PLOS ONE

RESEARCH ARTICLE

The use of artificial intelligence tools in cancer detection compared to the traditional diagnostic imaging methods: An overview of the systematic reviews

Helbert Eustáquio Cardoso da Silva*☯, Glaucia Nize Martins Santos☯, André Ferreira Leite‡, Carla Ruffeil Moreira Mesquita‡, Paulo Tadeu de Souza Figueiredo‡, Cristine Miron Stefani☯, Nilce Santos de Melo☯

Faculty of Health Science, Dentistry of Department, Brasilia University, Brasilia, Brazil

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work
* helbertcardososilva@gmail.com

OPEN ACCESS

Citation: Silva HECd, Santos GNM, Leite AF, Mesquita CRM, Figueiredo PTdS, Stefani CM, et al. (2023) The use of artificial intelligence tools in cancer detection compared to the traditional diagnostic imaging methods: An overview of the systematic reviews. PLoS ONE 18(10): e0292063. https://doi.org/10.1371/journal.pone.0292063

Editor: Yuchen Qiu, University of Oklahoma, UNITED STATES

Received: November 29, 2022

Accepted: September 12, 2023

Published: October 5, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0292063

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Abstract

Background and purpose

In comparison to conventional medical imaging diagnostic modalities, the aim of this overview article is to analyze the accuracy of the application of Artificial Intelligence (AI) techniques in the identification and diagnosis of malignant tumors in adult patients.

Data sources

The acronym PIRDs was used and a comprehensive literature search was conducted on PubMed, Cochrane, Scopus, Web of Science, LILACS, Embase, Scielo, EBSCOhost, and grey literature through Proquest, Google Scholar, and JSTOR for systematic reviews of AI as a diagnostic model and/or detection tool for any cancer type in adult patients, compared to the traditional diagnostic radiographic imaging model. There were no limits on publishing status, publication time, or language. For study selection and risk of bias evaluation, pairs of reviewers worked separately.

Results

In total, 382 records were retrieved in the databases, 364 after removing duplicates, 32 satisfied the full-text reading criterion, and 09 papers were considered for qualitative synthesis. Although there was heterogeneity in terms of methodological aspects, patient differences, and techniques used, the studies found that several AI approaches are promising in terms of specificity, sensitivity, and diagnostic accuracy in the detection and diagnosis of malignant tumors. When compared to other machine learning algorithms, the Super Vector Machine method performed better in cancer detection and diagnosis. Computer-assisted detection (CAD) has shown promising in terms of aiding cancer detection, when compared to the traditional method of diagnosis.

Funding: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Conclusions

The detection and diagnosis of malignant tumors with the help of AI seems to be feasible and accurate with the use of different technologies, such as CAD systems, deep and machine learning algorithms and radiomic analysis when compared with the traditional model, although these technologies are not capable of to replace the professional radiologist in the analysis of medical images. Although there are limitations regarding the generalization for all types of cancer, these AI tools might aid professionals, serving as an auxiliary and teaching tool, especially for less trained professionals. Therefore, further longitudinal studies with a longer follow-up duration are required for a better understanding of the clinical application of these artificial intelligence systems.

Trial registration

Systematic review registration. Prospero registration number: CRD42022307403.

Introduction

Since early diagnosis of cancer is associated with better treatment outcomes for the patient, there is substantial interest in using artificial intelligence (AI) technology in cancer screening and detection through image recognition, in the hope of reducing diagnosis times and increasing diagnostic accuracy [1]. AI has made significant advances in fields including medicine, biomedicine, and cancer research. To forecast cancer behavior and prognosis, AI employs mathematical approaches that aid in decision-making or action based on logical and autonomous thinking and effective adaptability [2–4].

AI has the potential to dramatically affect nearly all aspects of oncology–from enhancing diagnosis to personalizing treatment and discovering novel anticancer drugs. Thus, it is important to review the recent enormous progress in the application of AI and its potential in daily clinical practice, and also to highlight limitations and pitfalls for such purpose [1,2]. Several studies have attested to the potential of AI-based techniques to predict diagnosis, prognosis and response to treatment in some malignant tumors, including colorectal, breast, skin, and lung cancer [5–8].

Machine learning (ML), a branch of AI, has been shown to minimize intercurrences in dysplasia and cancer categorization, assuring uniformity and validity, and influencing treatment decisions [9]. Progress in Deep Learning (DL) approaches has shown gains in image-based diagnosis and illness detection in the study of cancer and oncology [10,11]. DL configurations are non-linear layered artificial neural networks that are hierarchically coupled. A range of DL architectures based on input data types have been developed during the last few years. Simultaneously, the model's performance was evaluated, and it was discovered that the use of DL in cancer prediction is superior than the standard procedures employed in ML [12].

In this context, these systems offer a lot of potential to support and enhance diagnostic methods, such as overcoming the limitations of human memory and attention, improving the effectiveness of computations and interpreting data, and preventing biases and prejudices from influencing judgments. However, radiologists find it challenging to assimilate and evaluate a significant volume of data to perform diagnosis and therapy because of the enormous volume and complexity of the picture data. The diagnosis takes longer, there is a higher risk of mistakes, and radiologists are prone to become fatigued. Automation in the field of radiological imaging can help to solve a number of issues, including a) improving the accuracy and

precision of picture analysis [13]; b) reducing interobserver variability [14]; and c) increasing the speed of image analysis and reports [15,16]. Thus, medical analysis demands the evolution of automated decision-making systems, with the aid of the use of computational intelligence for fast, accurate and efficient diagnosis [17], prognosis and treatment of diseases, such as brain tumors [18].

AI models, such as artificial neural networks (ANNs), have been popular in diagnostic and predictive decision-making procedures when clinical situations are complicated, such as liver cancer [19], malignant melanoma and breast cancer [20,21], and colon cancer [22]. Image processing, pattern recognition, artificial intelligence, and medical pictures are all combined in Computer-Aided Diagnosis (CADs) systems. Several computer-based solutions, such as Computer Aided Diagnosis (CADx) or Computer-Aided Detection (CADe), have been suggested to aid the radiologist in the process of interpreting computed tomography (CT) scans. CADe systems may detect and label suspicious regions as lesions in an image, while CADx systems not only highlight suspicious areas, but also point out the nature of the detected lesion as malignant or benign [23,24]. Therefore, CAD systems might potentially decrease the workload of radiologists, leading to fast and accurate diagnoses.

