Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Artificial Intelligence for Borderline Breast Lesions on Whole-Slide Images: A Systematic Review of Accuracy, ncertainty and Generalisation

Moulvi Gufran

moulvigufran@gmail.com

Government Medical College, Vizianagaram https://orcid.org/0009-0004-9450-5487

Sirisha Gunta

joysirisha03@gmail.com

Department of Pathology, Government Medical College, Vizianagaram

Systematic Review

Keywords: atypical ductal hyperplasia, flat epithelial atypia, computational pathology,whole-slide imaging, uncertainty, systematic review

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11573095/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Artificial Intelligence for Borderline Breast Lesions on Whole-Slide Images: A Systematic Review of Accuracy, Uncertainty and Generalisation

Title page

Authors: Gufran Moulvi (1), Sirisha Gunta (2)

(1) MBBS student, Government Medical College, Vizianagaram, Andhra Pradesh, India
(2) Department of Pathology, Government Medical College, Vizianagaram, Andhra Pradesh, India

Corresponding author: Gufran Moulvi, moulvigufran@gmail.com

ORCID: Gufran Moulvi, https://orcid.org/0009-0004-9450-5487

Protocol: registered on Zenodo, https://zenodo.org/records/23422795 (version 1.1)

Word count: about 2,450 (main text, excluding abstract, tables, legends and references) · **Tables:** 3 · **Figures:** 1

Abstract

Background: Borderline breast lesions, such as atypical ductal hyperplasia (ADH), flat epithelial atypia (FEA) and lobular neoplasia, are among the least reproducible diagnoses in breast pathology. We reviewed how well artificial intelligence (AI) on digitised haematoxylin and eosin slides classifies these lesions, whether it was tested at other hospitals, and whether it reports uncertainty.

Methods: Following PRISMA 2020 and a registered protocol, we searched PubMed, OpenAlex and Europe PMC to 22 October 2026, with citation and preprint searching. Screening was assisted by a large language model and checked by both authors. Risk of bias was assessed with QUADAS-2, and results were synthesised narratively (SWiM).

Results: Eighteen reports (2016–2026) from only seven patient cohorts were included. In multi-class settings, AI performed poorly on ADH (recall 0.19–0.69; F1 0.24–0.47 on the public BRACS dataset, where pathologists reached 0.38). In binary tasks, AI separated ductal carcinoma in situ (DCIS) from atypia with accuracy 0.85–0.90, versus 0.80 for pathologists on the same cases. Only 2 of 18 reports tested the model at another hospital

(external area under the curve 0.92 and 0.61–0.70). No report offered calibrated uncertainty or referral of uncertain cases to a pathologist. Risk of bias from patient selection was high in all reports.

Conclusions: AI for borderline breast lesions is promising in narrow tasks but weak for ADH in realistic multi-class settings, rarely validated externally, and never designed to say when it is unsure. Future studies need multi-centre test sets and calibrated uncertainty with a refer-to-pathologist option.

Registration: Zenodo, https://zenodo.org/records/23422795. **Funding:** None.

Keywords: atypical ductal hyperplasia; flat epithelial atypia; computational pathology; whole-slide imaging; uncertainty; systematic review

Introduction

Borderline (atypical) breast lesions sit between clearly benign proliferations and carcinoma. They include atypical ductal hyperplasia (ADH), flat epithelial atypia (FEA), lobular neoplasia (atypical lobular hyperplasia and lobular carcinoma in situ) and atypical papillary lesions. The line between ADH and low-grade ductal carcinoma in situ (DCIS) depends partly on lesion size and extent, which makes it hard to draw on a core biopsy [3,4].

These diagnoses are among the least reproducible in surgical pathology. In a large US study, 115 pathologists agreed with an expert consensus in 48% of atypia cases, compared with 96% of invasive carcinomas [1]. Agreement on ADH was also poor in earlier inter- and intra-observer studies [2,5]. The diagnosis matters clinically: ADH on core biopsy usually leads to surgical excision, yet most excisions show no cancer [6].

Artificial intelligence (AI) applied to digitised haematoxylin and eosin (H&E) slides could make these diagnoses more consistent. Many models now detect invasive cancer well, but borderline lesions are a harder test. Classes are small and imbalanced, labels are noisy because experts disagree, and models trained in one laboratory may fail at another because of differences in staining and scanners. For such a difficult task, a safe model should also be able to say when it is unsure and pass the case to a pathologist [7].

