

EMPIRICAL PAPER

The Promise and Pitfalls: A Systematic Review of AI-Powered Diagnostics in Modern Oncology

Akeem Moradehun Bamiro¹, Lateef Adebisi Odukoya², Ridwan Oladotun Ahmed³, Hafeez Abiola Afolabi⁴, Sameer Badri Al-Mhanna^{5,6}, Marwan Arkan Ghafoor⁷, Yusuf Wada⁸

Affiliations

¹Department of Anatomy, Faculty of Basic Medical Sciences, Lagos State University, Ikeja, 101233, Nigeria

²Translational Neuro-Oncology Laboratory, Department of Radiation Oncology Mayo Clinic Rochester MN USA.

³Locum Consultant Clinical Oncologist, Northwick Park Hospital & Mount Vernon Cancer Centre, East-North Hertfordshire NHS Trust, UK.

⁴Department of General Surgery, School of Medical Sciences, Universiti Sains Malaysia, Health Campus, Kelantan, 15160, Malaysia.

⁵Department of Physiology, School of Medical Sciences, Universiti Sains Malaysia, Health Campus, Kelantan, 15160, Malaysia.

⁶Center for Global Health Research, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.

⁷Department of Medical Laboratory Science, Komar University of Science and Technology, Sulaymaniyah 46001, Kurdistan Region, Iraq

⁸Department of Zoology, Faculty of Life Sciences, Ahmadu Bello University, Zaria, 810211, Kaduna, Nigeria.

ORCID Identifiers

Akeem Moradehun Bamiro
[0000-0002-9946-8206](https://orcid.org/0000-0002-9946-8206)

Lateef Adebisi Odukoya
[0000-0002-0427-7424](https://orcid.org/0000-0002-0427-7424)

Ridwan Oladotun Ahmed
[0000-0002-8152-5129](https://orcid.org/0000-0002-8152-5129)

Hafeez Abiola Afolabi
[0000-0002-1120-4100](https://orcid.org/0000-0002-1120-4100)

Sameer Badri Al-Mhanna
[0000-0001-9063-0185](https://orcid.org/0000-0001-9063-0185)

Marwan Arkan Ghafoor
[0009-0000-2544-7798](https://orcid.org/0009-0000-2544-7798)

Yusuf Wada
[0000-0003-0003-9581](https://orcid.org/0000-0003-0003-9581)

DOI

[10.65820/ejmsd-2vol2-issue2-2026](https://doi.org/10.65820/ejmsd-2vol2-issue2-2026)

Abstract

Purpose: Despite the rapid expansion of AI-powered diagnostics in oncology, most existing literature reports algorithm performance in isolated, retrospective settings, and no recent systematic review has jointly synthesised algorithm types, validation rigor, implementation barriers, and patient-outcome evidence. This leaves a gap in understanding how far AI diagnostics have actually progressed from development toward clinical use. This review addresses that gap by examining AI-powered oncology diagnostics across four areas: the types of AI algorithms developed, the validation techniques employed, the challenges and enablers of clinical implementation, and the evidence for improved patient outcomes.

Methodology: Following PRISMA guidelines, a systematic search of Scopus, BioMed Central, and Google Scholar identified peer-reviewed empirical studies published between 2020 and 2025. Of 1,265 records screened, 27 studies met inclusion criteria. Data extraction covered algorithm type, cancer type, dataset, validation technique, performance metrics, and implementation status; QUADAS-2 was used to assess risk of bias.

Results: Most studies used deep learning models, particularly CNNs, for imaging and histopathology diagnosis, with accuracies frequently exceeding 90% and AUCs up to 0.987; breast and lung cancers were most studied. However, validation was predominantly internal and retrospective, with only ~15% of studies reaching clinical trials or real-world implementation. Poor interpretability, limited external validation, and integration barriers constrain clinical adoption, alongside moderate-to-high risk of bias in patient selection and reference standards.

Novelty and Contribution: This review uniquely maps the gap between high-performing AI models and limited clinical adoption, offering a roadmap bridging development and implementation.

Implications: Interdisciplinary collaboration, prospective multicenter trials, and equitable, ethical integration frameworks are needed to advance clinical AI adoption in oncology.

Keywords: AI-powered diagnostics, oncology, algorithm, clinical implementations

*Corresponding author

E-mail address: bamiroakeem@yahoo.com or akeem.bamiro@lasucom.edu.ng

How to cite this article:

Bamiro, A.M., Odukoya, L.O., Ahmed, R.O., Afolabi, H.A., Al-Mhanna, S.B., Ghafoor, M.A., & Wada, Y. (2026). The Promise and Pitfalls: A Systematic Review of AI-powered Diagnostics in Modern Oncology. *Elicit Journal of Medical Science and Discovery*, 2(2), 17-36. <https://doi.org/10.65820/ejmsd-2vol2-issue2-2026>

1 Introduction

The World Health Organisation (WHO) and its International Agency for Research on Cancer (IARC) estimate that 9.8 million people died from cancer worldwide in 2024, while 20.6 million new cases were diagnosed (Sung et al., 2026). About 1 in 5 individuals will get cancer at some point in their lives, and 1 in 9 men and 1 in 13 women will die from the disease, making it the leading cause of disease burden worldwide. Cancer accounts for a significant share of deaths from non-communicable diseases (NCDs) globally and remains one of the top three causes of death for those aged 30 to 69 in most nations (Bray et al., 2024), underscoring its status as a major social, public health, and economic issue. Improving patient outcomes and lowering cancer-related mortality depend on the ability to identify cancer early, when treatment has a higher chance of success (Walter et al., 2025). Yet classic detection methods such as tissue biopsies, imaging, and tumour-specific screening tests frequently carry serious drawbacks, including invasiveness, high cost, and low sensitivity, particularly for malignancies that are asymptomatic or located in difficult-to-reach places (Rowe & Pomper, 2022; Tafazzoli et al., 2022).

These limitations have direct consequences for early detection and, by extension, survival, since delayed or missed diagnosis narrows the window for effective treatment. Closing this gap requires diagnostic tools capable of extracting subtle, high-dimensional patterns from imaging, histopathology, and genomic data at a scale and consistency that manual interpretation cannot reliably match. It is in this context that artificial intelligence (AI), an umbrella term spanning machine learning, neural networks, natural language processing, and rule-based systems has emerged as a candidate solution, offering the computational capacity to analyse large, complex datasets and detect patterns that may elude even experienced clinicians (Cabral et al., 2025). The AI-in-healthcare market reflects this momentum, projected to grow from US\$26.69 billion in 2024 to over US\$613.81 billion by 2034 (Cabral et al., 2025). Anticipated benefits include more accurate diagnoses, lower screening costs, broader healthcare access, and reduced physician workload (Rajpurkar & Lungren, 2023), particularly as the demand for imaging diagnostics has long outpaced the availability of specialists (Cabral et al., 2025).

Within oncology specifically, AI is increasingly positioned to transform diagnosis, biomarker evaluation, genomic characterisation, and treatment planning (Zhang et al., 2023), since AI algorithms can analyse large volumes of multi-modal data to surface signals that would otherwise be hard to discover (Hunter et al., 2022). AI systems incorporating electronic case data have already been shown to lower the missed diagnosis rate of lung cancer by 30.7% relative to standard clinical criteria (Lu et al., 2020), and AI-driven digital pathology and radiomics have improved tumour characterisation and prognosis prediction (Senthil Kumar et al., 2023). Despite these gains, a significant translational gap persists between algorithm development and routine clinical adoption. Oncology datasets are often scarce, heterogeneous, and biased, limiting model robustness and generalisability (Chua et al., 2021), while generalisability concerns, legal barriers, technical integration constraints, and clinician scepticism continue to impede clinical acceptance (Liu et al., 2022).

Prior reviews have examined AI in oncology from narrower vantage points for example, focusing on pathway-to-implementation barriers in general terms (Chua et al., 2021), summarising algorithmic applications across the oncology pipeline (Luchini et al., 2022), or restricting scope to a single cancer type such as ovarian (Maiorano et al., 2025) or digestive-tract cancers (Gao et al., 2025). This review differs by systematically and quantitatively mapping both the technical landscape (algorithm types, datasets, validation methods) and the clinical translation pathway (implementation stage, risk of bias) across a broad, multi-cancer sample, and by explicitly classifying studies into a four-stage implementation typology, asynthesis not offered by prior narrower or narrative reviews.

Given this background, a rigorous and critical examination of the existing evidence is needed to understand how AI technologies are being developed, validated, and applied in oncology diagnosis. This review synthesises the literature to investigate the kinds of AI algorithms currently in use, the techniques used to validate and test them, the obstacles and enablers to their clinical application, and the extent to which they have improved patient outcomes. Specifically, this review is guided by the following research questions:

- What types of AI algorithms are being developed for oncology diagnostics?
- How are these AI models validated and tested?
- What are the challenges and enablers for clinical implementation of AI in oncology?
- What evidence exists for improved patient outcomes due to AI-powered diagnostics?

This review attempts to provide insights for academics, clinicians, and policymakers who want to close the gap between innovation and clinical application by methodically examining the data and advancing our understanding of the current level of AI in oncology diagnosis.

2 Materials and Methods

In order to investigate the kinds of AI algorithms currently used for cancer detection, this systematic review was carried out in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria (Page et al., 2021). The PRISMA flowchart for the listed research is displayed in Figure 1.