The terms computer-aided detection (CADe) and computer-aided diagnosis (CADx) are frequently used to describe CAD in the literature. By calling radiologists' attention to questionable areas in an image, CADe schemes aim to eliminate observational oversight. On the other hand, CADx strategies aim to classify a worrisome area and characterize it. CAD schemes and ML-based prediction models for medical images, such as breast imaging, for example, have limited therapeutic relevance despite significant research efforts and the availability of marketed CAD solutions [25]. Radiomics, on the other hand, is a discipline that has emerged as a result of the recent quick breakthroughs in bioinformatics and the introduction of high-performance computers. Radiomics includes calculating numerical image-based features that can be mined and applied to forecast clinical outcomes [26]. To measure and define the size, shape, density, heterogeneity, and texture of the targeted tumors in medical imaging, radiomic techniques are utilized to extract a large number of features from a series of medical images [27]. Segmenting the tumor region and extracting features from there is one way to guarantee that the derived features have some clinical value. As a result, manual or partially automated tumor segmentation is used in several radiomics-based systems. New methods for creating CAD schemes are also being investigated and described in the literature due to the increasing enthusiasm for deep learning-based artificial intelligence (AI) technology [28]. Numerous research have contrasted CAD schemes employing deep learning techniques and traditional radiomics to examine their benefits and drawbacks [29,30].

Since deep learning models can directly extract characteristics from medical images, DL-based CAD schemes are appealing [31]. However, despite the difficulty in achieving high scientific rigor when creating AI-based deep learning models [32], using AI technology to create CAD schemes has emerged as the research standard. Aside from cancer detection and diagnosis, new AI-based models are being broadened to incorporate extensive clinical applications such short-term cancer risk and prognosis prediction and clinical outcome.

Currently, despite systematic reviews on the subject, there is still no overview in the literature that brings together the knowledge of published systematic reviews regarding the use of artificial intelligence in cancer detection in a single publication.

Considering the current potentialities of the aforementioned AI-driven systems for the oncologic field, the capability of these systems to detect malignant tumors based on different imaging modalities should be investigated. Therefore, this overview article aims to answer the following question: When compared to standard imaging diagnosis, how accurate are artificial intelligence applications for cancer detection in adult patients?

Materials and methods

Protocol registration

The protocol of this study was registered on the International Prospective Register of Systematic Reviews—PROSPERO (www.crd.york.ac.uk/PROSPERO/) under number CRD42022307403. This overview was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses, following the PRISMA checklist (http://www.prisma-statement.org/) and was developed according to the JBI Manual for Evidence Synthesis (https://synthesismanual.jbi.global) and the Cochrane Handbook for Systematic Reviews (www.training.cochrane.org/handbook).

The definition of systematic reviews considered was that established by the Cochrane Collaboration. A study was considered a systematic review when reporting or including:

i. research question;

ii. sources that were searched, with a reproducible search strategy (naming of databases, naming of search platforms/engines, search date and complete search strategy);

iii. inclusion and exclusion criteria;

iv. selection (screening) methods;

v. critically appraises and reports the quality/risk of bias of the included studies;

vi. information about data analysis and synthesis that allows the reproducibility of the results; [33,34]

Search strategy

On January 21th, 2022, a broad search of articles without language or time limits was performed in the following databases: PubMed, Cochrane Central Register of Controlled Studies (Cochrane), SciVerse Scopus (Scopus), Web of Science, Latin American and Caribbean Health Sciences (LILACS), Excerpta Medical Database (Embase), Scientific Electronic Library Online (Scielo), Business Source Complete (EBSCOhost) and grey literature through Proquest, Google Scholar and JSTOR. The following Medical Subject Headings (MeSH) terms "Cancer Early Diagnosis," "Artificial Intelligence," "remote technology," "neoplasm" and synonyms were used to develop the search strategy and acquire the main strategy in PubMed. When words with different spelling appeared, synonyms that were in the MeSH terms were used. This strategy was adapted for the other databases. The search strategy used is in S1 Table. Manual searches of reference lists of relevant articles were also performed.

Immediately after literature search, the references were exported to reference manager online Rayyan QCRI (https://rayyan.qcri.org/welcome) and duplicated references were removed.

Inclusion and exclusion criteria

PIRDs (Participants, Index test, Reference Test, Diagnosis of Interest and **S**tudies) acronym was used to define inclusion and exclusion criteria. As inclusion criteria, diagnostic models and/or detection tool of any type of cancer in adult patients (P) in systematic reviews using AI (I) compared to the traditional model of diagnostic radiographic imaging (R) were evaluated. For the diagnoses of interest (D), the following accuracy metrics for detecting and diagnosing cancer were considered: sensitivity, specificity, Receiver Operating Characteristic (ROC) curve, and Area Under the Curve (AUC).

Exclusion criteria comprised: 1—Studies evaluating diagnosis of areas other than medicine and dentistry (Physiotherapist, Nutritionist, Nursing, Caregivers etc.); 2 –Patients with a confirmed diagnosis of cancer; 3—Systematic Reviews on AI, ML, DL and CNN not evaluating the diagnostic accuracy of the systems; 4—Systematic Reviews with AI use for other diseases diagnosis (Diabetes, Hypertension, etc); 5—Systematic reviews in which AI was not compared to a reference test; 6—Systematic reviews evaluating other technologies for detection or cancer diagnosis (spectrometry, biomarkers, autofluorescence, Multispectral widefield optical imaging, optical instruments, robotic equipment etc.); 7—literature reviews, integrative reviews, narrative reviews, overviews; 8—Editorials/Letters; 9—Conferences, Summaries, abstracts and posters; 10—In vitro studies; 11—Studies of animal models; 12—Book chapters; 13—Pipelines, guidelines and research protocols; 14—Review papers that, despite self-styled systematic reviews, do not fulfill the criteria for the definition of Systematic Reviews; 15—Primary studies of any type.

Data extraction

The studies selection was performed in two phases. On phase 1, two independent reviewers (HECS and GNMS) evaluated titles and abstracts of all records, according to the eligibility criteria. On phase 2, both reviewers (HECS and GNMS) independently read the full texts according to the inclusion and exclusion criteria. In case of disagreements, both reviewers discussed and, if consensus was not reached, a third reviewer (AFL) was consulted to reach a final decision. At phase 2, the articles were excluded if they did not fulfill the key characteristics of systematic reviews according to the following criteria [33,34]:

1. Those carried out by a single reviewer
2. Those who do not propose a specific research question (e.g., using PICOS or another appropriate acronym);
3. Those who do not determine pre-specified eligibility criteria;
4. Those who do not use a pre-specified search strategy;
5. Those who do not apply the search strategy to at least two databases
6. Those that do not provide a clear description of the study selection process (methods used to include and exclude research at each level);
7. Those who do not use any method (qualitative/narrative or quantitative using instruments) to assess the methodological quality of included studies.