A previous systematic review covered AI for DCIS but did not focus on borderline lesions, external validation or uncertainty [8]. We therefore systematically reviewed AI for borderline breast lesions on H&E slides. We asked three questions: how accurate the models are for the borderline classes, whether they were tested on data from another hospital, and whether they report uncertainty or abstain on difficult cases.

Methods

Protocol and reporting

This review follows the PRISMA 2020 statement [9]. The protocol was registered on Zenodo before formal screening (https://zenodo.org/records/23422795). Version 1.1 (16 October 2026) replaced Embase, Scopus and Web of Science with OpenAlex because our institution has no access to those databases; eligibility criteria and outcomes were not changed.

Review question (PICO)

Element	Definition
Population	Breast tissue (core biopsies or excisions) containing borderline lesions: ADH, FEA, lobular neoplasia, atypical papilloma, or atypia versus low-grade DCIS
Index test (Intervention)	Any AI or machine-learning model applied to digitised H&E slides or regions of interest
Comparator / reference standard	Diagnosis by a pathologist or a consensus of pathologists; pathologists reading the same cases where reported
Outcomes	Accuracy for the borderline class (sensitivity, specificity, recall, precision, F1, AUC); testing at another hospital (external validation); reporting of uncertainty, calibration or abstention
Study design	Diagnostic accuracy, model development or validation studies

Eligibility criteria

We included studies that met all of the following:

1. Included at least one borderline lesion: ADH, FEA, lobular neoplasia (ALH or LCIS), atypical papilloma, an “atypia” class, or a task separating atypia from low-grade DCIS.
2. Applied AI or machine learning to digitised H&E histology images (whole-slide images or regions of interest).
3. Used a pathologist diagnosis or consensus as the reference standard.

4. Reported at least one result specific to the borderline group (for example per-class recall, F1 or area under the curve, or a binary task in which the borderline lesion was one of the two classes).
5. Was a peer-reviewed original study, either a journal article or a full paper in conference proceedings.

We excluded studies of radiology images, cytology, immunohistochemistry alone or other organs; reviews and editorials; conference abstracts; and preprints. Studies using only the public BRACS dataset were included and analysed as a subgroup.

Information sources and search

PubMed was searched on 14 October 2026, OpenAlex on 16 October 2026 and Europe PMC on 22 October 2026, from inception to the search date. Europe PMC was added after protocol registration to widen coverage, and this is reported as a deviation from the protocol. Each search combined four concept blocks: AI, breast, borderline lesions (including DCIS terms, so that atypia-versus-DCIS studies were captured) and histology. The full search strings are given in the protocol and in Supplementary File 1. On 22 October 2026 we searched the reference lists of all studies included from the databases and of the earlier DCIS review [8], and the articles citing them (forward citation searching), using OpenAlex.

As planned in the protocol, preprints were also searched on 22 October 2026: arXiv through its API (breast AND atypical ductal hyperplasia, flat epithelial atypia, atypia, lobular neoplasia or BRACS) and preprints indexed in OpenAlex, which covers medRxiv and bioRxiv, using the same concept blocks as the main search. Preprints were not eligible for inclusion and are listed separately.

A supplementary search of Semantic Scholar, which indexes computer-science conference papers, was run on the same date. A search of dblp was attempted but failed because the site blocks automated access. All search strings are given in Supplementary File 1.

Study selection

Records were de-duplicated by DOI and then by title. The first-pass screen of all 249 database records was carried out with the assistance of a large language model tool (reviewer 1), and every decision was recorded with a reason. The first author (GM, reviewer 2) independently screened 45 records: all 14 that reviewer 1 marked as include or uncertain and 31 that reviewer 1 excluded. Agreement was 39 of 45 records (Cohen's kappa 0.71). In addition, the supervising author (SG) checked the screening decisions for all 249 records. Disagreements were resolved by reading the full text and by discussion with SG, who accepted the final decisions. Europe PMC records not already screened,

citation-search records and supplementary-search records were screened on title and abstract against the same criteria, and possible inclusions were assessed in full text. One citation-search record (Pati 2022) had been excluded on title and abstract because its abstract does not mention atypia; it was re-assessed in full text after its preprint version was found in the preprint search, and it was included.

Data extraction

From each study we extracted the first author, year, country, venue, data source and number of centres, numbers of slides, patients and regions of interest, the borderline lesions and classes studied, the AI method, the unit of analysis, the validation design, the reference standard, the results for each borderline class, any comparison with pathologists, and any reporting of uncertainty, calibration or abstention. Numbers were copied from the text, tables or figures of the full paper, and the source table was recorded. Extraction was drafted with the help of a large language model tool and checked against the papers by the authors.

