Supplementary Table 1 Database search strategy

Database	Coverage	Limits/notes	Search strategy
PubMed/MEDLINE	Inception to 12 May 2026	No language restriction.	("Gallbladder Neoplasms"[Mesh] OR gallbladder cancer*[tiab] OR gallbladder carcinoma*[tiab] OR gallbladder neoplasm*[tiab] OR biliary tract cancer*[tiab]) AND ("Artificial Intelligence"[Mesh] OR artificial intelligence[tiab] OR machine learning[tiab] OR deep learning[tiab] OR radiomics[tiab] OR computer-aided diagnos*[tiab] OR convolutional neural network*[tiab] OR CNN[tiab] OR transformer*[tiab]) AND ("Diagnostic Imaging"[Mesh] OR "Ultrasonography"[Mesh] OR "Tomography, X-Ray Computed"[Mesh] OR "Magnetic Resonance Imaging"[Mesh] OR ultrasound[tiab] OR ultrasonography[tiab] OR CT[tiab] OR MRI[tiab] OR imaging[tiab])
Embase	Inception to 12 May 2026	No language restriction.	('gallbladder cancer'/exp OR 'gallbladder carcinoma':ti,ab OR 'gallbladder neoplasm*':ti,ab) AND ('artificial intelligence'/exp OR 'machine learning'/exp OR 'deep learning'/exp OR radiomics:ti,ab OR 'computer aided diagnosis':ti,ab OR cnn:ti,ab OR transformer*:ti,ab) AND ('diagnostic imaging'/exp OR 'ultrasonography'/exp OR 'computer assisted tomography'/exp OR 'magnetic resonance imaging'/exp OR ultrasound:ti,ab OR ct:ti,ab OR mri:ti,ab OR imaging:ti,ab)
Web of Science	Inception	Topic search.	TS=(("gallbladder cancer" OR "gallbladder

Core Collection	to 12 May 2026	carcinoma" OR "gallbladder neoplasm*" OR "biliary tract cancer") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR radiomics OR "computer-aided diagnosis" OR "convolutional neural network" OR CNN OR transformer*) AND (ultrasound OR ultrasonography OR CT OR MRI OR imaging))

Supplementary Table 2 Extended characteristics of the quantitative studies, including validation design, selected model, model-selection rationale, threshold handling, target contrast, and reconstruction notes

Study ID	Country	Design	Modality	Index AI model	Validation dataset	Target contrast	Reference standard	Validation N	Positive/negative	2x2 data source	Notes
S01	China	Retrospective diagnostic accuracy study	CT	ResNet50 deep learning	Internal validation	GBC vs benign gallbladder lesions	Pathology	84	33/51	Reconstructed from denominators and reported sensitivity/specificity	Threshold details incompletely reported
S02	India	Prospective diagnostic accuracy study	Ultrasound	Multiscale second-order pooling deep-learning classifier	Temporally independent prospe	GBC vs non-GBC	Histopathology, cytology, biopsy, or	273	183/90	Patient-level test set counts available	Compared with radiologists in original

					ctive test cohort		follow-up for benign cases				study
S03	China	Retrospective model development with external validation	CT	FCovResNet18 deep learning model	External validation	GBC vs xanthogranulomatous cholecystitis	Pathology	74	52/22	External cohort confusion matrix reconstructed from reported correct diagnoses	Imaging model used rather than combined nomogram
S04	India	Prospective independent test of	Ultrasound	RadFormer selected from RadFormer/	Independent test set	GBC vs xanthogranulomatous cholecystitis	Pathology or cytology	80	55/25	Exact confusion matrix available	RadFormer selected for higher

		pretrained models		GBCNet							AUC/specificity
S05	China	Prospective multi-institutional diagnostic accuracy study	Ultrasound	XGBoost ultrasound radiomics	External test A and B combined	Gallbladder carcinoma vs benign gallbladder masses	Pathology	104	22/82	External test counts combined to avoid multiple rows from one study	Radiomics/ML study

Supplementary Table 3 QUADAS-2 tailoring rules for AI imaging diagnostic accuracy studies

Domain	Signaling focus	Low-concern guidance	High/unclear-concern guidance
Patient selection	Representative patient spectrum	Consecutive or clearly defined clinical cohort; no inappropriate exclusions; not only highly selected public images	High concern if case-control, enriched differential, or unclear exclusions
Index test	AI test conduct and interpretation	Validation data independent from training/tuning; threshold/model selected before test evaluation; patient-level performance reported	High/unclear concern if post hoc thresholding, multiple models without prespecification, or possible data leakage
Reference standard	Correct classification of target condition	Histopathology/cytology/biopsy or clinical follow-up for benign disease; interpreted independently where reported	High/unclear concern if reference standard not described or incorporated index test
Flow and timing	Complete and comparable verification	All or nearly all validation patients included in 2 x 2 table; appropriate interval between index test and reference standard	High concern if exclusions after AI failure or missing reference standard are not explained
Applicability	Match to review question	Patient-level GBC detection using medical imaging and clinically relevant differential diagnoses	Concern if only image-level public dataset, non-GBC endpoint, polyp-only, staging, segmentation, or grading task

Supplementary Table 4 Reporting guideline mapping used for manuscript revision

Guideline/tool	Relevant content	Use in this manuscript
PRISMA 2020	Overall systematic-review reporting structure, abstract, flow diagram, search, synthesis, limitations	Used as general reporting framework
PRISMA-DTA	DTA-specific review question, index test, reference standard, 2 x 2 data, paired sensitivity/specificity synthesis	Primary reporting standard
PRISMA-S	Transparent reporting of database search strategies	Used for Supplementary Table S1
QUADAS-2	Risk of bias and applicability for diagnostic accuracy studies	Primary risk-of-bias tool, tailored to AI imaging
STARD 2015	Completeness of original diagnostic accuracy reporting	Used to guide extracted items from primary studies
CLAIM 2020/2024	AI imaging reporting completeness	Used as supplementary extraction logic, not as replacement for QUADAS-2
AMSTAR 2	Methodological quality concepts for reviews	Used as internal quality check only
GRADE-DTA	Certainty across a body of diagnostic-test evidence	Discussed narratively; formal rating not performed because of very small heterogeneous evidence base