based on a Kruskal–Wallis test[19]. Some models were also evaluated using the F1-score, which we chose not to tabulate (in Fig. 3) as the other metrics were reported more consistently. Survival model performance was typically reported using AUC, with other metrics including p-value, accuracy, hazard ratios, and C-index, which is similar to AUC but can account for censoring. Segmentation models were almost all evaluated differently from each other, with different studies reporting AUC, accuracy, Dice coefficient, intersection over union, sensitivity, specificity, and qualitative evaluations. Regression models were all evaluated using the coefficient of determination (R^2-statistic). For some models, performance was broken down per patient[39,61], per subtype[16], or per class[15,24,32,57], without an aggregated, holistic measure of model performance.

The variability of model performance was not frequently reported (33/94), and when it was reported it was often incomplete. This included cases where it was unclear what the intervals represented (95% confidence interval, one standard deviation, variation, etc.), or not clear what the exact bounds of the interval were due to results being plotted but not explicitly stated. Within the entire review, there were only three examples in which variability was reported during external validation[27,38,39], only one of which clearly reported both the bounds and the type of the interval[38]. No studies performed any Bayesian form of uncertainty quantification. Reported results are shown in Table 2, though direct comparisons between the performance of different models should be treated with caution due to the diversity of data and validation methods used to evaluate different models, the lack of variability measures, the consistently high risks of bias, and the heterogeneity in reported metrics.

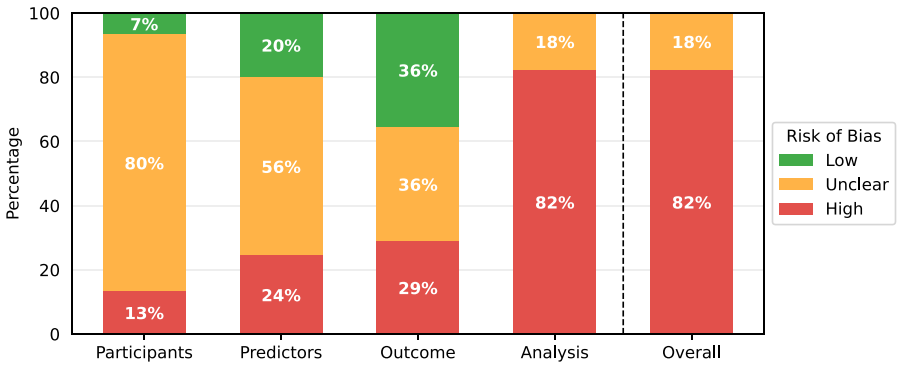

Fig. 2 PROBAST risk of bias results. PROBAST risk of bias results summarised for the 45 papers included in this review.

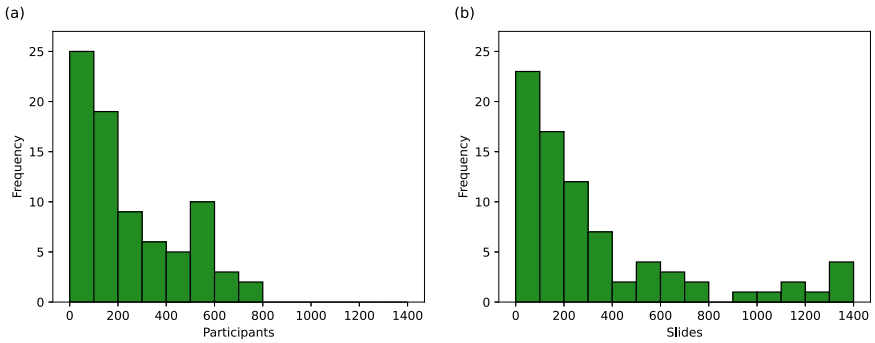

Fig. 3 Number of patients and slides per model. Histograms showing the number of **a** ovarian cancer patients and **b** ovarian cancer histopathology slides used in model development. Many of these values are uncertain due to incomplete reporting, as reflected in Table 2.

DISCUSSION

The vast majority of published research on AI for diagnostic or prognostic purposes in ovarian cancer histopathology was found to be at a high risk of bias due to issues within the analyses performed. Researchers often used a limited quantity of data and conducted analyses on a single train-test data split without using any methods to account for overfitting and model optimism (cross-validation, bootstrapping, external validation). These limitations are common in gynaecological AI research using other data types, with recent reviews pointing to poor clinical utility caused by predominantly retrospective studies using limited data[62,63] and limited methodologies with weak validation, which risk model performance being overestimated[64,65].