Risk of bias

Risk of bias and applicability were judged with QUADAS-2 [10] in four domains: patient selection, index test, reference standard, and flow and timing. Judgements were drafted with the assistance of large language model tools and then reviewed by the supervising author (SG), who confirmed or changed each judgement and approved the final ratings.

Synthesis

Tasks, class definitions and metrics differed widely, and several reports used the same patients. A meta-analysis was therefore not appropriate. We grouped studies by task (multi-class versus binary) and by validation design (internal versus external), and summarised them narratively following the Synthesis Without Meta-analysis (SWiM) guideline [11]. Reports from the same patient cohort were counted once when describing the evidence base. Reporting bias could not be assessed formally because there were too few comparable studies, and certainty of the evidence was not graded with GRADE because no pooled estimate was produced.

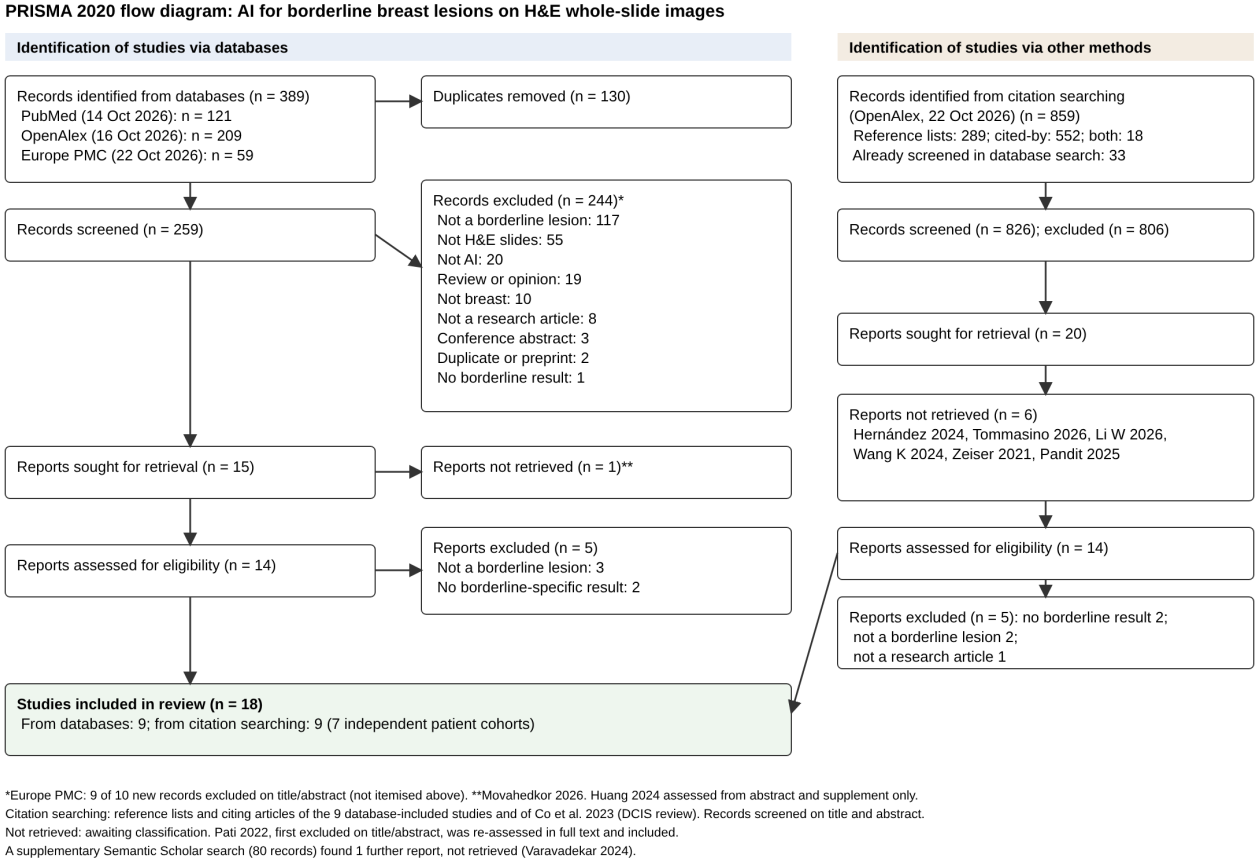
Results

Study selection

The database searches found 389 records (PubMed 121, OpenAlex 209, Europe PMC 59). After removing 130 duplicates, 259 were screened and 15 were sought in full text. One could not be retrieved, and of the 14 assessed, 9 were included. Citation searching found

859 further records, 33 of which had already been screened. Of the remaining 826, 20 were sought in full text. Six could not be retrieved, 5 were excluded, and 9 were included. In total, 18 reports were included (Figure 1). Excluded full-text reports and their reasons are listed in Supplementary File 2.