2.1 Search Strategy and Database

In addition to a manual search on Google Scholar, this study used two significant digital databases, Scopus and BioMed Central (BMC), to do a thorough search for peer-reviewed journal publications. Because of their extensive coverage of research on AI-powered diagnostic tools in oncology and their relevance to the subject, Scopus and BMC were chosen. In order to guarantee the inclusion of current and pertinent studies that reflected recent developments in AI technologies, the search was restricted to literature published between 2020 and 2025. These data sources made it possible to compile a substantial amount of academic literature, which raised the review's legitimacy and dependability. The inclusion of high-quality publications was also guaranteed by the use of two reputable databases, enhancing the validity of this systematic literature analysis and offering significant insights that add to the expanding corpus of knowledge in oncology research (Maiorano et al., 2025).

A thorough literature search was carried out by two independent reviewers using the two databases that were chosen. Only peer-reviewed publications and empirical research published between 2020 and 2025 were included in the search method for the targeted articles. The first search was launched on November 21, 2020, and it was revised on April 30, 2025, to include pertinent research that the first search might have overlooked (Zobair et al. 2023). The search terms deep learning, oncology, diagnosis, diagnostic, AND clinical implementation and "artificial intelligence, machine learning, and article extraction" are utilised in response to the PRISMA statement. A combination of terms, including operators AND and OR, were employed, such as ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("oncology" OR "cancer") AND ("diagnosis" OR "diagnostic") AND ("clinical implementation"). After a full-text assessment, the papers that were judged relevant were evaluated for eligibility using predetermined inclusion and exclusion criteria.

2.2 Eligibility and Inclusion Criteria

Strictly peer-reviewed published empirical journal papers in English that were relevant to the major body of research were included in this review (inclusion criteria); duplicates and unrelated materials were removed (exclusion criteria). Additionally, papers involving AI and machine learning applications, algorithm development, clinical trials, or deployment were considered in this study. Studies were excluded if they are not published in English, duplicate records across databases, non-empirical publications (editorials, commentaries, opinion pieces), conference proceedings, book chapters, and research notes lacking sufficient methodological detail, studies addressing AI or oncology in isolation, without an explicit application of AI to cancer diagnosis, studies unrelated to cancer diagnosis or detection (e.g., treatment-only or purely epidemiological studies) and studies published outside the 2020–2025 search window. In order to get the most pertinent items for analysis and synthesis, the authors conducted three rounds of screening and filtering procedures, excluding all extraneous publications, book chapters, conference papers, and research notes. This review upheld a high standard of article selection by using rigorous inclusion and exclusion criteria as well as careful screening and filtering procedures, which helped to ensure the correctness of the study's conclusions. The main themes of this study and the analysis's findings are covered in the section that follows.

2.3 Data Extraction

Two reviewers independently extracted the data using a standardised extraction form. Details of the AI approach and algorithm (architecture, datasets used), performance metrics (AUC, sensitivity, specificity, HR for OS and PFS), and an

overview of the study (authors, year of publication, country, study design, sample size, and cancer types) were among the extracted data. When practical, the validation approach internal, external, or cross-validation—was also recorded. To classify and comprehend the clinical validations, trials, and diagnostic accuracy, a narrative synthesis was also carried out. Table 1 displays a thorough data extraction table.

2.4 Risk of Bias Assessment

The chosen study's risk of bias was evaluated separately by three reviewers using QUADAS (Quality Assessment of Diagnostic Accuracy Studies)-2, a tool used for studies assessing diagnostic accuracy (Whiting et al., 2011). The QUADAS-2 method was used to evaluate risk bias in 26 of the 27 studies. Four areas of bias were investigated: data transparency, outcome measurement, model development, and patient selection. Any differences in evaluation were settled by discussion, and the studies were categorised as having a low, moderate, or high risk of bias.

3 Results

Initially, a total of 1265 studies were found using both manual and database searches. 628 full-text publications were examined for eligibility after duplicates were eliminated and titles and abstracts were evaluated for relevancy. 401 of this research were disqualified for the reasons listed below: Of the 258 non-empirical studies, 248 focused on either AI or cancer, but not both, and 95 were books, book chapters, or conference proceedings that lacked enough methodological information (Figure 1). Ultimately, 27 studies were examined after meeting the inclusion requirements. While Nittas et al. (2025) provide expert opinion on the usage of AI-powered tools in cancer diagnosis, 26 of the papers used AI algorithms for cancer diagnosis.

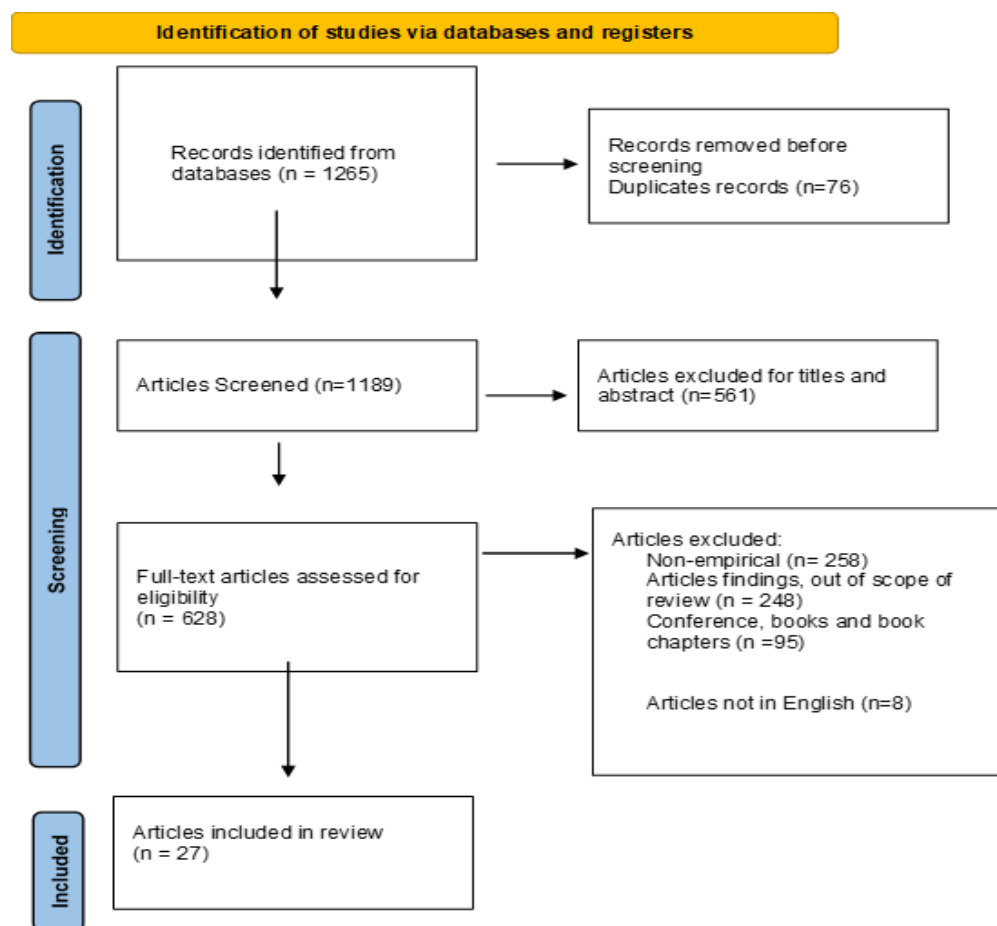


Figure 1 PRISMA flow diagram.

Table 1. Data Extraction table

S/ N	Author (Year)	Country	Cancer Type	AI Method / Algorithm	Dataset Description	Sample Size	Performance metrics	Stage of Implementation	Key Findings	Limitations Noted
1	Li et al. (2025)	Not specified	Gastric Cancer	AI-enhanced analytics for scRNA-seq and transcriptome data (specific model not stated)	TCGA (STAD) + GEO datasets: GSE183904, GSE84437, GSE26253 (public)	Not explicitly stated (TCGA and GEO samples)	Not numerically reported; mentions immune-related survival model	Development and partial validation (not yet in clinical use)	A survival model was created after APOE was discovered to be a crucial gene influencing the immunosuppressive tumour microenvironment through M2 macrophages.	Absence of clinical trials, unclear AI methodology, and an externally validated model
2	Alabi et al. (2025)	Sweden	Head and Neck Squamous Cell Carcinoma (HNSCC)	TabNet (DL), XGBoost, Voting Ensemble	Patient records from 3 Swedish university hospitals	419 patients	Accuracy: 88.1% (TabNet & XGBoost), 88.7% (ensemble)	Development & partial validation	ML models used integrative features to predict OS; important predictors included age, haemoglobin, cancer stage, and p16.	Needs clinical trial testing and external validation; there isn't a deployment yet.
3	Peeters et al. (2025)	Netherlands, Denmark, USA	Lung cancer	Ensemble of 2D & 3D CNNs with Mahalanobis distance out-of-distribution (OOD) detection	Multi-source CT scan datasets: NLST (ID), DLCST, RadboudUMC, DLR scans, and Pancreas dataset (OOD)	Over 10,000 scans across 5 datasets	High OOD detection performance across near- and far-OOD; no decline in model accuracy with increasing OOD score	Development & validation	The Mahalanobis-based OOD approach increases the safety and resilience of DL models in the diagnosis of lung cancer.	External generalisability to other clinical workflows is not fully established, and no patient outcomes have been evaluated.
4	Meshkov et al. (2025)	Not specified	Lung cancer	Logistic regression (L1/L2), Gaussian processes, kernel logistic regression	NGS data of cfDNA from 148 healthy + 138 LC subjects	286 total	AUC: 0.872–0.875; Sensitivity: 66.7–100%; Specificity: 79.3–90%	Development stage; proposed for clinical decision support	created a highly interpretable AI model for LC detection from cfDNA that is independent of clinical metadata and tumour biomarkers.	More validation is required; there is currently no real-world clinical implementation.
5	Sharma & Parvathi (2025)	Not specified	Cervical cancer	Hybrid DL model (DenseNet201 + InceptionV3); fine-tuning & feature fusion	Preprocessed medical image dataset (details unspecified)	Not reported	Accuracy: 96.54%, Precision: 95.91%, Recall: 96.44%, F1: 96.17%	Development stage, strong performance; no deployment	Outperformed ResNet-50, DenseNet201, Xception; effective early diagnostic classifier	Dataset source/size not stated; not yet validated in real-world settings
6	Wallis et al. (2025)	Netherlands	Pediatric cancers (52 tumor types, 96 subtypes)	Ensemble-based ML classifier trained on RNA-seq data	Internal (PMC biobank, 1256 samples), External	1256 training; 704 test samples	Precision: 99% (internal), 98% (external); Recall: 76%	Near-clinical readiness; diagnostic utility	M&M supports the diagnosis of rare cancers by accurately classifying a	Absent of retinoblastoma; does not report