Study selection

Data extraction was also performed by two independent reviewers (HECS and GNMS) and crosschecked. Extracted data comprised: Author, year, country; Design of included studies; N of included Studies/ N of select studies; Type of cancer; Index test; Reference test; True positives / N of images; True Negatives /N of images; Sensitivity and Specificity/ odds ratio Mean±SD, *p-value*; Diagnostic accuracy; and main conclusions of each paper. When necessary, request for additional information, via email, was made to the authors of the selected articles. Three authors did not provide consolidated data in the form of quantitative analysis. Despite contact via email and social networks, there were no responses from any of the three authors [35–37].

Assessing the methodological quality of included studies

The Critical Appraisal checklist for Systematic Reviews (Joanna Briggs Institute, 2014) was used to assess the methodological quality of the studies independently by two reviewers

(HECS and GNMS) [38]. It should be noted that critical appraisal/risk of bias tools classically indicated for systematic reviews, such as AMSTAR 2 and ROBIS, were designed for systematic reviews of intervention, while the articles included were systematic reviews of diagnostic accuracy. We opted for performing the methodological assessment, not the risk of bias in the selected studies.

Studies were characterized according to the scoring decisions agreed by reviewers previously. Systematic Reviews were considered of "low" methodological quality when only 1 to 4 tool items received "yes" answers; "moderate" quality with 5 to 8 "yes" answers; and "high" quality with 9 to 11 "yes" answers.

Considered outcomes

The indexes and reference tests were compared concerning to cancer detection and diagnosis (sensitivity, specificity, ROC, AUC). Despite previously planned on the protocol, meta-analysis of the data was unfeasible due to studies' high methodological heterogeneity.

Results

Description of included studies

The electronic search of five databases and grey literature retrieved 382 records. Removal of 18 duplicated studies resulted in 364 records. Titles and abstracts from these studies were read and those not fulfilling the eligibility criteria were excluded. In addition, 40 records retrieved from grey literature were considered. At the end of phase 1, 32 papers remained for full text reading (phase 2). Manual search of reference lists did not provide additional studies. Full text reading resulted in 09 eligible studies for qualitative analysis. S2 Table presents excluded articles and reasons for exclusion. A flowchart of the complete process inclusion is shown in Fig 1.

Included studies were conducted in EUA [28], Netherlands [36], Italy [40], Sweden [35], China [41,42], Indonesia [43], United Kingdom [44] and Denmark [37]. All included studies were published in English. One SR included descriptive studies [39], three RS included diagnostic accuracy studies [40,43,44], four SR included prospective or retrospective studies [35,36,41,42] and one SR included clinical trial studies [37]. The accuracy of AI for detecting cancer in adult patients was evaluated by sensitivity, specificity, ROC, and AUC.

Table 1 summarizes study details regarding participants, index test, reference test, outcomes (true positive, true negative, sensitivity, specificity and diagnostic accuracy) and conclusions.

Methodological quality within studies

None of the studies fulfilled all methodological quality criteria. However, five studies [39–42,44] were considered of "high" methodological quality, three studies [35,37,43] were of "moderate" methodological quality and only one study [36] was considered of "low" methodological quality.

In two studies [36,44], the review question was not considered clearly and explicitly stated. The inclusion criteria was not appropriate for the review question in one study [36], the sources and resources used to search for studies was not adequate in one study [39], the likelihood of publication bias was not assessed in four studies [35–37,39], the recommendations for policies and/or practices supported by the reported data were unclear for a study [37], and the specific directives for new research were inconclusive for three studies [36,37,41]. In all of studies the search strategy and the criteria for appraising studies were appropriate.

More information about the methodological quality assessment of included studies can be find in Table 2 (summarized assessment).

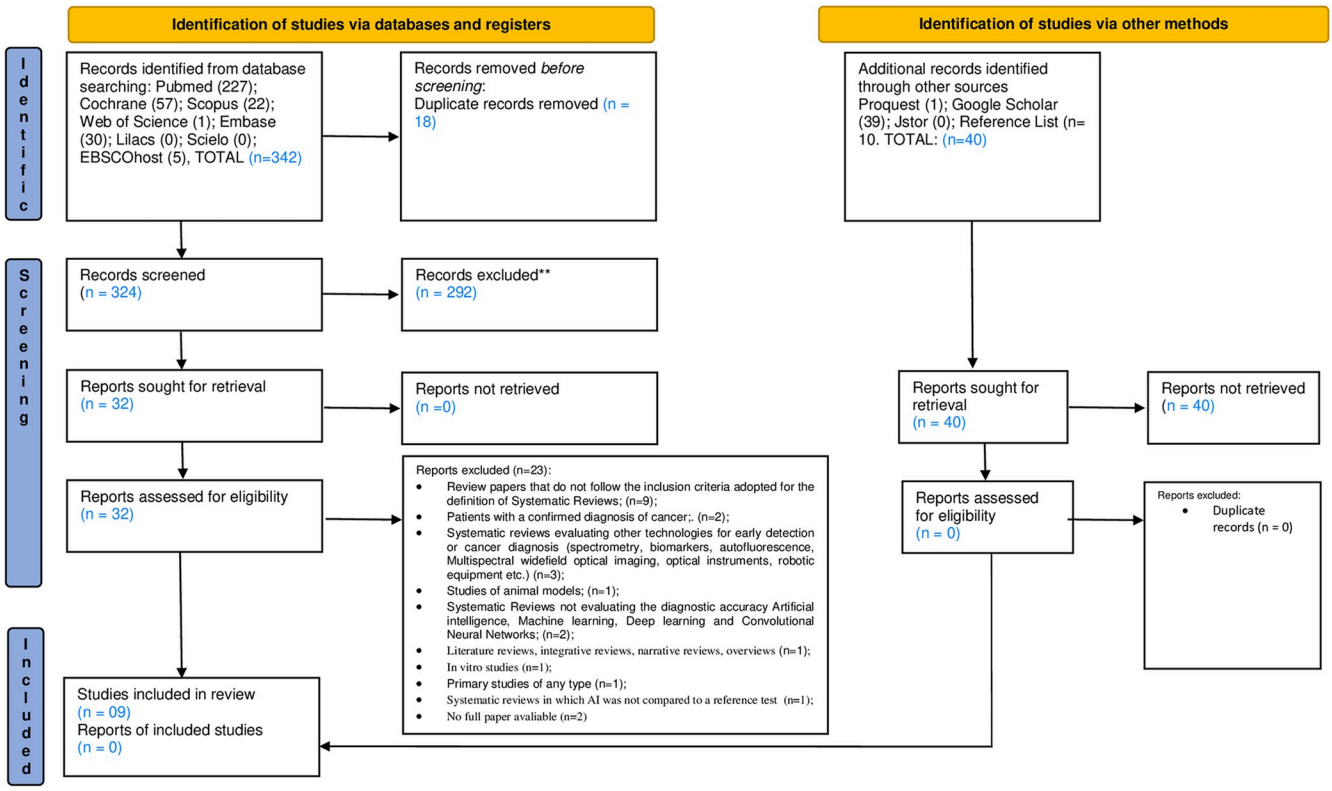

Fig 1. Flow diagram of the literature search and selection criteria.