Figure 1. PRISMA 2020 flow diagram. Records identified, screened, assessed and included, from databases (left) and citation searching (right).



*Europe PMC: 9 of 10 new records excluded on title/abstract (not itemised above). **Movahedkor 2026. Huang 2024 assessed from abstract and supplement only.
Citation searching: reference lists and citing articles of the 9 database-included studies and of Co et al. 2023 (DCIS review). Records screened on title and abstract.
Not retrieved: awaiting classification. Pati 2022, first excluded on title/abstract, was re-assessed in full text and included.
A supplementary Semantic Scholar search (80 records) found 1 further report, not retrieved (Varavadekar 2024).

Study characteristics

The 18 reports were published between 2016 and 2026 (Table 1). They came from only seven independent patient cohorts. Seven reports used the same 240-slide cohort from the US Breast Cancer Surveillance Consortium (the B-Path study) [12–18]. Three used a single-centre University of Pittsburgh cohort [19–21], and four used the public BRACS dataset from one hospital in Naples [22–24,29]. The other four cohorts came from Japan, Israel and France, and the USA [25–28].

ADH or an atypia class was studied in 17 reports, FEA in 8 and lobular neoplasia as a separate class in 1 [27]. In two reports ADH was pooled with low-grade DCIS [26,28]. Most models worked on regions of interest selected or annotated by pathologists rather than on whole slides. Methods ranged from hand-crafted nuclear and architectural features with classical classifiers to convolutional neural networks, graph neural networks and transformers.

Table 1. Characteristics of included studies

Study	Country	Data source (centres)	Size	Borderline lesions	Task	Validation
Gecer 2018 [12]	Turkey/USA	B-Path (BCSC, multi-state US)	240 WSIs	ADH	5-class slide	Held-out 60 WSIs
Mercan 2018 [13]	Turkey/USA	B-Path	240 WSIs	ADH (also FEA, ALH, papilloma in 14-class)	5- and 14-class	4-fold CV by slide
Mercan 2019 SPIE [14]	Turkey/USA	B-Path	240 WSIs; 437 ROIs	ADH	4-class ROI	Held-out 120 slides
Mercan 2019 JAMA [15]	USA	B-Path	240 WSIs	Atypia	DCIS vs atypia (stepwise)	Internal (same cohort)
Li 2021 [16]	USA	B-Path	240 biopsies; 428 ROIs	Atypia	4-class; DCIS vs atypia	Leave-one-out CV
Mercan 2021 [17]	Turkey/USA	B-Path	240 WSIs; 437 ROIs	ADH	4-class ROI	Held-out 120 slides
Mehta 2022 [18]	USA	B-Path	240 WSIs; 422 ROIs	Atypia	4-class ROI	Held-out 216 ROIs
Tosun 2017 [19]	USA	Univ. Pittsburgh (1)	93 WSIs; 1,759 ROIs	ADH, FEA	High-risk vs low-risk	Held-out ROIs
Parvatikar 2020 [20]	USA	Univ. Pittsburgh (1)	93 WSIs; 1,441 ROIs	ADH, FEA	High-risk vs low-risk	Held-out 607 ROIs
Parvatikar 2021 [21]	USA	Univ. Pittsburgh (1)	93 WSIs; 1,295 ROIs	ADH, FEA	ADH vs rest; FEA vs rest	10 random splits
Brancati 2022 [22]	Italy	BRACS, Naples (1)	547 WSIs; 4,539 ROIs	ADH, FEA	7-class ROI	Official split
Li 2025 [23]	China	BRACS	387 WSIs; 4,539 ROIs	ADH, FEA	7-class ROI; pairwise	Held-out 575 ROIs
Rinaldi 2026 [24]	Italy	BRACS	BRACS ROIs	ADH, FEA	7-class ROI	Official split
Pati 2022 [29]	Switzerland/Italy	BRACS (earlier release)	325 WSIs; 4,391 ROIs	ADH, FEA	7- and 4-class ROI; binary tasks	Held-out 626 ROIs (3 runs)