					(KidsFirst, 704 samples)		(internal), 77% (external)	demonstrate	variety of paediatric tumour types.	treatment results after diagnosis
7	Ruff et al. (2024)	Germany	IDH-mutant gliomas (brain tumors)	Deep Learning-based MRI Reconstruction (DLR)	25 MRI scans from glioma patients	25 patients	Qualitative preference (84–100%), higher SNR/CNR ($p < 0.001$); 29.6% time savings	Early clinical validation; implementation-ready	Clinicians preferred DLR because it offered faster scan times, higher-quality images, and no sacrifice in diagnostic utility.	Lack of long-term outcome data, small sample size, and non-randomized study design
8	Li, Wang and Wang (2024)	China	Bladder cancer	LASSO, SVM-RFE	GEO datasets: GSE13507, GSE3167, GSE37816; 1513 mitochondrial-related genes from MSigDB	Not explicitly stated, but based on multiple GEO datasets	Not numerically detailed, but confirmed diagnostic accuracy and gene relevance	Development and validation; not yet deployed	Four diagnostic genes (GLRX2, NMT1, OXSM, and TRAF3IP3) have been identified; the model predicts BC with diagnostic potential; NMT1 has been validated as a possible target in laboratory tests.	lacks real-world testing or trial data; needs clinical and external validation
9	Carles et al. (2024)	Europe	Lung cancer	nnU-Net (deep learning segmentation)	PET/CT images from 2 prospective trials (PET-Plan, STRIPE) and 1 retrospective cohort	560 images for training, 396 for testing (PET and CT, 3D & 4D)	PET: DSC (4D) = 0.74 ± 0.06 , (3D) = 0.82 ± 0.11 ; CT: DSC (4D) = 0.61 ± 0.28 , (3D) = 0.63 ± 0.34	Development and external validation stage; open-source release	AI models support 3D/4D integration in clinics, reduce inter-observer variability, and perform better than manual segmentation.	Absence of direct outcome data; deployment requires additional real-world testing
10	Li et al. (2024)	China	Breast cancer	YOLOv5 for detection + GOLO-CMSS (CNN-based segmentation)	USTC (28,477 images), UDIAT, BUSI, TUH (public)	28,477 (USTC) + 3 public datasets	YOLOv5 AP: up to 99.41%; GOLO-CMSS DSC: up to 93.19%; better than radiologists ($p < 0.001$)	Development with high potential for clinical implementation	When it came to identifying and classifying breast masses, AI was on par with or better than radiologists; it was reliable across several datasets.	Clinical integration procedures are required; not tested in prospective or real-time clinical settings
11	van Dooijewert et al. (2024)	Netherlands	Breast cancer	Visiopharm 'Metastasis Detection' App	Whole-slide images (H&E) of sentinel lymph node (SN) specimens	190 SN specimens (100 AI-assisted, 90 control)	Adjusted RR of IHC use: 0.68; ~€3,000 cost savings; up to 30% ↑ sensitivity; time reduction	Clinical implementation trial	AI improved diagnostic sensitivity while lowering pathologist time and IHC use.	Single-center trial; not randomized; patient outcome not measured
12	Morel et al. (2023)	France	Breast (n=719), Lung (n=541),	Convolutional Neural Networks (CNNs)	TCGA BRCA, LUAD, COAD histopathology image datasets	1,719 (combined)	Sensitivity up to 95%; PPV up to 90.6% in top 10% risk group;	Development and proposed diagnostic strategy	CNN-based image analysis can maintain high mutation detection sensitivity, prioritise	Limited datasets, no external validation, results based on mutation prevalence, and

			Colon (n=459)				sequencing reduction up to 49.9%		sequencing, and use fewer resources.	the use of conventional (not sophisticated) CNNs
13	Hill et al. (2023)	Not specified	Mantle Cell Lymphoma (MCL)	Gradient-boosted ML model; logistic regression and survival modeling	Curated clinical, genomic, and cytogenetic data from 862 MCL patients (2014–2022)	862 patients	AUC ROC = 0.83; feature robustness analysis	Development with prototype app interface; close to clinical readiness	Created the iMIPI and iMIPI-s indices with 20 and 10 of the best predictive features, respectively; the best predictors were TP53 mutations, LDH, and Ki-67.	There is currently no external validation, no discussion of the regulatory pathway, and no direct measurement of outcome improvement.
14	Ternov et al. (2022)	Denmark	Melanoma and other skin lesions	Convolutional Neural Networks (CNNs)	ISIC-2019 (25,331 images, open-source); AISC-2021 (26,591 images, novel dataset)	~51,922 total images	Melanoma misclassification: 31.1%; Triage sensitivity: 99.7%, specificity: 8.2%	Algorithm development and retrospective validation	CNNs misclassified a large number of melanomas, demonstrated high triage sensitivity but low specificity, and underlined the dangers of standalone AI.	Low specificity, restricted ability to generalise to new data, and unsuitability for clinical implementation
15	Liu et al. (2022)	USA	Oral epithelial dysplasia	Deep learning (CNN) using DeepLabv3+ and UNet+	Whole-slide pathology images annotated by expert pathologists	112 annotated cases (training/validation), 44 cases (test set)	Accuracy: 93.3%, F1-score: 90.9% (DeepLabv3+)	Algorithm development and validation (preclinical stage)	DeepLabv3+ The CNN model establishes the foundation for computer-assisted diagnosis by precisely detecting and segmenting OED.	Small dataset, no external validation, no clinical trial, and no real-time implementation as of yet
16	Jiang et al. (2022)	China	Lung cancer	Deep Learning Image Reconstruction (DLIR)	203 patients; 1066 nodules; ULD CT at 0.07 or 0.14 mSv + standard contrast-enhanced CT (CECT) for comparison	203 patients, 1066 nodules	Nodule detection: FBP = 62.5%, ASIR-V = 73.3%, DLIR-H = 75.8%; noise reduction also quantified	Clinical evaluation and validation stage (real-world imaging settings)	In ULD CT, DLIR-H performed noticeably better than ASIR-V and FBP in terms of noise reduction, nodule detection, and diameter measurement accuracy.	Only noncontrast CT was used to evaluate the AI method; no research was done on its long-term effects on clinical decision-making or results.
17	Zhao et al. (2020)	China	Colorectal Cancer (CRC)	Deep learning – CNN using transfer learning for 9-class tissue segmentation	Whole-slide H&E-stained pathology images (discovery: N=499,	814 total	Prognostic value (HR 1.72 and 2.08); improved model performance	Development + clinical prognosis validation (not yet deployed)	CNN-measured By classifying OS risk, TSR, an independent prognostic marker, can help with clinical decision-making.	Deployment of external clinical trials has not yet been completed, and generalisability