The more robust analyses included one study in which several relevant metrics were evaluated using 10 repeats of Monte Carlo cross-validation on a set of 406 WSIs, with standard deviations reported for each metric[26]. Other positive examples included the use of both internal cross-validation and external validation for the same outcome, giving a more rigorous analysis[28,34,39]. While external validations were uncommon, those which were conducted offered a real insight into model generalisability, with a clear reduction in performance on all external validation sets except one[28]. The only study which demonstrated high generalisability included the largest training set out of all externally validated approaches, included more extensive data labelling than many similar studies, and implemented a combination of three colour normalisation approaches, indicating that these factors may benefit generalisability.

Studies frequently had an unclear risk of bias within the participants and predictors domains of PROBAST due to incomplete reporting. Frequently missing information included where the patients were recruited, how many patients were included, how many samples/images were used, whether any patients/images were excluded, and the methods by which tissue was processed and digitised. Reporting was often poor regarding open-access datasets. Only three papers were found to be at low risk of bias for participants, with these including clear and reasonable patient recruitment strategies and selection criteria, which can be seen as positive examples for other researchers[37,43,56]. Information about the predictors (histopathology images and features derived thereof) was generally better reported, but still often missed key details which meant that it was unclear whether all tissue samples were processed similarly to avoid risks of bias from visual heterogeneity. It was found that when patient characteristics were reported, they often showed a high risk of bias. Many studies included very small quantities of patients with specific differences from the majority (e.g. less than 20 patients with a different cancer subtype to the majority), causing a risk of spurious correlations and results which are not generalisable to the wider population.

Reporting was particularly sparse in studies which used openly accessible data, possibly indicating that AI-focused researchers were not taking sufficient time to understand these datasets and ensure their research was clinically relevant. For example, many of the researchers who used TCGA data included frozen tissue sections without commenting on whether this was appropriate, despite the fact that pathologists do not consider them to be of optimal diagnostic quality. One paper handled TCGA data more appropriately, with a clear explanation of the positives and negatives of the dataset, and entirely separate models for FFPE and frozen tissue slides[15].

Sharing code can help to mitigate the effects of incomplete reporting and drastically improve reproducibility, but only 19 of the 45 papers did this, with some of these appearing to be incomplete or inaccessible. The better code repositories included detailed documentation to aid reproducibility, including environment set-up information[16,19], overviews of included functions[17,36,42], and code examples used to generate reported results[57].

Two papers were found to have major discrepancies between the reported data and the study design, indicating much greater risks of bias than those seen in any other research[46,54]. In one paper[46], it was reported that TCGA-OV data was used for subtyping with 5 classes, despite this dataset only including high-grade serous and low-grade serous carcinomas. In the other paper[54], it was reported that TCGA-OV data was used for slide-level classification into ovarian cancer and non-ovarian cancer classes using PAS-stained tissue, despite TCGA-OV only containing H&E-stained ovarian cancer slides.

Limitations of the review

While the review protocol was designed to reduce biases and maximise the quantity of relevant research included, there were some limitations. This review is restricted to published literature in the English language, however, AI research may be published in other languages or made available as pre-prints without publication in peer-reviewed journals, making this review incomplete. While most of the review process was completed by multiple independent researchers, the duplicate detection was performed by only a single researcher, raising the possibility of errors in this step of the review process, resulting in incorrect exclusions. Due to the significant time gap between the initial and final literature searches (approximately 12 months), there may have been inconsistencies in interpretations, both for data extraction and risk of bias assessments. Finally, this review focused only on light microscopy images of human histopathology samples relating to ovarian cancer, so may have overlooked useful literature outside of this domain.

Development of the field

The field of AI in ovarian cancer histopathology diagnosis is rapidly growing, with more research published since the start of 2020 than in all preceding years combined. The earliest research, published between 2010 and 2013, used hand-crafted features to train classical ML methods such as SVMs. These models were used for segmentation[49–51,53], malignancy classification[18,52], grading[21], and overall survival prediction[21]. Most of these early studies focused on IHC-stained tissue (5/7), which would be much less commonly used in subsequent research (6/38).

The field was relatively dormant in the following years, with only 6 papers published between 2014 and 2019, half of which had the same primary author[29,30,48]. These models still used traditional ML classifiers, though some used learned features rather than the traditional hand-crafted features. The models developed were used for histological subtyping[29,30,48] and cellular/tissue classification[40,57,66].

Since 2020, there has been a much greater volume of research published, most of which has involved the use of deep neural networks for automatic feature extraction and classification, with a minority using traditional machine learning model[17,24,31,61,67]. Recent research has investigated a broader array of diagnostic outcomes, including the classification of primary cancer type[15,54], mutation status[24,36,44], homologous recombination deficiency status[42], tumour–stroma reaction level[32], transcriptomic subtypes[19,24], microsatellite instability[24], and epithelial-mesenchymal transition status[41]. Three additional prognostic outcomes have also been predicted in more recent literature—progression-free survival[17,43,56], relapse[43,44], and treatment response[19,22,33–35].