https://doi.org/10.1371/journal.pone.0292063.g001

Results of individual studies

The systematic review conducted by the Department of Radiology at the University Medical Center Groningen in the Netherlands, looked at computer-assisted detection (CAD) in breast MRI and evaluated radiologists' accuracy in distinguishing benign from malignant breast lesions. Of the 587 papers assessed by the study authors, the 10 studies selected by eligibility criteria included a total of 895 patients with a total of 1264 breast lesions. Sensitivity and specificity were used to compare the performance accuracy of radiologists with and without CAD. Radiologists with experience attained a non-CAD sensitivity of 89% and a CAD sensitivity of 89%, respectively. On the other hand, the specificity was 86% without CAD and 82% specificity with CAD, respectively. Residents' sensitivity rose from 72% to 89% with CAD, while the difference was not statistically significant. In terms of specificity, the findings without CAD 79% and with CAD 78% were identical. The CAD in breast MRI has little bearing on the sensitivity and specificity of competent doctors. [39].

The reviewers from Universitas Gadjah Mada in Indonesia conducted a systematic review to establish the diagnostic accuracy of various ML algorithms for calculating breast cancer risk. There were 1,879 publications assessed in total, with 11 being included in systematic review and meta-analysis. Super Vector Machine (SVM), Artificial Neural Networks (ANN), Decision Tree (DT), Naive Bayes (NB), and K-Nearest Neighbor were identified as five types of ML algorithms used to detect breast cancer risk (KNN). The AUC of the Summary Receiver Operating Characteristic (SROC) for the SVM method was $> 90\%$, demonstrating the greatest performance among the algorithms studied in terms of calculating the risk of breast cancer, and thus having the best precision value compared to other machine learning algorithms [43].

Table 1. Summary of descriptive characteristics of included articles (n = 09).

Author, year, country and design studies	Included Studies	Type of cancer	Index test	Reference test	True positives / N of images	True Negatives /N of images	Sensitivity and Specificity/ odds ratio, Mean ±SD, *p' value*	Diagnostic accuracy (%), Mean±SD, *p' value*	Conclusions
Dorrius et al, 2011 [39], Netherlands, Descriptives studies	10	Breast Cancer	Computer-aided-detection (CADe)	Magnetic Resonance Imaging (MRI)	-	-	Sensitivity Radiologist no CAD, general 82% (95% CI: 72%–90%) Radiologist with CAD, general 89% (95% CI: 83%–93%) Specificity Radiologist no CAD, general 81% (95% CI: 74%–87%) Radiologist with CAD, general 81% (95% CI: 76%–85%)	-	MR images CAD has little influence on the sensitivity and specificity of the performance of radiologists experienced in breast MRI diagnosis. Breast MRI interpretation by radiologists remains essential. Radiologists with less experience seem to benefit from a CAD system when performing breast MRI evaluation.
Henriksen EL et al, (2018) [37], Denmark Clinical trials	13	Breast cancer	CAD system.; Single Reading (SR) SR vs SR + CAD; Double Reading (DR) DR vs SR þ CAD;	MM	-	-	-	-	In conclusion, all but two studies found that SR CAD improves mammography screening RRs, sensitivity, and CDR when compared to SR alone. No statistically significant variations in sensitivity or CDR were seen when compared to DR. More research is needed to assess the impact of CAD in a population-based screening program with high-volume readers. Longer follow-up studies are required for a thorough assessment of cancer rates. And studies based on digital mammography are required to assess the efficacy of CAD in the current standard of care technology.
Nindrea et al, 2018 [43], Indonesia, Diagnostic Accuracy studies	11	Breast cancer	Machine Learning Algorithms Super Vector Machine (SVM); Artificial Neural Networks (ANN); Decision Tree (DT'); Naive Bayes (NB); K-Nearest Neighbor (KNN)	Mammography (MM)	SVM 40,37%/ 3532; ANN 1,30%/ 63325 DT 33,19%/738 NB 35,32%/ 1039 KNN 41%/1568	SVM 46,40%/ 3532 ANN 97,88%/ 63325 DT 61,38%/738 NB 54,66%/ 1039 KNN 44,89%/ 1568	Sensitivity SVM: 0.67–0.99 (95% CI: ([0.41–0.87]-[0.95–1.00]); ANN: 0.84–0.97 (95% CI: ([0.60–0.97]-[0.95–98]); DT: 0.90–0.92 (95% CI: ([0.68–0.99]-[0.88–.95]); NB: 0.76–0.91 (95% CI: ([0.68–0.83]-[0.87–.95]); KNN: 0.56–0.95 (95% CI: ([0.48–0.64]-[0.92–0.97]); Specificity SVM: 0.60–0.98 (95% CI: ([0.36–0.81]-[0.96–1.00]); ANN: 0.71–0.99 (95% CI: ([0.48–0.89]-[0.99–0.99]); DT: 0.79–0.97 (95% CI: ([0.54–0.94]-[0.9–0.98]); NB: 0.78–0.99 (95% CI: ([0.52–0.94]-[0.9–1.00]); KNN: 0.53–0.99 (95% CI: ([0.44–0.61]-[0.93–0.97];	SVM: 99.51%; ANN: 97.3%; DT: 95.13%; NB: 95.99%; KNM: 95.27%;	Therefore, the early diagnosis of breast cancer will be more effective, and the mortality rate of breast cancer will decrease. Additionally, if the present method is designed in the form of a web-based or smartphone application, women who want to know their own risk of breast cancer will be able to access this information easily in daily life.