Study	Country	Data source (centres)	Size	Borderline lesions	Task	Validation
Yamada 2016 [25]	Japan	Shinshu Univ. (1)	175 specimens; 1,419 ROIs	ADH	ADH vs LG-DCIS (pairwise)	Random ROI splits; 4 case splits
Sandbank 2022 [26]	Israel/France	Maccabi + Institut Curie (2)	436 cases external	ADH pooled with LG-DCIS	LG-DCIS/ADH vs HG/IG-DCIS	**External, 2 sites**
Tahir 2025 [27]	USA	Ohio State Univ. (1)	104 cases	Lobular neoplasia	LN vs other classes; reader study	Single test set
Yoo 2026 [28]	USA/S. Korea	UVA + GWU (2)	108 patients	ADH pooled with LG-DCIS	4-class tile and patient	**External, 39 patients**

B-Path, Breast Pathology Study cohort; BCSC, Breast Cancer Surveillance Consortium; CV, cross-validation; HG/IG, high/intermediate grade; LG, low grade; LN, lobular neoplasia; ROI, region of interest; WSI, whole-slide image.

Performance for borderline lesions

Results are shown in Table 2. Two patterns stood out.

Multi-class tasks. When ADH had to be picked out from several classes, performance was low. ADH recall was 0.25 on whole slides in Gecer 2018 (4 of 16 ADH slides) [12] and 0.48 at region level in Mercan 2018 [13]. Later B-Path models reached ADH recall of 0.62–0.69 and F1 of 0.70–0.72 [14,17]. On BRACS, ADH was the hardest class, with F1 of 0.24 in the baseline model [22], 0.27–0.46 with graph networks [24], 0.40 with a hierarchical cell-to-tissue graph network [29] and 0.47 with knowledge distillation [23]. In the one BRACS study that also tested three pathologists on the same images, their mean ADH F1 was 0.38 [29]. FEA was easier (F1 0.59–0.76) [22–24,29].

Binary tasks. Narrower tasks gave better results. Separating DCIS from atypia reached accuracy of 0.85 [15] and 0.90 [16], higher than the 0.80 achieved by pathologists on the same cases. Separating high-risk lesions (ADH and FEA) from low-risk lesions gave recall of 0.65–0.76, close to the 0.77 of a single expert [19,20]. Yamada and colleagues separated ADH from low-grade DCIS with 88.1% accuracy on random region splits, but accuracy varied from 71.8% to 95.9% when whole cases were held out [25]. For lobular neoplasia, Tahir and colleagues reported an AUC of 0.953 [27].

Table 2. Results for borderline lesions

Study	Borderline task	AI result	Pathologists (same cases)	Uncertainty reported
Gecer 2018	ADH, 5-class slide	Recall 0.25, precision 0.67	Overall accuracy 65.4% vs AI 55%	No
Mercan 2018	ADH, 5-class ROI	Precision 0.36, recall 0.48	Not head-to-head	Partial (low-score windows dropped)
Mercan 2019 SPIE	ADH, 4-class ROI	Precision 0.75, recall 0.69, F 0.72	Overall 0.70 vs AI 0.67	No
Mercan 2019 JAMA	DCIS vs atypia	Sens 0.89, spec 0.80, acc 0.85	Sens 0.70, spec 0.82, acc 0.80	No
Li 2021	DCIS vs atypia	Sens 0.91, spec 0.89, acc 0.90	Sens 0.70, spec 0.82, acc 0.80	No
Mercan 2021	ADH, 4-class ROI	Precision 0.82, recall 0.62, F 0.70	Overall 70% vs AI 72.7%	No
Mehta 2022	Atypia, 4-class ROI	AUC 0.80–0.83	Overall 0.70 vs AI 0.70	No
Tosun 2017	High-risk vs low-risk	Recall 0.65–0.69	Expert recall 0.77	No
Parvatikar 2020	High-risk vs low-risk	Recall 0.76	Expert recall 0.77	No
Parvatikar 2021	ADH vs rest; FEA vs rest	ADH recall 38–80%; FEA 35–68%	Not compared	No
Brancati 2022	ADH, FEA, 7-class ROI	ADH F1 24.0; FEA F1 64.0	Not compared	No
Li 2025	ADH, FEA, 7-class ROI	ADH F1 47.4; FEA F1 75.8	Not compared	No
Rinaldi 2026	ADH, FEA, 7-class ROI	ADH F1 0.27–0.46; FEA 0.59–0.71	Not compared	No
Pati 2022	ADH, FEA, 7-class ROI	ADH F1 40.4; FEA F1 74.2	3 pathologists: ADH F1 37.9; FEA F1 46.2	No