Despite progress within a few specific outcomes, there was no obvious overall trend in the sizes of datasets used over time, either in terms of the number of slides or the number of participants. Similarly, there was no evidence that recent research included more rigorous internal validations, though external validations have been increasing in frequency—no research before 2021 included any external validation with ovarian cancer data, but seven studies published more recently did[16,24,27,28,34,38,39]. While these external validations were typically limited to small quantities of data, the inclusion of any external validation demonstrates progress from previous research. Such validations are essential to the clinical utility of these models as real-world implementation will require robustness to different sources of visual heterogeneity, with variation occurring across different data centres and within data centres over time. As this field continues to mature, we hope to see more studies conduct thorough validations with larger, high-quality independent datasets, including clearly reported protocols for patient recruitment and selection, pathology slide creation, and digitisation. This will help to reduce the biases, limited reproducibility, and limited generalisability identified in most of the existing research in this domain.

Current limitations and future recommendations

A large proportion of published work did not provide sufficient clinical and pathological information to assess the risk of bias. It is important that AI researchers thoroughly report data provenance to understand the extent of heterogeneity in the dataset, and to understand whether this has been appropriately accounted for in the study design. Modelling and analysis methods must also be thoroughly reported to improve reliability and reproducibility. Researchers may find it useful to refer to reporting checklists, such as *transparent reporting of a multivariable prediction model for individual prognosis or diagnosis* (TRIPOD)[68], to ensure that they have understood and reported all relevant details of their studies. In many studies, it is not clear how AI would fit in the clinical workflow, or whether there are limitations in how these methods could be applied. AI researchers should ensure they understand the clinical context of their data and potential models before undertaking research to reduce bias and increase utility. Ideally, this will involve regular interactions with expert clinicians, including histopathologists and oncologists.

To further improve reproducibility, we recommend that researchers should make code and data available where possible. It is relatively easy to publish code and generate documentation to enhance usability, and there are few drawbacks to doing so when publishing research. Making data available is more often difficult due to data security requirements and the potential storage costs, but it can provide benefits beyond the primary research of the original authors. Digital pathology research in ovarian cancer is currently limited by the lack of openly accessible data, leading to over-dependence on TCGA, and causing many researchers to painstakingly collate similar but distinct datasets. These datasets often contain little of the heterogeneity seen in multi-centre, multi-scanner data, making it difficult for researchers to train robust models or assess generalisability. Where heterogeneous data is included, it often includes small quantities of data which are different to the majority, introducing risks of bias and confounding rather than helping to overcome these issues. TCGA-based studies are prone to this, with significant differences between TCGA slides originating from different data centres[69], but with many of these centres only providing small quantities of data. Many researchers are reliant on open-access data, but there is a severe shortage of suitable open-access ovarian cancer histopathology data. Making such data available, with detailed protocols describing data creation, allows researchers to conduct more thorough analyses and significantly improve model generalisability and clinical implementability.

For AI to achieve clinical utility, it is essential that more robust validations are performed, especially considering the limitations of the available datasets. We recommend that researchers should always conduct thorough analyses, using cross-validation, bootstrapping, and/or external validations to ensure that results are robust and truly reflect the ability of their model(s) to generalise to unseen data, and are not simply caused by chance. This should include reporting the variability of results (typically in a 95%

confidence interval), especially when comparing multiple models to help to distinguish whether one model is genuinely better than another or whether the difference is due to chance. Statistical tests can also be beneficial for these evaluations. Another option for capturing variability is Bayesian uncertainty quantification, which can be used to separate aleatoric (inherent) and epistemic (modelling) uncertainty.

Current literature in this field can be largely characterised as model prototyping with homogeneous retrospective data. Researchers rarely consider the reality of human-machine interaction, perhaps believing that these models are a drop-in replacement for pathologists. However, these models perform narrow tasks within the pathology pipeline and do not take into consideration the clinical context beyond their limited training datasets and siloed tasks. We believe these models would be more beneficial (and more realistic to implement) as assistive tools for pathologists, providing secondary opinions or novel ancillary information. While current research is typically focused on assessing model accuracy without any pathologist input, different study designs could be employed to better assess the real-world utility of these models as assistive tools. For example, usability studies could investigate which models are most accessible and most informative to pathologists in practice, and prospective studies could quantify any benefits to diagnostic efficiency and patient outcomes, and investigate the robustness of models in practice. Understanding the effects of AI on the efficiency of diagnosis is particularly important given the limited supply of pathologists worldwide. As such, this type of research will significantly benefit clinical translation.