(*Continued*)

Table 1. (Continued)

Author, year, country and design studies	Included Studies	Type of cancer	Index test	Reference test	True positives / N of images	True Negatives /N of images	Sensitivity and Specificity/ odds ratio, Mean ±SD, *p' value*	Diagnostic accuracy (%), Mean±SD, *p' value*	Conclusions
Azavedo et al, 2012 [35], Sweden, Prospective or Retrospective studies	4	Breast cancer	Computer-aided-detection (CAD)	MM	-	-	-	-	The scientific evidence is insufficient to determine whether CAD + single reading by one breast radiologist would yield results that are at least equivalent to those obtained in standard practice, i.e. double reading where two breast radiologists independently read the mammographic images.
Eadie et al, 2012 [44], United Kingdom, Diagnostic Accuracy studies	48	Breast cancer, lung cancer, liver cancer, prostate cancer, bone cancer, bowel cancer, skin cancer, neck cancer.	CADe; Diagnostic CAD (CADx)	MM; Breast ultrasound (BUS); BUS + mammogram; Lung Conputered Tomography (LCT); Dermatologic;	-	-	Sensitivity (SD) CADe overall Radiologist alone: 80.41±1.46 With CAD: 84.02±1.30 CADx overall Radiologist alone: 2.79±6.12 With CAD: 90.66±4.07 Specificity (SD) CADe overall Radiologist alone: 90.10±1.97 With CAD: 87.08±2.75 CADx overall Radiologist alone: 83.00±14.46) With CAD: 88.04±15.03	Diagnostic odds ratio (DOR) (SD) CADe overallRadiologist alone3.63±0.16With CAD:3.58±0.20CADx overallRadiologist alone3.44±0.79With CAD: 4.75±0.91	Certain types of CAD did offer diagnostic benefit compared with radiologists diagnosing alone: significantly better ln DOR scores were seen with CADx systems used with mammography and breast ultrasound. Applications such as lung CT and dermatologic imaging do not seem to benefit overall from the addition of CAD. These findings therefore offer suggestions about how CAD can be best applied in the diagnosis of cancer using imaging.
Zhao et al, 2019 [42], China, Prospective or Retrospective studies	5	Thyroid (nodules) cancer	CADx system	US	positive likelihood ratio CADx system 4.1 (95% CI 2.5–6.9); CADx by Samsung 4.9 (95% CI 3.4–7.0); radiologists 11.1 (95% CI 5.6–21.9);	negative likelihood ratio CADx sistem 0.17 (95% CI 0.09–0.32); CADx by Samsung 0.22 (95% CI 0.12–0.38); radiologists 0.13 (95% CI 0.08–0.21);	Sensitivity CADx system 0.87 (95% CI: 0.73–0.94; I^2 = 93.53%); CADx by Samsung 0.82 (95% CI: 0.69–0.91; I^2 = 79.62%); radiologists 0.88 (95% CI: 0.80–0.93; I^2 = 81.66%); Specificity CADx system 0.79 (95% CI: 0.63–0.89; I^2 = 89.67%); CADx by Samsung 0.83 (95% CI: 0.76–0.89; I^2 = 27.52%); radiologists 0.92 (95% CI: 0.84–0.96; I^2 = 84.25%);	DOR CADx system25 (95% CI: 15–42; I^2 = 15.5%, p = 0.315);CADx by Samsung23 (95% CI: 11–46; I^2 = 35.9%, p = 0 .197);radiologists86 (95% CI: 47–158; I^2 = 41.1%, p = 0.147)	The sensitivity of the CAD system in thyroid nodules was similar to that of experienced radiologists. However, the CAD system had lower specificity and DOR than the experienced radiologist. The CAD system may play the potential role as a decision-making assistant alongside radiologists in the thyroid nodules' diagnosis.
Cuocolo et al, 2020 [40], Italy, Diagnostic Accuracy studies	12	PCa	Machine learning (ML) ANN; SVM; LDA; NB; Linear regression (LIR); Random forest (RF); Logistic regression (LOR); Convolutional neural network (CNN); Deep transfer learning (DTL);	MRI	-	-	ML in PCa identification–overall (95%CI: 0.81–0.91; I^2 = 92%, p <0.0001); Biopsy group (95%CI: 0.79–0.91; I^2 = 87%, p <0.0001); Radical prostatectomy group (95%CI: 0.76–0.99; I^2 = 93%, p <0.0001); Deep learning (95%CI: 0.69–0.86; I^2 = 86%, p = 0.0001); Non-deep learning (95%CI: 0.85–0.94; I^2 = 89%, p <0.0001);	AUC overall AUC = 0.86Biopsy groupAUC = 0.85; Rradical prostatectomy groupAUC = 0.88;Deep learningAUC = 0.78; Non-deep learningAUC = 0.90;	The findings show promising results for quantitative ML-based identification of csPCa. The results suggest that the overall accuracy of ML approached might be comparable with that reported for traditional Prostate Imaging Reporting and Data System scoring. Nevertheless, these techniques have the potential to improve csPCa detection accuracy and reproducibility in clinical practice.

(*Continued*)

Table 1. (Continued)

Author, year, country and design studies	Included Studies	Type of cancer	Index test	Reference test	True positives / N of images	True Negatives /N of images	Sensitivity and Specificity/ odds ratio, Mean ±SD, *p' value*	Diagnostic accuracy (%), Mean±SD, *p' value*	Conclusions
Tabatabaei et al, 2021 [36], USA Retrospectives studies	18	Glioma	DT; KNN; SVM; RF; LOR; LDA; LIR; Least Absolute Shrinkage and Selection Operator (LAS/SO); Elastic Net (EN); Gradient Descent Algorithm (GDA); Deep Neural Network (DNN)	MRI	-	-	-	-	The results appear promising for grade prediction from MR images using the radiomics techniques. However, there is no agreement about the radiomics pipeline, the number of extracted features, MR sequences, and machine learning technique. Before the clinical implementation of glioma grading by radiomics, more standardized research is needed.
Xing et al, 2021 [41], China, Retrospective studies	15	prostate cancer (PCa); Peripheral zone (PZ); Transitional zone (TZ); Central gland (CG);	CAD system.; ANN; SVM; Linear Discriminant Analysis (LDA); Radiomic Machine Learning (RML); Non—specific classifier (NSC);	MRI	SVM 42,76%/608; ANN 34,55%/301; RML 34,78%/738; NSC 19,41%/1586; PZ 51,95%/256; TZ 59,67%/186; CG 32,39%/71;	SVM 41,94%/608; ANN 37,54%/301; RML 32,60%/738; NPC 65,15%/1586; PZ 32,81%/256; TZ 26,34%/186; CG 46,47%/71;	Sensitivity: 0.47 to 1.00 0.87(95% CI: 0.76–0.94; I^2 = 90.3%, p = 0.00) ANN: 0.66 to 0.77 SVM: 0.87 to 0.92 LDA: NR RML: 0.96 Prostate zones PZ: 0.66 to 1.00 TZ: 0.89 to 1.00 CG: 0.66 Specificity: 0.47 to 0.89 0.76(95% CI: 0.62–0.85; I^2 = *95.8%*, p = 0.00) ANN: 0.64 to 0.92 SVM: 0.47 to 0.95 LDA: NR RML: 0.51 Prostate zones PZ: 0.48 to 0.89; TZ:0.38 to 0.85; CG:0.92	AUC 0.89 (95% CI: 0.86–0.91)	The study indicated that the use of CAD systems to interpret the results of MRI had high sensitivity and specificity in diagnosing PCa. We believe that SVM should be recommended as the best classifier for the CAD system.