					validation: N=315)		when integrated with Cox model			outside of CRC has not been examined.
18	Sartor et al. (2020)	Sweden	Cervical and Anorectal Cancers	Convolutional Neural Network (CNN) for CT segmentation	CT images and structure sets (2014–2018) from Skåne University Hospital	75 cervical, 191 anorectal cancer cases	Dice: 0.91– 0.94 (femoral heads), 0.84– 0.94 (bladder), 0.82–0.88 (bowel bag, CTVNs); MSD: ~1.5–6.8 mm	Developmen t and internal validation; not yet implemented clinically	CNN segmented some organs with high overlap, which could lessen the workload for oncologists.	Restricted by inconsistent annotations in training data; lymph node performance is lower; additional data is required
19	Gao et al. (2024)	China	Lung Cancer	3D Convolutional Neural Network (DenseNet)	Retrospective low-dose CT scan data (2011– 2013); local participants aged 40–74	570	AI AUC: 0.921; Panel AUC: 0.873; Kappa = 0.883; CCC = 0.809	Pilot evaluation; not deployed in clinical workflows	AI can help with the diagnosis of pulmonary nodules and has shown excellent agreement with expert evaluations.	Retrospective data only; no prospective or real-time clinical trials; little patient diversity observed
20	Deng et al. (2024)	China	Prostate cancer	AI-guided MRI fusion for targeted biopsy (AI-cTB)	380 patients in a prospective randomized controlled trial (RCT)	380 (191 AI-cTB, 189 cTB)	csPCa detection: AI- cTB (69.39%) vs. cTB (49.71%), $p < 0.001$	Fully integrated in clinical workflow for trial; potential for broader use	AI-cTB greatly enhanced identification of clinically relevant PCa; effective clinical application of AI tool	Study conducted at a single institution; details of the AI algorithm are not made public; generalisability to multi-center settings is not yet verified.
21	Zhang et al. (2023)	China	Prostate Cancer	AutoML using Autogluon; Logistic Regression, Random Forest, XGBoost	Multicenter data from 9 hospitals; demographics, lab, clinical, ultrasound data	4,747 cases (4,312 biopsy cases; 108 features)	AUC: 0.853 (train), 0.820 (validation), 0.807 (Changhai), 0.850 (Zhongda)	Clinical deployment in testing; not yet globally validated	PCAIDS outperformed PSA/fPSA-tPSA in predicting csPCa; avoided up to 32.2% of unnecessary biopsies with <5% csPCa missed	Needs more extensive prospective validation; currently only available for Chinese cohorts.
22	Pan et al. (2023)	China	Bladder Cancer	Deep Learning (convolutional neural network (CNN) based on ScanNet	854 Whole Slide Images (WSIs) from 692 patients	854 images / 692 patients	AUC: 0.878 (patch); 0.870 (WSI); 0.847 (pathologist comparison); High interpretability	Developmen t and retrospective validation	Useful for transurethral resection, accurate, and not less accurate than skilled pathologists in predicting invasion depth and histologic grade	Retrospective, single-center, not yet generalisable, heavy labelling burden, inability to identify carcinoma in situ, and future plans to broaden the dataset and scope

23	Li et al. (2023)	China	Laryngopharyngeal Cancer (LPC)	Deep Convolutional Neural Network (CNN)	31,543 internal laryngoscopic images; 25,063 external images from 5 hospitals; 551 videos (real-time evaluation)	2,382 patients (internal); 5 hospitals (external)	Accuracy: 0.956 (internal), 0.949 (videos); AUC: 0.965–0.987; Accuracy vs experts: 0.940	Real-time tested; not yet routine clinical use	High real-time diagnostic accuracy; consistent across clinical settings; LPAIDS performed better than or on par with human laryngologists, especially non-experts.	Since LPAIDS is a retrospective study, it cannot fully represent real-world clinical use.
24	Akintola et al. (2025)	Not specified	Lung cancer	Hybrid DL model: Mask R-CNN (segmentation) + CNN (classification)	Curated public datasets; 1000 images, expanded to 4000 via augmentation	4000 images (post-augmentation)	Accuracy: 95.6%, Precision: 94.8%, Recall: 94.3%, F1-score: 97.5%, AUC: 94.5%	Development and validation (no clinical deployment yet)	The hybrid model performed better than the baseline. CNN and Mask R-CNN alone; potential for high accuracy and scalability	Small dataset; no clinical or external validation; Not yet examined in a clinical context
25	Zhang et al. (2024)	China	Breast cancer	LightGBM (gradient boosting decision tree ML model)	Clinical blood biomarker data from 239 patients across two centers	239	AUC: 0.945 (train), 0.892 (internal), 0.908 (external); high accuracy, sensitivity, specificity	Late development / preclinical validation	De novo bone metastasis in breast cancer was accurately predicted by the LightGBM model; key characteristics included CA153, CEA, ApoB, CK, etc.	Limited sample size; testing is required for generalisability and geographic diversity; clinical decision support system deployment is not yet complete.
26	Lu et al. (2020)	USA	Lung cancer	convolutional neural network (CXR-LC)	data commonly available in the electronic medical record (EMR) (chest radiograph, age, sex, and whether currently smoking).	Cancer Screening Trial ($n = 41\,856$),	(PLCO AUC, 0.755 vs. 0.634; $P < 0.001$).	Clinical trial	The model improved CMS eligibility criteria, according to the study, with an AUC of 0.755 as opposed to 0.634. It also attained parity with more complex risk scores that required 11 data points. (PLCOM 2012)	Validation outside of a clinical setting in lung cancer screening trials.

3.1 Overview of Included Studies

The 27 studies in this systematic review, which were published between 2020 and 2025, demonstrate the quick development and rising interest in artificial intelligence (AI) applications for oncology diagnostics over the previous five years. With the bulk of studies ($n = 16$, 59.25%) published in 2024 and 2025, the distribution of publication years clearly demonstrates an increase in research activity, suggesting a recent uptick in clinical and computational innovation in the field.

Regarding the study's geographic distribution, it was carried out in a variety of places, with a significant global presence. With 10 studies (37.07%) contributed, China became the top country in AI-oncology research, demonstrating the nation's strong focus on AI-driven healthcare innovation. In cooperation with other nations like Denmark and the USA, the Netherlands provided three studies. Strong research efforts in Europe and North America are reflected in the two to three studies that the USA, Sweden, and Denmark each participated in. In addition to cross-national studies involving several institutions, Germany and France are among the other nations represented. Some studies ($n = 8$) inferred the country from institutional affiliations or dataset origin rather than explicitly stating it.

The 27 studies cover a wide variety of cancer types, demonstrating the adaptability of AI techniques in oncological fields. As seen in figure 2, lung cancer was the most commonly studied cancer type ($n = 8$), followed by breast cancer ($n = 4$), prostate cancer ($n = 2$), bladder cancer ($n = 2$), and cervical cancer ($n = 2$). Other cancer types were only studied once.

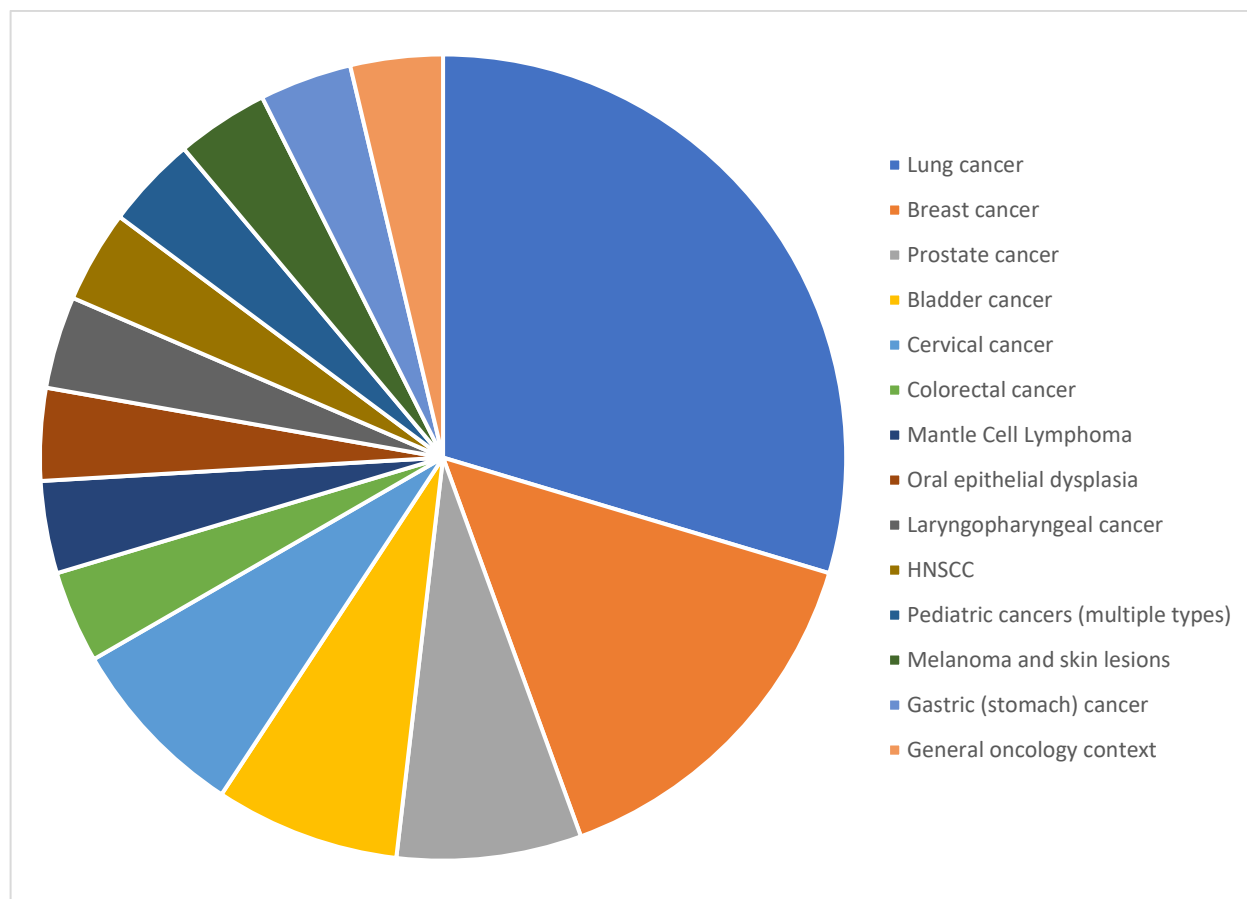


Figure 2 Cancer types identified in the reviewed articles

Depending on the type of cancer, study design (retrospective, prospective, clinical trial), and data modality (imaging, genomics, clinical), sample sizes varied significantly amongst the studies. These are a few important patterns. For instance, thousands of patients or data points were used in large-scale studies. More than 28,477 ultrasound pictures were used by Li et al. (2024), 4,747 prostate cancer cases from nine hospitals were used by Zhang et al. (2023), and more than 10,000 were used by Peeters et al. (2025).