Study	Borderline task	AI result	Pathologists (same cases)	Uncertainty reported
Yamada 2016	ADH vs LG-DCIS	Acc 88.1% (ROI); 71.8–95.9% (case)	Not compared	No
Sandbank 2022	LG-DCIS/ADH vs HG/IG-DCIS	AUC 0.921 (0.878–0.965), external	Not head-to-head	No
Tahir 2025	Lobular neoplasia	AUC 0.953; sens 94.1%; spec 95.8%	Accuracy 97.1% → 100% with AI	No
Yoo 2026	ADH/LG-DCIS, 4-class	External tile AUC 0.70–0.81; patient AUC 0.61–0.70	Not compared	No

Acc, accuracy; sens, sensitivity; spec, specificity.

Generalisation to other hospitals

Only 2 of 18 reports tested the model on patients from a hospital not used for training. Sandbank and colleagues validated a commercial algorithm at two sites with different scanners and reported an AUC of 0.921 for separating low-grade DCIS or ADH from higher-grade DCIS [26]. Yoo and colleagues found that their H&E model dropped to a patient-level AUC of 0.61–0.70 at the external hospital [28]. All other reports used cross-validation or a held-out split from the same cohort.

Uncertainty and abstention

No report gave calibrated uncertainty estimates, prediction sets or an option to refer uncertain cases to a pathologist. Mercan 2018 discarded image windows with low classifier scores before classification [13], and two commercial systems displayed probability scores [26,27], but none measured whether these scores were reliable.

Risk of bias

QUADAS-2 judgements are summarised in Table 3. Risk of bias in patient selection was high in all 18 reports, because case sets were enriched for atypia and DCIS, drawn from curated public datasets, or chosen to contain ADH. Risk of bias for the index test and for flow and timing was unclear in 12 reports, mostly because thresholds, blinding or the handling of excluded cases and regions were not reported; one report was at high risk for the index test (random region-level splits) [25] and one for flow and timing (discordant regions discarded) [24]. The reference standard, an expert consensus in most studies,

was at low risk in all reports. Applicability concerns about the index test were low in 17 reports and unclear for the BRACS dataset paper, which did not evaluate a single index test.

Table 3. QUADAS-2 summary (number of reports, n = 18)

Domain	Low	High	Unclear
Risk of bias: patient selection	0	18	0
Risk of bias: index test	5	1	12
Risk of bias: reference standard	18	0	0
Risk of bias: flow and timing	5	1	12
Applicability: patient selection	0	18	0
Applicability: index test	17	0	1
Applicability: reference standard	18	0	0

Study-level judgements are given in Supplementary File 3.

Discussion

Main findings

This review found 18 reports of AI for borderline breast lesions on H&E slides, but they rest on only seven patient cohorts, and almost half come from a single 240-slide US cohort. The evidence is therefore much thinner than the number of papers suggests.

AI did well in narrow binary tasks. Separating DCIS from atypia, models matched or beat pathologists on the same cases [15,16]. This fits the idea that computers can measure nuclear and architectural features more consistently than the eye. In realistic multi-class settings, however, ADH was consistently the weakest class. On BRACS, the only public benchmark, ADH F1 stayed below 0.5 in all four studies [22–24,29]. In the one BRACS study that tested pathologists on the same images, three pathologists did no better (mean ADH F1 0.38) [29], which shows how hard ADH is for people and machines alike. Even so, a model that misses half of ADH lesions is not ready to triage biopsies.

Two further gaps matter for clinical use. First, only two reports tested a model at another hospital [26,28]. In one of them, an H&E model lost much of its accuracy at the external site [28]. Stain, scanner and case-mix differences are well known to degrade pathology AI, and borderline lesions, which depend on subtle features, are likely to be especially sensitive. Second, no report measured whether the model knew when it was unsure. For

lesions on which experts themselves agree only about half of the time [1], a model that gives a confident answer for every case is risky. Methods such as calibration and conformal prediction can give a model a statistically controlled way to say “I am not sure; refer to a pathologist” [7], but none of the included studies used them.

Comparison with earlier reviews

The earlier review of AI for DCIS searched to December 2022 and did not analyse atypia separately or assess risk of bias [8]. Our review adds borderline-specific results, citation searching, QUADAS-2 assessment, and a focus on external validation and uncertainty.

Implications for research

Future studies should:

1. Report per-class results for ADH, FEA and lobular neoplasia, not only overall accuracy.
2. Test models on slides from at least one other hospital and scanner.
3. Report calibration and an abstention or referral option, with the accuracy of the cases the model keeps.
4. Use whole slides or automatically selected regions, so that results reflect real workflow.
5. Share data and code, so that results can be compared on common test sets.