Subtitles: CADe = Computer-aided-detection; MRI = Magnetic Resonance Imaging; SVM = Super Vector Machine; ANN = Artificial Neural Networks; DT = Decision Tree; NB = Naive Bayes; KNN = K-Nearest Neighbor; MM = Mammography; CADx = Diagnostic CAD; BUS = Breast ultrasound; DOR = Diagnostic odds ratio; LCT = Lung Conputered Tomography; CDR = CAD on cancer detection rate (CDR); DR = double reading; RR = Recall Rate; Pca = Prostate cancer; PZ = Peripheral zone; TZ = Transitional zone; CG = Central gland; LDA = Linear Discriminant Analysis; RML = Radiomic Machine Learning; NSC = Non—specific classifier; ML = Machine learningA; LIR = Linear regression; RF = Random forest; LOR = Logistic regression; CNN = Convolutional neural network; DTL = Deep transfer learning; LAS/SO = Least Absolute Shrinkage and Selection Operator; EN = Elastic Net; GDA = Gradient Descent Algorithm; DNN = Deep Neural Network; SR = Single Reading; DR = Double Reading.

https://doi.org/10.1371/journal.pone.0292063.t001

The systematic review carried out by researchers from the University College London, United Kingdom, searched the literature for evidence of the effectiveness of a CAD systems in cancer imaging to assess their influence in the detection and diagnosis of cancer lesions by radiologists. A total of 9,199 articles were reviewed, of which 16 papers with radiologists using CAD to detect lesions (CADe) and 32 papers with radiologists using CAD to classify or diagnose lesions (CADx) were included for analysis. CADx was observed to significantly improve diagnosis in mammography, with a diagnostic odds ratio (DOR) value of 4.99 (0.53), with an average increase of 8 and 7% between without and with CADx for sensitivity and specificity, respectively; and for the breast ultrasound DOR was 4.45 (1.40), with a mean increase of 4 and 8% for sensitivity and specificity, respectively. In cases where CADx were applied to

Table 2. Evaluation of methodological quality of included systematic reviews (n = 9).

Study	Methodological quality items assessed											Overall quality[a]
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	
Dorrius (2011) [39]	N	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	High
Nindrea (2018) [43]	N	Y	Y	Y	Y	Y	Y	U	N	Y	N	Moderate
Eadie (2012) [44]	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	High
Zhao (2019) [42]	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	High
Henriksen (2019) [37]	Y	Y	Y	Y	Y	N	Y	Y	N	U	U	Moderate
Azavedo (2012) [35]	N	Y	Y	Y	Y	Y	U	N	N	Y	Y	Moderate
Cuocolo (2020) [40]	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	High
Xing (2021) [41]	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	U	High
Tabatabaei (2021) [36]	N	U	Y	Y	Y	U	U	U	N	Y	U	Low

Note: JBI Critical Appraisal Tool for Systematic Reviews—Q1. Is the review question clearly and explicitly stated? Q2. Were the inclusion criteria appropriate for the review question? Q3. Was the search strategy appropriate? Q4. Were the sources and resources used to search for studies adequate? Q5. Were the criteria for appraising studies appropriate? Q6. Was critical appraisal conducted by two or more reviewers independently? Q7. Were there methods to minimize errors in data extraction? Q8. Were the methods used to combine studies appropriate? Q9. Was the likelihood of publication bias assessed? Q10. Were recommendations for policy and/or practice supported by the reported data? Q11. Were the specific directives for new research appropriate?

[a]Low quality: 1 to 5 "yes" answers; Moderate quality: 6 to 10 "yes" answers; High quality: 11 to 13 "yes" answers

Abbreviations: N, no; U, unclear; Y, yes.

https://doi.org/10.1371/journal.pone.0292063.t002

pulmonary CT, DOR was 2.79 (1.45) and to dermatological images DOR was 3.41 (1.00). It was found diagnostic contradictions with a mean decrease in specificity on pulmonary CT of 7% and on dermatological images of 17%. There was no evidence of benefit from using CADe. The review showed that CADx may offer some benefit to radiologists in specific imaging applications for breast cancer diagnosis although there is no evidence that it can be used in a generalized way, suggesting its application in some types of cancer diagnosis [44].

Based on a study of the current literature, reviewers from Sichuan University in Sichuan, China, conducted a meta-analysis to determine the accuracy of CAD for thyroid nodule diagnosis. A total of 1,206 publications were screened, with 5 of them being chosen for systematic review and meta-analysis in a set of 536 patients and 723 thyroid nodules. The CAD system's sensitivity in diagnosing thyroid nodules was 0.87, which was comparable to expert radiologists' 0.88. However, the CAD system had lower specificity of 0.79 and DOR of 25 when compared to specificity of 0.92 and DOR of 86 of experienced radiologists. The CAD system has potential as an auxiliary tool in decision making, being a possible ally of radiologists in the diagnosis of thyroid nodules [42].

The accuracy and recall rates (RR) of single reading (SR) vs SR + CAD and double reading (DR) vs SR + CAD were examined in a systematic study undertaken by authors from Metropolitan University College in Copenhagen, Denmark. They looked at 1,522 papers of which 1,491 were excluded by abstract. Of the remaining 31 articles, 18 were excluded after full text reading, and therefore 13 matched the review's inclusion criteria. Except for two publications in the SR vs. SR + CAD comparison, adding CAD increased sensitivity and/or cancer detection rate (CDR). There were no significant variations in sensitivity or CDR between the DR group and the SR + CAD group. In all but one research, adding CAD to SR raised RR and lowered specificity. Only one study found a significant difference between the DR and SR+CAD groups. To assess the efficacy of CAD, more research is needed based on coordinated population-based screening programs with extended follow-up times, high-volume readers, and digital mammography [37].

Researchers from Lund University, Skne University Hospital Malmö, Sweden, conducted a systematic review to verify whether readings of mammographic images by a single breast radiologist plus CAD were at least as accurate as readings by two breast radiologists. The authors looked over 1,049 papers of which 996 were excluded. 53 full-text articles were assessed for eligibility and only four met the inclusion criteria, with a population of 271,917 women being investigated. The findings suggested that there was inadequate scientific evidence to establish whether a single mammography reading by a breast radiologist plus CAD is as accurate as the present method of double reading by two breast radiologists. Similarly, the scientific evidence in the literature was insufficient to investigate cost-effectiveness, and the study's quality was deemed low [35].