Summary of recommendations

To improve clinical utility, researchers should understand their data and ensure planned research is clinically relevant before any modelling, ideally involving clinicians throughout the project. They should also consider different study designs, including usability studies and/or prospective studies. When evaluating models, researchers should conduct thorough analyses using cross-validation, external validation, and/or bootstrapping. When reporting research, researchers should clearly report the context of any histopathology data, including how patients were recruited/selected, and how tissue specimens were processed to generate digital pathology images. Finally, researchers should make all code openly accessible, and make data available where possible.

METHODS

Literature search

Searches were conducted in three research databases, PubMed, Scopus and Web of Science, and two trial registries, Cochrane Central Register of Controlled Trials (CENTRAL) and the World Health Organisation International Clinical Trial Registry Platform (WHO-ICTRP). The research databases only include journals and conference proceedings which have undergone peer review, ensuring the integrity of included research. The initial searches were performed on 25/04/2022 and were most recently repeated on 19/05/2023. The search strategy was composed of three distinct aspects—artificial intelligence, ovarian cancer, and histopathology. For each aspect, multiple relevant terms were combined using the *OR* operator (e.g. "artificial intelligence" OR "machine learning"), and then these were combined using the *AND* operator to ensure that retrieved research met all three aspects. The widest possible set of search fields was used for each search engine except for Scopus, where restrictions were imposed to avoid searching within the citation list of each article, which is not an available field in the other search engines. The terms "ML" and "AI" were restricted to specific fields due to the diversity of their possible meanings. To ensure the most rigorous literature search possible, no restrictions were placed on the publication date or article type during searching.

Many AI approaches build on statistical models, such as logistic regression, which can blur the lines between disciplines. When conducting searches, a previously reported methodology was adopted[70] whereby typical AI approaches were searched by name (e.g. neural networks), and other methods were searched by whether the authors described their work as *artificial intelligence*. Full details of the search implementation for each database are provided in Supplementary Note 1. The review protocol was registered with PROSPERO before the search results were screened for inclusion (CRD42022334730).

Literature selection

One researcher (J.B.) manually removed duplicate papers with the assistance of the referencing software *EndNote X9*. Two researchers (J.B., K.A.) then independently screened articles for inclusion in two stages, the first based on title and abstract, the second based on full text. Disagreements were discussed and arbitrated by a third researcher (N.R. or N.M.O.). Trials in WHO-ICTRP do not have associated abstracts, so for these studies, only titles were available for initial screening.

The inclusion criteria required that research evaluated the use of at least one AI approach to make diagnostic or prognostic inferences on human histopathology images from suspected or confirmed cases of ovarian cancer. Studies were only included where AI methods were applied directly to the digital pathology images, or to features which were automatically extracted from the images. Fundamental tasks, such as segmentation and cell counting, were included as these could be used by pathologists for computer-aided diagnosis. Only conventional light microscopy images were considered, with other imaging modalities, such as fluorescence and hyperspectral imaging, excluded. Publications which did not include primary research were excluded (such as review papers). Non-English language articles and research where a full version of the manuscript was not accessible were excluded.

A model in an included study was considered to be a *model of interest* if it met the same inclusion criteria. Where multiple models were compared against the same outcome, the model of interest was taken to be the newly proposed model, with the best performing model during validation taken if this was unclear. If multiple model outcomes were assessed in the same study, a model of interest was taken for each model outcome, regardless of any similarity in modelling approaches. The same model outcome at different levels of precision (e.g. patch-level, slide-level, patient-level) were not considered to be different model outcomes. Models did not need to be entirely independent, for example, the output of one model of interest could have been used as the input of another model of interest on the condition that model performance was separately evaluated for each model.

Risk of bias assessment

The risk of bias was assessed for models of interest using the Prediction model Risk Of Bias ASsessment Tool (PROBAST)[71], where *risk of bias* is the chance of reported results being distorted by limitations within the study design, conduct, and analysis. It includes 20 guiding questions which are categorised into four domains (participants, predictors, outcome, and analysis), which are summarised as either high-risk or low-risk, or unclear in the case that there is insufficient information to make a comprehensive assessment and none of the available information indicates a high risk of bias. As such, an unclear risk of bias does not indicate methodological flaws, but incomplete reporting.

The **participants** domain covers the recruitment and selection of participants to ensure the study population is consistent and representative of the target population. Relevant details include the participant recruitment strategy (when and where participants

were recruited), the inclusion criteria, and how many participants were recruited.