Strengths and limitations

Strengths of this review are a registered protocol, searches of three databases (one of which, OpenAlex, indexes engineering and computer-science venues), a supplementary Semantic Scholar search, a preprint search, backward and forward citation searching, and extraction of borderline-specific numbers from full texts.

There are several limitations. We could not search Embase, Scopus or Web of Science because we had no access to them; we partly offset this with Europe PMC, OpenAlex and Semantic Scholar, which added few new records. Eight reports could not be retrieved (Supplementary File 2), and one database report (Huang 2024) was assessed from its abstract and supplementary material only. One eligible study (Pati 2022) was missed at title-and-abstract screening and found only through the preprint search, so other BRACS studies whose abstracts do not mention atypia may have been missed. Preprints were excluded by protocol, so the preprint search was used only to look for missed or emerging work; it found three relevant preprints, each reporting borderline-lesion results only on the public BRACS dataset, and one of them (CorePath, 2026) adds a deferral option to report generation (Supplementary File 4). Screening was assisted by large language model tools; all 249 decisions were checked by the supervising author, but agreement (kappa) was measured only on the 45 records screened independently by two reviewers.

Some values were read from figures, and values for one study were taken from the authors' preprint version of the paper [29]. Risk-of-bias judgements were drafted with large language model assistance and finalised by one author (SG), not by two independent reviewers. Finally, tasks, classes and metrics varied too much for meta-analysis, and overlapping cohorts mean the reports are not independent.

Conclusion

AI for borderline breast lesions on H&E slides performs well in narrow tasks, such as separating DCIS from atypia, but poorly for ADH when several classes must be told apart. The evidence comes from few cohorts, rarely includes testing at another hospital, and never includes a reliable way for the model to flag cases it is unsure about. Multi-centre studies that report calibrated uncertainty and refer difficult cases to pathologists are needed before these tools are used in practice.

Statements and Declarations

Use of AI tools: Large language model tools were used to assist with study screening and data checking. All eligibility decisions were reviewed and finalised by the authors (GM and SG), who take full responsibility for the content.

Funding: None.

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics: Not required; this review uses published data only.

Data availability: The screening sheet, citation-search log, extraction tables and QUADAS-2 judgements are available as supplementary files.

Author contributions: GM: conception, searching, screening, data extraction, writing. SG: supervision, resolution of disagreements, review of the manuscript.

Supplementary files: (1) search strings; (2) screening sheet and citation-search log with exclusion reasons; (3) results tables and study-level QUADAS-2 judgements; (4) preprints found by the preprint search.

References

1. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. JAMA. 2015;313(11):1122–1132. doi:10.1001/jama.2015.1405

2. Jain RK, Mehta R, Dimitrov R, et al. Atypical ductal hyperplasia: interobserver and intraobserver variability. Mod Pathol. 2011;24(7):917–923. doi:10.1038/modpathol.2011.66
3. Sapino A, Marchiò C, Kulka J. “Borderline” epithelial lesions of the breast: what have we learned in the past three decades? Pathologica. 2021;113(5):354–359. doi:10.32074/1591-951X-374
4. Onega T, Weaver DL, Frederick PD, et al. The diagnostic challenge of low-grade ductal carcinoma in situ. Eur J Cancer. 2017;80:39–47. doi:10.1016/j.ejca.2017.04.013
5. Allison KH, Rendi MH, Peacock S, Morgan T, Elmore JG, Weaver DL. Histological features associated with diagnostic agreement in atypical ductal hyperplasia of the breast: illustrative cases from the B-Path study. Histopathology. 2016;69(6):1028–1046. doi:10.1111/his.13035
6. Schiaffino S, Calabrese M, Melani EF, et al. Upgrade rate of percutaneously diagnosed pure atypical ductal hyperplasia: systematic review and meta-analysis of 6458 lesions. Radiology. 2020;294(1):76–86. doi:10.1148/radiol.2019190748
7. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med. 2021;4(1):4. doi:10.1038/s41746-020-00367-3
8. Co M, Lau YCC, Qian YXY, et al. Artificial intelligence in histologic diagnosis of ductal carcinoma in situ. Mayo Clin Proc Digit Health. 2023;1(3):267–275. doi:10.1016/j.mcpdig.2023.05.008
9. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71
10. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–536. doi:10.7326/0003-4819-155-8-201110180-00009
11. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368:l6890. doi:10.1136/bmj.l6890
12. Geçer B, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. Pattern Recognit. 2018;84:345–356. doi:10.1016/j.patcog.2018.07.022
13. Mercan C, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans Med Imaging. 2018;37(1):316–325. doi:10.1109/TMI.2017.2758580
14. Mercan C, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. From patch-level to ROI-level deep feature representations for breast histopathology classification. In: Proc SPIE 10956, Medical Imaging 2019: Digital Pathology. 2019:109560H. doi:10.1117/12.2510665