Authors from the Italian University of Naples "Federico II" conducted a systematic evaluation to assess the diagnostic accuracy of ML systems for diagnosing prostate cancer (csPCa) using magnetic resonance imaging. After the final editing, a total of 3,224 articles were evaluated, of which 3,164 were excluded. Thus, 60 full-text articles were blindly evaluated by each investigator for eligibility, with 12 articles included, with a total of 1979 imaging screenings evaluated. As in the general analysis, statistical heterogeneity was considerable in all subgroups. In the identification of csPCa, the overall AUC for ML was 0.86. The AUC for the biopsy subgroup was 0.85. The AUC for the radical prostatectomy subgroup was 0.88 and Deep learning had an AUC of 0.78. The systematic review presents promising results for the quantitative identification of csPCa based on ML, with the potential to generate improvements in the detection of csPCa in terms of accuracy and reproducibility in clinical practice [40].

The diagnosis accuracy of CAD systems based on magnetic resonance imaging for PCa was investigated in a systematic review conducted by Gansu University of Traditional Chinese Medicine in China. A total of 3107 articles were examined. Of these, 3070 were excluded and of the remaining 37 articles, 15 were included for analysis with a total of 1945 patients. The overall sensitivity of the CAD system varied from 0.47 to 1.00, with specificity ranging from 0.47 to 0.89, according to the meta-analysis. The CAD system's sensitivity was 0.87, specificity was 0.76 and AUC was 0.89. Among the CAD systems, the SVM exhibited the best AUC, with sensitivity ranging from 0.87 to 0.92 and specificity ranging from 0.47 to 0.95. In terms of prostate zones, the CAD system exhibited the highest AUC in the transitional zone, with sensitivity ranging from 0 to 1. The review points out the advantage of using CAD systems for prostate cancer detection due to its high sensitivity and specificity, and the best performance of SVM algorithm for the aforementioned detection purpose [41].

The authors of a systematic review undertaken by the University of Alabama at Birmingham (UAB), Birmingham, AL, USA, analyzed the most current studies in the classification of gliomas by radiomics based on machine learning, evaluating the clinical utility and technical flaws. At the end of the screening phase, a total of 2858 patients were analyzed, from 18 articles that were chosen from 1177 publications, with 1159 papers excluded in the selection process according to the eligibility criteria adopted. The results were promising for predicting the quality of MRI images using radiomics approaches. However, there was no consensus on the radiomics pipeline, considering that the selected articles have employed a wide range of software, large amount of extracted features, different sequences and machine learning techniques. As a result, the authors urge that more standardized research should be done before radiomic glioma categorization is used in clinical practice [36].

Certainty of the evidence in the systematic review's included

Only two articles [35,41] used the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) method to assess the evidence, which examines five factors: risk of

bias, indirectness, inconsistency, imprecision, and publication. Due to the risk of bias and inconsistency, one paper [41] discovered low quality evidence for the following outcomes: true positives (patients with prostate cancer), true negatives (patients without prostate cancer), false negatives (patients incorrectly classified as not having prostate cancer), and false positives (patients incorrectly classified as having prostate cancer).

The second systematic review [35] evaluated only one study regarding the certainty of evidence for the following outcomes: Cancer detection rate and Recall rate, and the quality of the evidence found was very low due to the risk of bias and Indirectness.

Overlapping

Within the RS reviews, included in this overview, a total of 136 primary studies were found. Approximately 3.67% of these main studies were included in multiple SRs. Only five studies were mentioned more than once. S3 Table provides more details on the overlap and features of the primary studies.

Discussion

To the best of the authors' knowledge, this is the first overview article that critically appraise the scientific evidence of AI use for detecting and diagnosing malignant tumors on different imaging modalities. As this is a current and relatively novel topic, nine recent published SRs were retrieved in the literature search. These SRs found high accuracy metric results for the aforementioned diagnostic purpose, demonstrating the potential of AI tools for the oncologic field. The selected studies demonstrated the use of computer-assisted detection (CAD) [35,37,39,41,42,44], machine learning algorithms [40,41,43]and radiomic analysis [36] for detection and diagnosis of malignant tumors based on radiological images.

AI-driven methods for detecting and diagnosing cancer were analyzed by accuracy metrics, such as sensitivity, specificity, AUC, and ROC. The SVM algorithm showed better performance in the detection and diagnosis of prostate cancer and breast cancer when compared to other machine learning algorithms [41,43]. In four studies, CAD systems demonstrated some benefit in helping to detect cancer [39,41,42,44]. Nevertheless, the use of this tool did not present evidence that it can be used in a generalized way, with better indication for some types of cancer, such as breast cancer [44]. In addition, two studies found promising evidence on the use of ML and radiomic analysis in prostate cancer detection and glioma classification, with potential applicability in clinical practice [36,40].

Two questions that were often addressed in the selected articles were which professional can benefit most from the use of AI systems and how these tools should be used. The CAD systems demonstrated high values of sensitivity and sensitivity for diagnosing prostate cancer and this performance may be related to the location of the tumor in the prostate, for example, central gland, peripheral zone and transition zone. It was observed that the sensitivity and specificity in the transition zone was higher than in the peripheral zone and in the central gland [41]. Some papers corroborate the findings that radiologists benefit most from the use of CAD systems in the detection of prostate cancer lesions [45–48].

However, in other study, less experienced radiologists benefited more from the use of artificial intelligence than experienced professionals [39]. Residents or radiologists with little or no experience had greater sensitivity when accompanied by a CAD system for discriminating between breast lesions on MRI. On the other hand, the performance of experienced radiologists showed a non-significant decrease in specificity from 86% (95% CI: 79–91%) without CAD to 82% (95% CI: 76–87%) with CAD. This observation is due to the fact that CAD systems are based only on the dynamics of enhancement, without considering the morphology of

the lesion, which suggests that experienced radiologists may be misled by the enhancement pattern of CAD, resulting in decreased specificity [39]. The literature agrees with the findings that less experienced radiology professionals and residents benefit most from the use of CAD systems in the detection of lesions. [49–52]. Another study demonstrated that when evaluating thyroid nodules for malignancy using ultrasound imaging, a CAD system had similar sensitivity and negative likelihood ratios compared to experienced radiologists [42].

Two studies [35,37] found no significant evidence regarding sensitivity, specificity, and diagnostic accuracy, between single-reading or double-reading mammography compared with single-reading plus CAD or double-reading plus CAD. The use of CADe to detect lesions on images added less value to radiologists than CADx, used to diagnose lesions, with a small increase in weighted mean sensitivity but a decrease in mean specificity However, CADx did not improve diagnosis in combined mammography and breast ultrasound systems. Thus, CADx can be help radiologists that are looking for breast cancer in mammograms or ultrasounds, but it cannot be assumed that its use may be generalized, with applications in other types of cancer [44].