CT scans from five different datasets. Hundreds to low thousands of patients or samples were included in moderate-scale studies. For instance, Hill et al. (2023) examined 862 MCL patients, Meshkov et al. (2025) assessed 286 cfDNA samples, and Carles et al. (2024) included 396 test images and 560 training images. Finally, smaller-scale studies ($n < 100$) were usually pilot evaluations, qualitative studies, or early development studies. For instance, Sartor et al. (2020) analysed 266 CT cases for classification, Liu et al. (2022) used 112 training/validation and 44 test pathology images, Ruff et al. (2024) used 25 MRI scans for validation, and Nittas et al. (2025) conducted expert interviews in a qualitative framework. Some studies described dataset sources, such as multi-center hospital networks or public repositories (TCGA, GEO, and ISIC), but they did not provide exact sample sizes.

3.2 AI Methods/Algorithm Details

The reviewed studies show how advanced and varied AI algorithms used in oncology diagnostics can be. These models include hybrid frameworks, sophisticated deep learning architectures, and traditional machine learning techniques.

3.2.1 Common AI Models

Because of their efficiency in processing imaging and histopathology data, deep learning (DL) techniques in particular, Convolutional Neural Networks (CNNs) were used in the majority of studies. Numerous other sophisticated and comprehensible models were also well-known. At least 13 out of the 27 studies used CNNs, making them the most widely used AI models. They were used for image segmentation and classification (e.g., bladder, melanoma, colorectal, lung, and breast cancers). Whole-slide images (WSIs) and radiography image enhancement (such as Deep Learning Image Reconstruction, or DLIR) are used in histopathology slide analysis. These include YOLOv5 + GOLO-CMSS (Li et al., 2024 – Breast cancer), DenseNet, InceptionV3 (Sharma & Parvathi, 2025 – Cervical cancer), UNet, DeepLabv3+, nnU-Net for segmentation tasks, and Mask R-CNN in conjunction with CNNs (Akintola et al., 2025 – Lung cancer).

In order to increase predictive power, a number of studies used ensemble techniques, including stacked CNN ensembles (Peeters et al., 2025 – Lung cancer; Wallis et al., 2025 – Paediatric tumours) and voting ensembles, XGBoost, and TabNet (Alabi et al., 2025 – HNSCC). Furthermore, LASSO and SVM-RFE (Li, Wang & Wang, 2024 – Bladder cancer), Logistic Regression, Gaussian Processes (Meshkov et al., 2025 – Lung cancer), and Gradient Boosting (LightGBM, AutoML, XGBoost) (Zhang et al., 2024; Hill et al., 2023) all used machine learning (ML) models, frequently in conjunction with feature selection techniques. Other studies made use of clinical-grade AI tools or applications, such as the Visiopharm 'Metastasis Detection' App (van Doijeweert et al., 2024) and AI-cTB for MRI-guided biopsy (Deng et al., 2024).

3.2.2 Dataset Used

Numerous data sources, including institutional, multicenter, and publicly accessible datasets, were used in the studies. More than half of the studies used open-access datasets, such as NLST, PLCO, national datasets used in lung cancer screening (Lu et al., 2020; Peeters et al., 2025), ISIC-2019, AISC-202, for dermoscopic skin lesion classification (Ternov et al., 2022), and TCGA (The Cancer Genome Atlas), used for gene expression and histopathology data (e.g., Morel et al., 2023).

Numerous studies made use of institutional or proprietary datasets from clinics and hospitals, including Princess Máxima Centre (Wallis et al., 2025 – paediatric tumours), Swedish University Hospitals (Alabi et al., 2025 – HNSCC), PET-Plan, STRIPE Trials (Carles et al., 2024 – lung cancer segmentation), and prospective MRI studies (Deng et al., 2024 – prostate cancer).

Some studies made use of extensive private datasets. For example, Li et al. (2024) used 28,477 ultrasound images from the USTC to detect breast cancer, and Li et al. (2023) used 25,000 external images and more than 31,000 laryngoscopic images to diagnose LPC. Custom-curated or newly labelled datasets were also used; for instance, Pan et al. used 854 whole-slide images from 692 bladder cancer patients, and Meshkov et al. built models on cfDNA NGS data from lung cancer and healthy subjects.

3.2.3 Training and Validation Techniques

Across studies, strong training and validation techniques were used to evaluate the generalisability and reliability of AI models. To reduce overfitting and evaluate model stability, K-fold cross-validation (e.g., 5-fold, 10-fold) was widely employed. On extensive, multi-center imaging datasets, Peeters et al. employed 10-fold CV, while Carles et al. used 5-fold CV.

While external validation used entirely separate datasets, frequently from different institutions, internal validation was carried out using training/test splits or bootstrapping on the same dataset. For instance, Zhang et al. (2023) validated their AutoML model on two external hospital cohorts, and Wallis et al. validated their paediatric cancer classifier on KidsFirst external data.

Few studies proceeded to possible real-world validation. Van Dooijeweert et al. (2024) carried out a non-randomized clinical trial for AI-assisted metastasis detection in breast cancer pathology, while Deng et al. (2024) carried out a randomised controlled trial (RCT) on MRI-guided prostate biopsy. Lastly, in order to evaluate real-world competitiveness, a number of models were compared to human experts. Li et al. (2024) evaluated AI breast mass detection against radiologists, and Li et al. (2023) used real-time videos to validate the LPC AI system against human laryngologists.

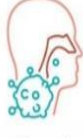
AI-Powered Diagnostics in Oncology Cancer types and AI Algorithms for Diagnosis		
 Lung Cancer CNN, Logistic, regression, deep learning, hybrid DL	 Breast Cancer CNN, Visiopharm, gradient boosting decision tree ML model	 Prostate Cancer AI-guided MRI fusion, random forest
 Bladder Cancer CNN, Lasso	 Cervical Cancer Hybrid DL model, CNN	 Colorectal Cancer CNN
 Mantle cell lymphoma Gradient-boosted ML model	 Oral epithelial dysplasia CNN	 Laryngopharyngeal Cancer CNN
 Head and neck cancer TabNet, XGBoost	 Pediatric Cancer Ensemble-based ML classifier	 Melanoma and other skin lesions CNN
	 Gastric Cancer AI-enhanced analytics	

Figure 3 Cancer type and AI models employed (Source: Author creation)

Table 2 Categories of AI algorithm and observations

Category	Observations	Source
Most common AI models	CNNs (including YOLOv5, UNet, DeepLabv3+, Mask R-CNN); Ensemble methods; LightGBM; AutoML; Logistic Regression; TabNet	Li et al. (2024), Liu et al (2022), Zhang et al. (2024), Meshkov et al. (2025), Alabi et al. (2025)
Data sources	Public repositories (TCGA, GEO, ISIC); Institutional clinical datasets; National screening cohorts (NLST, PLCO); Custom-curated datasets	Morel et al. (2023), Li, Wang and Wang. (2024), Peeters et al. (2025), Lu et al. (2020)
Validation approaches	Internal (train-test split, cross-validation); External (independent cohort testing); Prospective trials (e.g., RCTs); Expert comparisons	Akintola et al. (2025), Peeters et al. (2025), van Dooijeweert et al. (2024)
Notable datasets	USTC (28,477 images), KidsFirst (pediatric cancers), PET-Plan/STRIPE, cfDNA NGS, NLST, GEO gene expression sets, Whole-slide pathology images (WSIs)	Li et al. (2024), wallis et al. (2025), Meshkov et al. (2025), Peeters et al. (2025), lu et al. (2020)

3.3 Diagnostic Accuracy

A variety of performance metrics were used in the reviewed studies to evaluate the diagnostic accuracy of AI models. AUC (Area Under the Curve), F1-score, Dice Similarity Coefficient (DSC), accuracy, sensitivity, specificity, and other metrics provide information about the predictive accuracy and resilience of AI tools in a variety of oncology domains.

At least one performance metric was reported in most of the studies. Using a hybrid DL model (DenseNet201 + InceptionV3), Sharma & Parvathi (2025) reported an accuracy of 96.54% for the classification of cervical cancer. Similarly, Akintola et al. (2025) used a hybrid Mask R-CNN + CNN model to classify lung cancer with 95.6% accuracy. However, Li et al. (2023) found that their internal accuracy for detecting laryngopharyngeal cancer in more than 30,000 images was 0.956.