15. Mercan E, Mehta S, Bartlett J, Shapiro LG, Weaver DL, Elmore JG. Assessment of machine learning of breast pathology structures for automated differentiation of breast cancer and high-risk proliferative lesions. JAMA Netw Open. 2019;2(8):e198777. doi:10.1001/jamanetworkopen.2019.8777
16. Li B, Mercan E, Mehta S, et al. Classifying breast histopathology images with a ductal instance-oriented pipeline. In: Proc 25th International Conference on Pattern Recognition (ICPR). 2021:8727–8734. doi:10.1109/ICPR48806.2021.9412824
17. Mercan C, Aygunes B, Aksoy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. Deep feature representations for variable-sized regions of interest in breast histopathology. IEEE J Biomed Health Inform. 2021;25(6):2041–2049. doi:10.1109/JBHI.2020.3036734
18. Mehta S, Lu X, Wu W, et al. End-to-end diagnosis of breast biopsy images with transformers. Med Image Anal. 2022;79:102466. doi:10.1016/j.media.2022.102466
19. Tosun AB, Nguyen L, Ong N, et al. Histological detection of high-risk benign breast lesions from whole slide images. In: MICCAI 2017. Lecture Notes in Computer Science, vol 10434. Springer; 2017:144–152. doi:10.1007/978-3-319-66185-8_17
20. Parvatikar A, Choudhary O, Ramanathan A, et al. Modeling histological patterns for differential diagnosis of atypical breast lesions. In: MICCAI 2020. Lecture Notes in Computer Science, vol 12265. Springer; 2020:550–560. doi:10.1007/978-3-030-59722-1_53
21. Parvatikar A, Choudhary O, Ramanathan A, et al. Prototypical models for classifying high-risk atypical breast lesions. In: MICCAI 2021. Lecture Notes in Computer Science, vol 12908. Springer; 2021:143–152. doi:10.1007/978-3-030-87237-3_14
22. Brancati N, Anniciello AM, Pati P, et al. BRACS: a dataset for BReAst Carcinoma Subtyping in H&E histology images. Database (Oxford). 2022;2022:baac093. doi:10.1093/database/baac093
23. Li X, Chen J, Luo B, et al. Classifying breast intraductal proliferative lesions via a knowledge distillation framework using convolutional neural network-based nuclei-segmentation-assisted classification (KDCNN-NSAC). Heliyon. 2025;11(10):e42756. doi:10.1016/j.heliyon.2025.e42756
24. Rinaldi AM, Russo C, Tommasino C. Hierarchical graph networks for breast cancer subtype classification. Data Min Knowl Discov. 2026;40(6). doi:10.1007/s10618-026-01259-y
25. Yamada M, Saito A, Yamamoto Y, et al. Quantitative nucleic features are effective for discrimination of intraductal proliferative lesions of the breast. J Pathol Inform. 2016;7:1. doi:10.4103/2153-3539.175380
26. Sandbank J, Bataillon G, Nudelman A, et al. Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. NPJ Breast Cancer. 2022;8(1):129. doi:10.1038/s41523-022-00496-w

27. Tahir M, Hu YP, Kumar H, et al. A comprehensive AI-based approach in classifying breast lesions: focusing on improving pathologists' accuracy and efficiency. Clin Breast Cancer. 2025;25(6):e818–e825. doi:10.1016/j.clbc.2025.03.016

28. Yoo J, Karthikeyan R, Kamat K, et al. Mechanosensitive TRPV4 immunohistochemistry improves deep learning-based classification of ductal carcinoma in situ beyond H&E morphology. J Pathol Clin Res. 2026;12(4):e70106. doi:10.1002/2056-4538.70106

29. Pati P, Jaume G, Foncubierta-Rodríguez A, et al. Hierarchical graph representations in digital pathology. Med Image Anal. 2022;75:102264. doi:10.1016/j.media.2021.102264

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BBLSupplementarymaterial20261022.docx
- BBLPRISMA2020checklist20261022.docx