The **predictors** domain covers the consistent definition and measurement of predictors, which in this field typically refers to the generation of digital pathology images. This includes methods for fixing, staining, scanning, and digitally processing tissue before modelling.

The **outcome** domain covers the appropriate definition and consistent determination of ground-truth labels. This includes the criteria used to determine diagnosis/prognosis, the expertise of any persons determining these labels, and whether labels are determined independently of any model outputs.

The **analysis** domain covers statistical considerations in the evaluation of model performance to ensure valid and not unduly optimistic results. This includes many factors, such as the number of participants in the test set with each outcome, the validation approaches used (cross-validation, external validation, bootstrapping, etc.), the metrics used to assess performance, and methods used to overcome the effects of censoring, competing risks/confounders, and missing data. The risks caused by some of these factors are interrelated, for example, the risk of bias from using a small dataset is somewhat mitigated by cross-validation, which increases the effective size of the test set and can be used to assess variability, reducing optimism in the results. Further, the risk caused by using a small dataset depends on the type of outcome being predicted, for example, more data is required for a robust analysis of 5-class classification than binary classification. There must also be sufficient data within all relevant patient subgroups, for example, if multiple subtypes of ovarian cancer are included, there must not be a subtype that is only represented by a few patients. Due to these interrelated factors, there are no strict criteria to determine the appropriate size of a dataset, though fewer than 50 samples per class or fewer than 100 samples overall is likely to be considered high-risk, and more than 1000 samples overall is likely to be considered low-risk.

Risks of bias often arise due to inconsistent methodologies. Inconsistency in the participants and predictors domains may cause heterogeneity in the visual properties of digital pathology slides which may lead to spurious correlations, either through random chance or systematic differences between subgroups in the dataset. Varied data may be beneficial during training to improve model generalisability when using large datasets, though this must be closely controlled to avoid introducing systematic confounding. Inconsistent determination of the outcome can mean that the results of a study are unreliable due to spurious correlations in the ground truth labels, or invalid due to incorrect determination of labels.

While PROBAST provides a framework to assess risks of bias, there is some level of subjectivity in the interpretation of signalling questions. As such, each model was analysed by three independent researchers (any of J.B., K.A., N.R., K.Z., N.M.O.), with at least one computer scientist and one clinician involved in the risk of bias assessment for each model. The PROBAST applicability of research analysis was not implemented as it is unsuitable for such a diverse array of possible research questions.

Data synthesis

Data extraction was performed independently by two researchers (J.B., K.A.) using a form containing 81 fields within the categories *Overview*, *Data*, *Methods*, *Results*, and *Miscellaneous*. Several of these fields were added or clarified during data extraction with the agreement of both researchers and retroactively applied to all accepted literature. The final data extraction form is available at www.github.com/scjjb/OvCaReview, and is summarised in Supplementary Table 1.

Information was sought from full-text articles, as well as references and supplementary materials where appropriate. Inferences were made only when both researchers were confident that this gave the correct information, with disagreements resolved through discussion. Fields which could not be confidently completed were labelled as being *unclear*.

All extracted data were summarised in two tables, one each for study-level and model-level characteristics. Only models of interest were included in these tables. The term *model outcome* refers to the model output, whether this was a clinical outcome (diagnosis/prognosis), or a diagnostically relevant outcome that could be used for computer-aided diagnosis, such as tumour segmentation. The data synthesis did not include any meta-analysis due to the diversity of included methods and model outcomes. The PRISMA 2020 guidelines for reporting systematic reviews were followed, with checklists provided in Supplementary Tables 2 and 3.

DATA AVAILABILITY

The authors declare that the main data supporting the findings of this study are available within the article and its Supplementary Information files. Extra data are available from the corresponding author upon request.

Received: 31 March 2023; Accepted: 1 August 2023;
Published online: 31 August 2023