Using LightGBM, Zhang et al.'s (2024) study on breast cancer yielded area under the curve (AUC) values of 0.945 (train), 0.892 (internal), and 0.908 (external). Meshkov et al. (2025) showed a strong discriminatory ability with an AUC of 0.872–0.875 for lung cancer cfDNA. Utilising the CXR-LC model, Lu et al. (2020) (lung cancer screening) outperformed CMS eligibility (0.634) and matched PLCOM2012 (0.751) with an AUC of 0.755. According to Pan et al. (2023), the expert comparison was 0.847, the WSI-level was 0.870, and the patch-level AUC was 0.878 for bladder cancer. An AUC ROC of 0.83 was obtained for mantle cell lymphoma by Hill et al. (2023). However, the AUC range of 0.965–0.987, which is exceptionally high for real-time video evaluation, was produced by Li et al. (2023) (LPAIDS for LPC).

Ternov et al. (2022) demonstrated a high false-positive rate with a triage sensitivity of 99.7% and specificity of 8.2%. Meshkov et al. (2025) obtained a specificity of 79.3–90% and a sensitivity of 66.7–100%. The prognostic model with HR 1.72–2.08 by Zhao et al. (2020) improves survival stratification. According to Jiang et al. (2022), DLIR-H outperformed ASIR-V (73.3%) and FBP (62.5%) with a nodule detection rate of 75.8%. With less than 5% of csPCa missed, Zhang et al. (2023) (PCAIDS for prostate cancer) decreased needless biopsies by as much as 32.2%.

The study by Sharma and Parvathi (2025) yielded an F1-score of 96.17%, a precision score of 95.91%, and a recall score of 96.44%. An F1-score of 97.5%, a precision score of 94.8%, and a recall score of 94.3% were obtained from the Akintola et al. (2025) study. Lastly, Liu et al. (2022) generated an F1-score of 90.9% for the detection of oral dysplasia.

To assess AI's practical diagnostic utility, a number of studies directly contrasted its performance with that of skilled medical professionals, including radiologists, pathologists, and oncologists. Li et al. (2024) produced an AI performance that outperformed three radiologists using YOLOv5 + GOLO-CMSS for breast ultrasound, with AP reaching 99.41% and DSC reaching 93.19%. In a similar vein, Li et al. (2023) employed LPAIDS for laryngopharyngeal cancer, with AI reaching an AUC of up to 0.987, which is higher or equivalent to the accuracy of skilled laryngologists. When it came to identifying tumour invasion and grade, the AI model used in Pan et al.'s (2023) study on bladder cancer pathology

performed no worse than skilled pathologists. According to van Dooijeweert et al. (2024), AI-assisted diagnostics demonstrated improved efficiency and quality in a real-world pathology workflow, with a 30% increase in sensitivity and a 30% reduction in IHC usage.

The results of this review also demonstrated that, according to the study by Ruff et al. (2024) (MRI DLR), the Multidisciplinary Tumour Board preferred DLR-enhanced scans (84–100%) over standard MRI for image quality. However, a study by Jiang et al. (2022) demonstrated that AI-enhanced CT reconstructions outperformed radiologist-reviewed conventional techniques in terms of nodule detection.

The CNN model for melanoma, however, had poor specificity (8.2%) and 31.1% misclassification, despite having extremely high sensitivity, which raised concerns about overdiagnosis (Ternov et al., 2022). In a similar vein, CNN-based mutation prediction models were sensitive to mutation prevalence and had a PPV of up to 90.6%, but they lacked external validation (Morel et al., 2023).

AI-powered oncology diagnostic tools showed high diagnostic accuracy across the 27 studies; in certain tasks, many models were able to outperform or match expert clinicians. Accuracy rates often surpassed 90% (Akintola et al., 2025; Sharma & Parvathi, 2025), and reported AUCs reached as high as 0.987 (Li et al., 2023). While machine learning classifiers like LightGBM, TabNet, and logistic regression were common in the analysis of genomic and clinical data, deep learning models particularly CNN variants dominated imaging tasks. Performance benchmarks are positive, but there are still significant gaps for clinical adoption in external validation, real-world trials, and interpretability. However, taken as a whole, these metrics highlight AI's revolutionary potential to increase oncology diagnostic accuracy and efficiency.

3.4 Clinical Validation and Trials

With multiple studies showcasing remarkable model performance in terms of accuracy, sensitivity, and specificity, the incorporation of artificial intelligence (AI) into oncology diagnostics has advanced quickly. Clinical implementation is still a major bottleneck, though. While many AI models are in the development or preclinical validation stage, only a small number have advanced into real-world clinical validation, according to our review of 27 studies. This underscores a crucial translation gap between algorithm development and bedside adoption.

Notably, eight studies showed real-world testing or clinical validation of some kind. The most thorough evidence to date is found in the randomised controlled trial (RCT) on AI-guided targeted prostate biopsy (AI-cTB) carried out by Deng et al. (2024). This study demonstrated how AI can improve diagnostic accuracy in a high-stakes clinical setting by showing a statistically significant improvement in the detection of clinically significant prostate cancer (69.39% vs. 49.71%, $p < 0.001$). This study is a good candidate for wider clinical implementation because it uses a prospective design and is directly integrated into the clinical workflow.

A non-randomized clinical trial assessing an AI-based metastasis detection tool in standard pathology practice was also carried out by van Dooijeweert et al. (2024). Their findings demonstrate the operational benefits of AI beyond predictive accuracy, including a 30% increase in diagnostic sensitivity and cost savings of about €3,000. This study is a practical step towards adoption and offers a replicable model for other institutions, despite its limitations, which include its single-center design and lack of patient outcome data.

Li et al.'s (2023) real-time validation of the Laryngopharyngeal AI Diagnostic System (LPAIDS) across multiple clinical sites demonstrated diagnostic performance that met or exceeded that of human laryngologists, and Jiang et al.'s (2022) prospective evaluation of Deep Learning Image Reconstruction (DLIR) in lung cancer imaging showed superior nodule detection and image quality. These studies offer important insights into model usability in real-world clinical settings, despite their retrospective design.

We selected studies with high implementation potential based on a number of criteria. For instance, in order to obtain regulatory approval and clinical trust, studies such as Deng et al. (2024), Jiang et al. (2022), and Li et al. (2023) went beyond retrospective datasets and conducted live clinical evaluation. Furthermore, policymakers and healthcare providers are more likely to accept AI systems that have proven to be on par with or better than human clinicians, such as Li et al. (2024) with radiologists and van Dooijeweert et al. (2024) with pathologists. Both retrospective and prospective cohorts were used to test the PCAIDS model (Zhang et al., 2023) in nine hospitals, improving its clinical relevance and generalisability.

The majority of the review's studies are still in the preclinical validation or algorithm development stages, despite these encouraging examples. Without showing how the model would function in real-world, diverse clinical settings, many relied on internal datasets, retrospective designs, or public data repositories. Furthermore, only a small percentage of studies addressed issues that affect real-world performance, such as data shift, patient diversity, model explainability, and integration costs. Additionally, some AI models had low specificity (8.2%), which increased the risk of overdiagnosis and needless interventions, even though they achieved high sensitivity (see Ternov et al., 2022; melanoma triage, 99.7%). This demonstrates the necessity of weighing model trade-offs fairly, particularly in situations involving population-wide screening.

3.5 Risk of Bias Assessment

Significant variation was found among the included studies when the risk of bias was evaluated using QUADAS-2 for diagnostic accuracy studies (Whiting et al., 2011). With 25 out of 26 studies classified as low risk, the index test domain which concentrates on the AI model itself, was consistently the strongest across studies. The majority of models were created using appropriate internal validations, performance metrics (such as AUC, accuracy, and F1-score), and well-described architectures. In order to ensure both model transparency and clinical interpretability, Alabi et al. (2025), for instance, used sophisticated explainability techniques like SHAP and LIME to predict head and neck cancer outcomes using a multi-hospital dataset. Similar to this, Li et al. (2024) used CNN segmentation and YOLOv5 to detect breast cancer, outperforming expert radiologists on several datasets.

However, nine studies found that patient selection was a high-risk domain. Numerous studies made use of highly curated samples or retrospective datasets that did not accurately represent the diversity of clinical practice. For example, Sharma and Parvathi (2025) used an unidentified pre-processed image dataset without disclosing patient demographics or selection criteria, which limited generalisability. This calls into question spectrum bias and whether these models can be applied to a variety of patient populations in clinical settings.

More than one-third of the studies in the reference standard domain showed unclear or high risk, indicating significant variability. The diagnostic ground truth was frequently either ill-defined or lacked blinding between reference evaluations and AI predictions. For instance, Morel et al. (2023) may have compromised diagnostic accuracy by using TCGA mutation annotations without external histopathological confirmation. The validity of the model was weakened by studies like Meshkov et al. (2025) that did not specify whether cfDNA labels were validated by independent diagnostic evaluations.

The flow and timing domain also exhibited frequent methodological shortcomings. Nine studies were at high or unclear risk due to a lack of transparency in patient inclusion, timing between AI assessment and reference standard, or unreported dropouts. Akintola et al. (2025) exemplifies these issues by relying solely on public datasets without clinical testing or reporting of workflow continuity. In contrast, Wallis et al. (2025) demonstrated strong methodological rigor by validating a pediatric cancer diagnostic model on both internal and external clinical datasets collected at the Princess Máxima Center.