REFERENCES

1. Sung, H. et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Menon, U. et al. Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* **397**, 2182–2193 (2021).
3. Ebell, M. H., Culp, M. B. & Radke, T. J. A systematic review of symptoms for the diagnosis of ovarian cancer. *Am. J. Prev. Med.* **50**, 384–394 (2016).
4. Berek, J. S., Renz, M., Kehoe, S., Kumar, L. & Friedlander, M. Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. *Int. J. Gynecol. Obstet.* **155**, 61–85 (2021).
5. Köbel, M. et al. Ovarian carcinoma subtypes are different diseases: implications for biomarker studies. *PLoS Med.* **5**, e232 (2008).
6. Prat, J. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. *Int. J. Gynecol. Obstet.* **124**, 1–5 (2014).
7. Matsuno, R. K. et al. Agreement for tumor grade of ovarian carcinoma: analysis of archival tissues from the surveillance, epidemiology, and end results residual tissue repository. *Cancer Causes Control* **24**, 749–757 (2013).
8. Köbel, M. et al. Ovarian carcinoma histotype determination is highly reproducible, and is improved through the use of immunohistochemistry. *Histopathology* **64**, 1004–1013 (2014).
9. Barnard, M. E. et al. Inter-pathologist and pathology report agreement for ovarian tumor characteristics in the nurses' health studies. *Gynecol. Oncol.* **150**, 521–526 (2018).
10. Wilson, M. L. et al. Access to pathology and laboratory medicine services: a crucial gap. *Lancet* **391**, 1927–1938 (2018).
11. Royal College of Pathologists. Meeting pathology demand: histopathology workforce census. https://www.rcpath.org/static/952a934d-2ec3-48c9-a8e6e00fcdca700f/Meeting-Pathology-Demand-Histopathology-Workforce-Census-2018.pdf (2018).
12. Baidoshvili, A. et al. Evaluating the benefits of digital pathology implementation: time savings in laboratory logistics. *Histopathology* **73**, 784–794 (2018).
13. Stenzinger, A. et al. Artificial intelligence and pathology: from principles to practice and future applications in histomorphology and molecular profiling. *Semin. Cancer Biol.* **84**, 129–143 (2022).
14. Raciti, P. et al. Clinical validation of artificial intelligence–augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. *Arch. Pathol. Lab. Med.* https://doi.org/10.5858/arpa.2022-0066-OA (2022).
15. Kalra, S. et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *npj Digit. Med.* **3**, 31 (2020).
16. Meng, Z. et al. A deep learning-based system trained for gastrointestinal stromal tumor screening can identify multiple types of soft tissue tumors. *Am. J. Pathol.* **193**, 899–912 (2023).
17. Boehm, K. M. et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat. Cancer* **3**, 723–733 (2022).
18. Kothari, S., Phan, J. H., Osunkoya, A. O. & Wang, M. D. Biological interpretation of morphological patterns in histopathological whole-slide images. In *Proceedings of*

the ACM Conference on Bioinformatics, Computational Biology and Biomedicine 218–225 (ACM, 2012).
19. Yu, K. H. et al. Deciphering serous ovarian carcinoma histopathology and platinum response by convolutional neural networks. *BMC Med.* **18**, 1–14 (2020).
20. Liu, T., Su, R., Sun, C., Li, X. & Wei, L. EOCSA: Predicting prognosis of epithelial ovarian cancer with whole slide histopathological images. *Expert Syst. Appl.* **206**, 117643 (2022).
21. Poruthoor, A., Phan, J. H., Kothari, S. & Wang, M. D. Exploration of genomic, proteomic, and histopathological image data integration methods for clinical prediction. In *2013 IEEE China Summit and International Conference on Signal and Information Processing* 259–263 (IEEE, 2013).
22. Yaar, A., Asif, A., Raza, S. E. A., Rajpoot, N. & Minhas, F. Cross-domain knowledge transfer for prediction of chemosensitivity in ovarian cancer patients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 928–929 (IEEE, 2020).
23. Ghoniem, R. M., Algarni, A. D., Refky, B. & Ewees, A. A. Multi-modal evolutionary deep learning model for ovarian cancer diagnosis. *Symmetry* **13**, 643 (2021).
24. Zeng, H., Chen, L., Zhang, M., Luo, Y. & Ma, X. Integration of histopathological images and multi-dimensional omics analyses predicts molecular features and prognosis in high-grade serous ovarian cancer. *Gynecol. Oncol.* **163**, 171–180 (2021).
25. Holback, C. et al. The cancer genome atlas ovarian cancer collection (TCGA-OV) (version 4) [data set]. The Cancer Imaging Archive. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=7569497 (2016).
26. Levine, A. B. et al. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *J. Pathol.* **252**, 178–188 (2020).
27. Boschman, J. et al. The utility of color normalization for ai-based diagnosis of hematoxylin and eosin-stained pathology images. *J. Pathol.* **256**, 15–24 (2022).
28. Farahani, H. et al. Deep learning-based histotype diagnosis of ovarian carcinoma whole-slide pathology images. *Mod. Pathol.* **35**, 1983–1990 (2022).
29. BenTaieb, A., Li-Chang, H., Huntsman, D. & Hamarneh, G. Automatic diagnosis of ovarian carcinomas via sparse multiresolution tissue representation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18* 629–636 (Springer, 2015).
30. BenTaieb, A., Nosrati, M. S., Li-Chang, H., Huntsman, D. & Hamarneh, G. Clinically-inspired automatic classification of ovarian carcinoma subtypes. *J. Pathol. Informatics* **7**, 28 (2016).
31. Jiang, J. et al. Digital pathology-based study of cell- and tissue-level morphologic features in serous borderline ovarian tumor and high-grade serous ovarian cancer. *J. Pathol. Informatics* **12**, 24 (2021).
32. Jiang, J. et al. Computational tumor stroma reaction evaluation led to novel prognosis-associated fibrosis and molecular signature discoveries in high-grade serous ovarian carcinoma. *Front. Med.* **9**, 994467 (2022).
33. Wang, C.-W. et al. A weakly supervised deep learning method for guiding ovarian cancer treatment and identifying an effective biomarker. *Cancers* **14**, 1651 (2022).
34. Wang, C.-W. et al. Weakly supervised deep learning for prediction of treatment effectiveness on ovarian cancer from histopathology images. *Comput. Med. Imaging Graphics* **99**, 102093 (2022).
35. Wang, C.-W. et al. Interpretable attention-based deep learning ensemble for personalized ovarian cancer treatment without manual annotations. *Comput. Med. Imaging Graphics* **107**, 102233 (2023).
36. Ho, D. J. et al. Deep interactive learning-based ovarian cancer segmentation of h&e-stained whole slide images to study morphological patterns of brca mutation. *J. Pathol. Informatics* **14**, 100160 (2023).
37. Paijens, S. T. et al. Prognostic image-based quantification of cd8cd103 t cell subsets in high-grade serous ovarian cancer patients. *Oncoimmunology* **10**, 1935104 (2021).
38. Shin, S. J. et al. Style transfer strategy for developing a generalizable deep learning application in digital pathology. *Comput. Methods Programs Biomed.* **198**, 105815 (2021).
39. Mayer, R. S. et al. How to learn with intentional mistakes: Noisyensembles to overcome poor tissue quality for deep learning in computational pathology. *Front. Med.* **9**, 959068 (2022).
40. Du, Y. et al. Classification of tumor epithelium and stroma by exploiting image features learned by deep convolutional neural networks. *Ann. Biomed. Eng.* **46**, 1988–1999 (2018).
41. Hu, Y. et al. Predicting molecular traits from tissue morphology through self-interactive multi-instance learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II* 130–139 (Springer, 2022).
42. Lazard, T. et al. Deep learning identifies morphological patterns of homologous recombination deficiency in luminal breast cancers from whole slide images. *Cell Rep. Med.* **3**, 100872 (2022).
43. Yokomizo, R. et al. O3c glass-class: a machine-learning framework for prognostic prediction of ovarian clear-cell carcinoma. *Bioinformatics Biol. Insights* **16**, 11779322221134312 (2022).
44. Nero, C. et al. Deep-learning to predict brca mutation and survival from digital H&E slides of epithelial ovarian cancer. *Int. J. Mol. Sci.* **23**, 11326 (2022).
45. Wu, M. et al. Exploring prognostic indicators in the pathological images of ovarian cancer based on a deep survival network. *Front. Genet.* **13**, 1069673 (2023).
46. Kasture, K. R., Choudhari, D. & Matte, P. N. Prediction and classification of ovarian cancer using enhanced deep convolutional neural network. *Int. J. Eng. Trends Technol.* **70**, 310–318 (2022).
47. Kowalski, P. A., Błoniarz, J. & Chmura, Ł. Convolutional neural networks in the ovarian cancer detection. In *Computational Intelligence and Mathematics for Tackling Complex Problems 2* 55–64 (Springer, 2022).