4 Discussion

27 peer-reviewed studies on the development, verification, and clinical application of artificial intelligence (AI) tools in oncology diagnostics were assessed in this systematic review. The results show promising performance metrics, a variety of application domains, and notable algorithmic advancements. Nonetheless, there is still a significant disconnect between routine clinical use and technical advancement. AI models for a variety of cancers, such as lung, breast, prostate, cervical, colorectal, bladder, gastric, and paediatric tumours, were created in the reviewed studies. The most popular architecture was Convolutional Neural Networks (CNNs), especially for image-based diagnostics like radiology and histopathology tasks (Ternov et al., 2022; Liu et al., 2022). For structured clinical or genomic data, machine learning models such as logistic regression, XGBoost, and LightGBM were commonly used (Meshkov et al., 2025; Zhang et al., 2024). Additionally, investigated were hybrid and ensemble models, which frequently showed better accuracy and resilience (Alabi et al., 2025; Akintola et al., 2025).

A customised approach to oncology diagnostics is reflected in the proliferation of various AI architectures, where the choice of model is dictated by the requirements for interpretability, diagnostic objective, and data modality. This pattern is consistent with earlier reviews that highlight task-specific model design as a crucial component that makes precision oncology possible (Esteva et al., 2019; Liu et al., 2019).

The majority of studies used stringent validation methods, such as external dataset testing, internal cross-validation, and occasionally benchmarking against medical professionals. Across studies, reported performance metrics were consistently high: accuracy, sensitivity, and F1-scores frequently exceeded 90%, and the area under the receiver operating characteristic curve (AUC) varied between 0.80 and 0.98 (Li et al., 2024; Sharma & Parvathi, 2025; Pan et al., 2023). Notably, a number of models demonstrated the ability to enhance or automate diagnostic decision-making by outperforming or matching human experts in diagnostic accuracy (Li et al., 2023; Zhang et al., 2023; van Dooijeweert et al., 2024). However, questions concerning generalisability and real-world robustness are raised by the fact that many studies lack prospective validation. These results are in line with earlier research that called for more robust external validation in clinical AI studies (Liu et al., 2019).

Just 15% or so of the studies advanced to clinical trials or real-time deployment. For instance, Deng et al. (2024) showed noticeably higher detection rates of clinically significant cancers in a randomised controlled trial that contrasted AI-guided prostate biopsies with traditional techniques. An AI-assisted pathology tool was validated in a real clinical workflow by Van Dooijeweert et al. (2024), which resulted in increased sensitivity and cost savings. In a clinical imaging setting, Jiang et al. (2022) prospectively assessed AI-enhanced image reconstruction for lung nodule detection. The viability and advantages of incorporating AI into clinical pathways are demonstrated by these implementation-stage studies. They are still exceptions, though. Due to obstacles like data silos, ethical issues, a lack of regulatory frameworks, and difficulties integrating workflows, the majority of studies are still limited to the development or preclinical validation stages (Nittas et al., 2025).

Many studies provided indirect indicators of better care, but few reported direct patient outcomes like survival or quality of life. These include improved detection of subtle pathology (Li et al., 2023; Ruff et al., 2024), fewer needless biopsies (Zhang et al., 2023), and quicker diagnostic workflows (van Dooijeweert et al., 2024). If expanded, these enhancements may help achieve important oncology objectives such as earlier diagnosis, more focused treatment choices, and improved resource allocation. However, conclusive findings regarding clinical benefit are limited by the lack of long-term outcome data. Longitudinal studies that evaluate the effects of AI tools on treatment decisions, disease progression, and survival must be given top priority in future research.

Four implementation stages can be used to broadly classify the studies. Model training and internal validation were the only activities in Stage 1 development ($n = 9$) (e.g., Sharma & Parvathi, 2025; Akintola et al., 2025). In the second stage, known as experimental validation ($n = 10$), external datasets or functional validation without clinical workflow testing were used (e.g., Li et al., 2025). The third stage, clinical readiness ($n = 5$), involved testing real-world data that had not yet been deployed (e.g., Wallis et al., 2025; Zhang et al., 2023). The deployment of trials or live environments was part of the final stage, which also includes clinical implementation ($n = 3$) (Deng et al., 2024; van Dooijeweert et al., 2024; Jiang et al., 2022). This distribution emphasises the need for translational frameworks that facilitate the transfer from research to practice by highlighting a persistent gap between innovation and adoption.

The following are the main drawbacks of the existing body of work. Underreporting of implementation frameworks, regulatory compliance, and integration into electronic health records (EHRs), a lack of prospective and multicenter validation both crucial for evaluating generalizability and a lack of emphasis on interpretability and clinician trust, particularly in deep learning models. To overcome these obstacles, multicenter randomised trials, explainable AI for clinical transparency, regulatory guidelines (e.g., FDA, CE), and ethical considerations (e.g., bias, equity, and data privacy) should all be incorporated into future AI research in oncology (Char et al., 2018).

5 Conclusion

The evidence from 27 peer-reviewed studies on the creation, verification, and application of AI-powered oncology diagnostics was compiled in this systematic review. The results show that deep learning models like convolutional neural networks (CNNs) and artificial intelligence (AI) have great potential to improve cancer diagnostics in a number

of fields, including imaging, histopathology, and genomics. AI systems have shown excellent diagnostic accuracy, frequently matching or outperforming expert clinician performance. In some instances, classification accuracies have surpassed 95%, and AUC values have reached 0.987. The examined studies used a variety of datasets and validation techniques and cover a broad spectrum of cancer types, including bladder, prostate, breast, and lung cancers. External validation and practical application are still scarce, despite the fact that algorithmic development and internal validation were strong in the majority of studies. There is a persistent translational gap between technical innovation and clinical integration, as evidenced by the small percentage of studies that progressed to clinical trials or live clinical environments. The lack of model interpretability, regulatory uncertainty, difficulties integrating AI into current clinical workflows, and the scarcity of prospective and multicenter trials are major obstacles to broad adoption. Furthermore, many AI models are less generalisable due to biases in data selection and underreporting of patient demographics.

Recommendation for Future Studies and Clinical Practice

In order to assess clinical utility and generalisability, future research should give priority to real-world testing through prospective, multicenter randomised controlled trials. Explainable AI should be developed further in research to increase clinician trust and promote adoption, particularly for deep learning models. Studies evaluating long-term patient outcomes related to AI-assisted diagnostics, such as survival, treatment choices, and quality of life, are required. Researchers must address potential biases in model training and validation by incorporating a wider range of patient populations and clinical settings to guarantee equity and robustness. To create clinically feasible and morally sound AI solutions, future research should incorporate viewpoints from oncologists, AI developers, ethicists, and regulatory specialists.

Institutions should use structured frameworks for implementing AI in clinical practice, which should include workflow integration, clinical validation, user training, and monitoring. To enable safe and efficient implementation in clinical settings, AI tools should be created in compliance with new regulatory standards (such as those set forth by the FDA and CE). In order to ensure a collaborative approach that capitalises on both human expertise and machine precision, AI should be positioned as a decision-support tool rather than a substitute for clinicians. Lastly, in order to iteratively improve AI tools after implementation, clinical deployments should incorporate mechanisms for ongoing performance monitoring and feedback.

Conflict of Interest

The authors declare that there is no conflict of interest.

Funding

The authors declare that no financial support was received for the research, authorship, and publication of this article.

References

- Akintola, A. G., Obiwusi, K. Y., Olatunde, Y. O., Usman, M., Aberuagba, F. G., Adebisi, M. A., & Adebayo, S. A. (2025). Integrated Deep Learning Paradigm for Comprehensive Lung Cancer Segmentation and Classification Using Mask R-CNN and CNN Models. *Franklin Open*, 100278.
- Alabi, R. O., Mäkitie, A. A., Elmusrati, M., Almangush, A., Ehrsson, Y. T., & Laurell, G. (2025). Machine learning explainability for survival outcome in head and neck squamous cell carcinoma. *International Journal of Medical Informatics*, 199, 105873.
- Bhinder, B., Gilvary, C., Madhukar, N. S., & Elemento, O. (2021). Artificial intelligence in cancer research and precision medicine. *Cancer discovery*, 11(4), 900-915.
- Biot, J. (2019). How will clinical practice be impacted by artificial intelligence?. *European Journal of Dermatology*, 29(Supp), 8-10.

- Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3), 229-263.
- Cabral, B. P., Braga, L. A. M., Conte Filho, C. G., Penteado, B., Freire de Castro Silva, S. L., Castro, L., ... & Mota, F. (2025). Future use of ai in diagnostic medicine: 2-wave cross-sectional survey study. *Journal of Medical Internet Research*, 27, e53892.
- Carles, M., Kuhn, D., Fechter, T., Baltas, D., Mix, M., Nestle, U., ... & Gkika, E. (2024). Development and evaluation of two open-source nnU-Net models for automatic segmentation of lung tumors on PET and CT images with and without respiratory motion compensation. *European radiology*, 34(10), 6701-6711.
- Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *New England Journal of Medicine*, 378(11), 981-983.
- Chiu, H. Y., Chao, H. S., & Chen, Y. M. (2022). Application of artificial intelligence in lung cancer. *Cancers*, 14(6), 1370.
- Chua, I. S., Gaziel-Yablowitz, M., Korach, Z. T., Kehl, K. L., Levitan, N. A., Arriaga, Y. E., ... & Hassett, M. (2021). Artificial intelligence in oncology: Path to implementation. *Cancer Medicine*, 10(12), 4138-4149.
- Deng, R., Liu, Y., Wang, K., Ruan, M., Li, D., Wu, J., ... & Gong, K. (2024). Comparison of MRI artificial intelligence-guided cognitive fusion-targeted biopsy versus routine cognitive fusion-targeted prostate biopsy in prostate cancer diagnosis: a randomized controlled trial. *BMC medicine*, 22(1), 530.
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1), 24-29.
- Gao, S., Xu, Z., Kang, W., Lv, X., Chu, N., Xu, S., & Hou, D. (2024). Artificial intelligence-driven computer aided diagnosis system provides similar diagnosis value compared with doctors' evaluation in lung cancer screening. *BMC Medical Imaging*, 24(1), 141.
- Gao, Y., Wen, P., Liu, Y., Sun, Y., Qian, H., Zhang, X., ... & Fu, H. (2025). Application of artificial intelligence in the diagnosis of malignant digestive tract tumors: focusing on opportunities and challenges in endoscopy and pathology. *Journal of Translational Medicine*, 23(1), 412.
- Hill, H. A., Jain, P., Ok, C. Y., Sasaki, K., Chen, H., Wang, M. L., & Chen, K. (2023). Integrative prognostic machine learning models in mantle cell lymphoma. *Cancer research communications*, 3(8), 1435-1446.
- Hunter, B., Hindocha, S., & Lee, R. W. (2022). The role of artificial intelligence in early cancer diagnosis. *Cancers*, 14(6), 1524.
- Li, J., Wang, Z., & Wang, T. (2024). Machine-learning prediction of a novel diagnostic model using mitochondria-related genes for patients with bladder cancer. *Scientific Reports*, 14(1), 9282.
- Li, N., Lin, J., Wang, C., Ran, X., & Zhao, Z. (2025). Artificial intelligence-driven integration of single-cell RNA sequencing and transcriptome analysis to decipher APOE's role in gastric cancer prognosis and therapy. *Discover Oncology*, 16(1), 1-19.
- Li, W., Ye, X., Chen, X., Jiang, X., & Yang, Y. (2024). A deep learning-based method for the detection and segmentation of breast masses in ultrasound images. *Physics in Medicine & Biology*, 69(15), 155027.
- Liu, J. A., Yang, I. Y., & Tsai, E. B. (2022). Artificial intelligence (AI) for lung nodules, from the AJR special series on AI applications. *American Journal of Roentgenology*, 219(5), 703-712.
- Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., ... & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6), e271-e297.
- Liu, Y., Bilodeau, E., Pollack, B., & Batmanghelich, K. (2022). Automated detection of premalignant oral lesions on whole slide images using convolutional neural networks. *Oral Oncology*, 134, 106109.
- Lu, M. T., Raghu, V. K., Mayrhofer, T., Aerts, H. J., & Hoffmann, U. (2020). Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. *Annals of Internal Medicine*, 173(9), 704-713.

- Luchini, C., Pea, A., & Scarpa, A. (2022). Artificial intelligence in oncology: current applications and future perspectives. *British Journal of Cancer*, 126(1), 4-9.
- Maiorano, M. F. P., Cormio, G., Loizzi, V., & Maiorano, B. A. (2025). Artificial Intelligence in Ovarian Cancer: A Systematic Review and Meta-Analysis of Predictive AI Models in Genomics, Radiomics, and Immunotherapy. *AI*, 6(4), 84.
- Meshkov, I. O., Koturigin, A. P., Ershov, P. V., Safonova, L. A., Remizova, J. A., Maksyutina, V. V., ... & Skvortsova, V. I. (2025). Diagnostics of lung cancer by fragmented blood circulating cell-free DNA based on machine learning methods. *Frontiers in Medicine*, 12, 1435428.
- Morel, L. O., Derangère, V., Arnould, L., Ladoire, S., & Vinçon, N. (2023). Preliminary evaluation of deep learning for first-line diagnostic prediction of tumor mutational status. *Scientific Reports*, 13(1), 6927.
- Nittas, V., Ormond, K. E., Vayena, E., & Blasimme, A. (2025). Realizing the promise of machine learning in precision oncology: expert perspectives on opportunities and challenges. *BMC cancer*, 25(1), 276.
- Oala, L., Murchison, A. G., Balachandran, P., Choudhary, S., Fehr, J., Leite, A. W., ... & Wiegand, T. (2021). Machine learning for health: algorithm auditing & quality control. *Journal of medical systems*, 45, 1-8.
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.
- Pan, J., Hong, G., Zeng, H., Liao, C., Li, H., Yao, Y., ... & Lin, T. (2023). An artificial intelligence model for the pathological diagnosis of invasion depth and histologic grade in bladder cancer. *Journal of Translational Medicine*, 21(1), 42.
- Peeters, D., Venkadesh, K. V., Dinnessen, R., Saghir, Z., Scholten, E. T., Vliegenthart, R., ... & Jacobs, C. (2025). Towards safe and reliable deep learning for lung nodule malignancy estimation using out-of-distribution detection. *Computers in Biology and Medicine*, 186, 109633.
- Rajpurkar, P., & Lungren, M. P. (2023). The current and future state of AI interpretation of medical images. *New England Journal of Medicine*, 388(21), 1981-1990.
- Rowe, S. P., & Pomper, M. G. (2022). Molecular imaging in oncology: Current impact and future directions. *CA: a cancer journal for clinicians*, 72(4), 333-352.
- Ruff, C., Bombach, P., Roder, C., Weinbrenner, E., Artzner, C., Zerweck, L., ... & Gohla, G. (2024). Multidisciplinary quantitative and qualitative assessment of IDH-mutant gliomas with full diagnostic deep learning image reconstruction. *European Journal of Radiology Open*, 13, 100617.
- Sartor, H., Minarik, D., Enqvist, O., Ulén, J., Wittrup, A., Bjurberg, M., & Trägårdh, E. (2020). Auto-segmentations by convolutional neural network in cervical and anorectal cancer with clinical structure sets as the ground truth. *Clinical and Translational Radiation Oncology*, 25, 37-45.
- Senthil Kumar, K., Miskovic, V., Blasiak, A., Sundar, R., Pedrocchi, A. L. G., Pearson, A. T., ... & Ho, D. (2023). Artificial intelligence in clinical oncology: from data to digital pathology and treatment. *American Society of Clinical Oncology Educational Book*, 43, e390084.
- Sung, H., Filho, A. M., Laversanne, M., Ferlay, J., Siegel, R. L., Soerjomataram, I., ... & Bray, F. (2026). Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA: A Cancer Journal for Clinicians*, 76(4), e70090.
- Tafazzoli, A., Ramsey, S. D., Shaul, A., Chavan, A., Ye, W., Kansal, A. R., ... & Fendrick, A. M. (2022). The potential value-based price of a multi-cancer early detection genomic blood test to complement current single cancer screening in the USA. *Pharmacoeconomics*, 40(11), 1107-1117.
- Ternov, N. K., Christensen, A. N., Kampen, P. J., Als, G., Vestergaard, T., Konge, L., ... & Hannemose, M. R. (2022). Generalizability and usefulness of artificial intelligence for skin cancer diagnostics: an algorithm validation study. *JEADV Clinical Practice*, 1(4), 344-354.
- van Doijeweert, C., Flach, R. N., Ter Hoeve, N. D., Vreuls, C. P. H., Goldschmeding, R., Freund, J. E., ... & van Diest, P. J. (2024). Clinical implementation of artificial-intelligence-assisted detection of breast cancer metastases in sentinel lymph nodes: the CONFIDENT-B single-center, non-randomized clinical trial. *Nature Cancer*, 5(8), 1195-1205.

- Wallis, F. S., Baker-Hernandez, J. L., van Tuil, M., van Hamersveld, C., Koudijs, M. J., Verwiël, E. T., ... & Kester, L. A. (2025). M&M: an RNA-seq based pan-cancer classifier for paediatric tumours. *EBioMedicine*, 111.
- Walter, N., Groth, J., & von und zu Zwerger, B. (2025). Evaluation of an innovative multi-cancer early detection test: high sensitivity and specificity in differentiating cancer, inflammatory conditions, and healthy individuals. *Frontiers in Oncology*, 15, 1520869.
- Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... & QUADAS-2 Group*. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, 155(8), 529-536.
- Yang, D., Fineberg, H. V., & Cosby, K. (2021). Diagnostic excellence. *Jama*, 326(19), 1905-1906.
- Zhang, H., Ji, J., Liu, Z., Lu, H., Qian, C., Wei, C., ... & Wang, F. (2023). Artificial intelligence for the diagnosis of clinically significant prostate cancer based on multimodal data: a multicenter study. *BMC medicine*, 21(1), 270.
- Zhang, W. H., Tan, Y., Huang, Z., Tan, Q. X., Zhang, Y. M., & Wei, C. Y. (2024). Development and validation of an artificial intelligence model for predicting de novo distant bone metastasis in breast cancer: a dual-center study. *BMC Women's Health*, 24(1), 442.
- Zhao, K., Li, Z., Yao, S., Wang, Y., Wu, X., Xu, Z., ... & Liu, Z. (2020). Artificial intelligence quantified tumour-stroma ratio is an independent predictor for overall survival in resectable colorectal cancer. *EBioMedicine*, 61.
- Zobair, K. M., Houghton, L., Tjondronegoro, D., Sanzogni, L., Islam, M. Z., Sarker, T., & Islam, M. J. (2023). Systematic review of Internet of medical things for cardiovascular disease prevention among Australian first nations. *Heliyon*, 9(11).