48. BenTaieb, A., Li-Chang, H., Huntsman, D. & Hamarneh, G. A structured latent model for ovarian carcinoma subtyping from histopathology slides. *Med. Image Anal.* **39**, 194–205 (2017).
49. Dong, J., Li, J., Lu, J. & Fu, A. Automatic segmentation for ovarian cancer immunohistochemical image based on chroma criterion. In *2010 2nd International Conference on Advanced Computer Control, vol. 2* 147–150 (IEEE, 2010).
50. Dong, J., Li, J., Fu, A. & Lv, H. Automatic segmentation for ovarian cancer immunohistochemical image based on YUV color space. In *2010 International Conference on Biomedical Engineering and Computer Science* 1–4 (IEEE, 2010).
51. Signolle, N., Revenu, M., Plancoulaine, B. & Herlin, P. Wavelet-based multiscale texture segmentation: application to stromal compartment characterization on virtual slides. *Signal Process.* **90**, 2412–2422 (2010).
52. Janowczyk, A., Chandran, S., Feldman, M. & Madabhushi, A. Local morphologic scale: application to segmenting tumor infiltrating lymphocytes in ovarian cancer tmas. In *Medical Imaging 2011: Image Processing, vol. 7962* 827–840 (SPIE, 2011).
53. Janowczyk, A. et al. High-throughput biomarker segmentation on ovarian cancer tissue microarrays via hierarchical normalized cuts. *IEEE Trans. Biomed. Eng.* **59**, 1250–1252 (2012).
54. Ramasamy, S. & Kaliyaperumal, V. A hybridized channel selection approach with deep convolutional neural network for effective ovarian cancer prediction in periodic acid-Schiff-stained images. *Concurrency Comput. Pract. Exp.* **35**, e7568 (2023).
55. Gentles, L. et al. Integration of computer-aided automated analysis algorithms in the development and validation of immunohistochemistry biomarkers in ovarian cancer. *J. Clin. Pathol.* **74**, 469–474 (2021).
56. Laury, A. R., Blom, S., Ropponen, T., Virtanen, A. & Carpén, O. M. Artificial intelligence-based image analysis can predict outcome in high-grade serous carcinoma via histology alone. *Sci. Rep.* **11**, 19165 (2021).
57. Heindl, A. et al. Microenvironmental niche divergence shapes BRCA1-dysregulated ovarian cancer morphological plasticity. *Nat. Commun.* **9**, 3917 (2018).
58. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning* 2127–2136 (PMLR, 2018).
59. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
60. He, K. et al. Transformers in medical image analysis: a review. *Intell. Med.* **3**, 59–78 (2022).
61. Elie, N. et al. Impact of automated methods for quantitative evaluation of immunostaining: towards digital pathology. *Front. Oncol.* **12**, 931035 (2022).
62. Shrestha, P. et al. A systematic review on the use of artificial intelligence in gynecologic imaging–background, state of the art, and future directions. *Gynecol. Oncol.* **166**, 596–605 (2022).
63. Zhou, J., Cao, W., Wang, L., Pan, Z. & Fu, Y. Application of artificial intelligence in the diagnosis and prognostic prediction of ovarian cancer. *Comput. Biol. Med.* **146**, 105608 (2022).
64. Fiste, O., Liontos, M., Zagouri, F., Stamatakos, G. & Dimopoulos, M. A. Machine learning applications in gynecological cancer: a critical review. *Crit. Rev. Oncol. Hematol.* **179**, 103808 (2022).
65. Xu, H.-L. et al. Artificial intelligence performance in image-based ovarian cancer identification: a systematic review and meta-analysis. *EClinicalMedicine* **53**, 101662 (2022).
66. Lorsakul, A. et al. Automated wholeslide analysis of multiplex-brightfield ihc images for cancer cells and carcinoma-associated fibroblasts. In *Medical Imaging 2017: Digital Pathology, vol. 10140* 41–46 (SPIE, 2017).
67. Salguero, J. et al. Selecting training samples for ovarian cancer classification via a semi-supervised clustering approach. In *Medical Imaging 2022: Digital and Computational Pathology, vol. 12039* 20–24 (SPIE, 2022).
68. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Ann. Intern. Med.* **162**, 55–63 (2015).
69. Dehkharghanian, T. et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn. Pathol.* **18**, 1–12 (2023).
70. Dhiman, P. et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med. Res. Methodol.* **22**, 101 (2022).
71. Wolff, R. F. et al. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **170**, 51–58 (2019).

72. Köbel, M. et al. Diagnosis of ovarian carcinoma cell type is highly reproducible: a transcanadian study. *Am. J. Surg. Pathol.* **34**, 984–993 (2010).

ACKNOWLEDGEMENTS

There was no direct funding for this research. J.B. is supported by the UKRI Engineering and Physical Sciences Research Council (EPSRC) [EP/S024336/1]. K.A. and P.A. are supported by the Tony Bramall Charitable Trust. A.S. is supported by Innovate UK via the National Consortium of Intelligent Medical Imaging (NCIMI) [104688], Cancer Research UK [C19942/A28832] and Leeds Hospitals Charity [9R01/1403]. The funders had no role in influencing the content of this research. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

J.B. created the study protocol with feedback and contributions from all other authors. J.B., K.A., K.Z., N.M.O., and N.R. performed the risk of bias assessments. J.B. and K.A. performed data extraction. J.B. analysed extracted data and wrote the manuscript, with feedback and contributions from all other authors.

COMPETING INTERESTS

G.H. receives research funding from IQVIA. N.M.O. receives research funding from 4D Path. The other authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-023-00432-6.

Correspondence and requests for materials should be addressed to Jack Breen.

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.