The literature is still controversial regarding the issue of single reading with the presence of CAD and double reading. A previous study found equivalent performance of CAD systems when a single reading was compared to double reading in the detection of cancer lesions [53]. However, for detecting pulmonary nodules, the performance of a CAD system was comparable to a second opinion reading [54]. However, there are works that demonstrate that the single reading of a reader with the help of the CAD as a second reader produces a significantly higher sensitivity than the single reading and the simulated conventional double reading, being a valuable tool for the detection of pulmonary nodules and can be used as a second opinion reading [54]. As there are also works that attest that the independent double reading produces a better detection performance, the presence and probability of CAD mass markers can improve the interpretation of mammography [55,56].

On the other hand, a recent study stated that the quality and amount of the evidence on the use of AI systems in breast cancer screening is still far from what is needed for its incorporation into clinical practice. In screening programs, AI systems are not sufficiently specialized to take the position of radiologist double reading. Larger research do not confirm promising outcomes from smaller ones [57].

Support vector machines (SVM) exhibited the best AUC among the CAD system classifiers for the detection of prostate cancer (CaP) in magnetic resonance imaging, with a range of 0.47 to 1.00 and specificity of 0.47 to 0.89, with an AUC of 0.89 (0.86–0.91). The AUC curve demonstrated stronger sensitivity and specificity in the transition zone than in the peripheral zone and the core gland of the organ, according to the location of the tumor in the prostate. As a result, the sensitivity of different regions of the human body to screening methods may be explained. Other screening methods, with the exception of CAD-assisted MRI, may not detect it due to limited sensitivity [41].

In another study, SVM was compared to four additional classification algorithms: artificial neural network (ANN), decision tree (DT), naive bayes (NB), and K-Nearest Neighbor (KNN). In the breast cancer risk calculation, SVM was shown to generate the best area under the curve (AUC), with AUC $>$ 90%. The SVM has a 97.13% accuracy rate, demonstrating its effectiveness in predicting and detecting breast cancer and having the greatest accuracy and low error rate. In this approach, the SVM algorithm can predict breast cancer risk and outperforms other algorithms in terms of accuracy. Different machine learning algorithms, on the other hand, can aid in the diagnosis of breast cancer. They serve to decrease the risk of errors caused by weariness or inexperienced professionals, and they allow medical data to be analyzed in less time and with more precision [43].

With a combined AUC of 0.86, machine learning paired with radiomics demonstrated excellent results in the characterization of prostate cancer (csPCa). Deep learning analyses, on the other hand, were less accurate than artisanal radiomics and non-deep ML techniques, with AUCs of 0.78 and 0.90, respectively. While deep learning excels with big datasets with hundreds or even millions of examples, this is rarely the case in medical image analytics. In this case, the datasets are often made up of hundreds of patients at most, and the artisan technique outperforms deep learning in this scenario. As deep learning is also computationally more expensive and less understandable, it should be used with caution in medical image analysis and only when it significantly outperforms alternative approaches [40].

The radiomic study of gliomas using radiomic feature extraction in conjunction with various forms of machine learning has yielded encouraging findings with high sensitivity, specificity, accuracy, and AUC. Radiomics systems that used an external dataset had AUCs of 94% and 72%, respectively, indicating a more realistic performance [35]. The ability to translate DL models into real-world applications, in order to improve acceptance and the performance of DL clinically applied by physicians through the generalization of its applications, the interpretability of its algorithms, access to data, and medical ethics, is one of the challenges for the future of AI use in the medical field, particularly oncology, regarding the diagnosis and detection of cancer. The process of application generalization involves building a multimodal model using information other than the evaluated image itself, such as sample size, age, sex, ethnicity, incomplete data collection and a lack of a standard clinical protocol, clinical manifestations, laboratory tests, image data, and epidemiological histories. Due to the complexity of neural networks and the use of these unrepresentative datasets, overfit models that do not generalize to other populations and biased algorithms are produced [58].

The capacity of algorithms to do activities that call for intelligence is referred to as artificial intelligence. Machine learning is a subset of AI, and it refers to algorithms that learn from data in order to perform better. There are two ways that data given into an ML program may be represented: as features or as raw data. Lesion length is an example of a feature, which is a variable in data that may be measured. Digital mammography (DM), ultrasound (US), and magnetic resonance imaging (MRI) scans are examples of raw data in cancer imaging [59].

Learning features poses a challenge for these algorithms even though they often outperform handcrafted features in terms of performance. The subset of ML methods known as DL can be used to overcome this issue. The ability to recognize complicated patterns is the strength of machine learning and deep learning based approaches. Through feature engineering or feature learning, more detailed picture attributes, such as texture, form, border, location, etc., may be acquired. Higher accuracy can be achieved by segmentation based on detailed picture properties. By categorizing picture blocks of a particular size using a sliding window, typical machine learning based algorithms (such as RBFNN, SVM, etc.) get the whole segmentation image. This leads to unnecessary computation, misclassification, and jagged segmentation borders. On the other hand, deep learning-based approaches (such 3D U-Net CNN) outperform conventional machine learning-based methods in terms of performance and segmentation. Deep learning-based methods have greater discriminating abilities in pixel categorization because they can learn more useful picture attributes. However, many machine learning-based approaches require a large amount of labelled training data [59–61].

Features are represented in terms of other, more basic features in DL. Since DL algorithms are made up of many (deep) layers of linked neurons, they are sometimes referred to as deep neural networks (DNNs). CNNs are a specific kind of DNN. CNNs are frequently employed in cancer image analysis since they were created particularly to detect important characteristics in pictures [62,63]. Different criteria are employed for various activities in order to compare the performance of DL networks with human standards. The metrics used in categorization

are founded on receiver operating characteristic analysis. AUC, accuracy, sensitivity, and specificity all have a significant impact in this situation. Thus, accuracy represents the proportion of correctly classified samples, sensitivity represents the likelihood that the model or radiologist will output a positive (and thus malignant) result if the sample is malignant, specificity represents the probability that the model or radiologist will output a negative (and thus benign) result if the sample is benign, and AUC represents the average sensitivity for all possible specificity values [60,61].

Oncologists find it challenging to comprehend how DL models assess data and make judgments since the sheer number of parameters involved make it challenging for professionals to interpret algorithms. Data access and quality are frequently negatively impacted by a deficient data sharing network, as well as competition between different institutions. Building an open data-sharing platform with the participation of numerous institutes is the first step in overcoming these challenges. Governments and businesses must create a formal structure in the future to enable secure data sharing. Examples include privacy-preserving distributed DL (DDL), which offers a way to protect privacy and enables several participants to train jointly using a deep model without explicitly sharing local datasets. Additionally, the Cancer Imaging Archive, which compiles clinical images from many hospitals and institutes, is another excellent illustration of data sharing and can support radiomic studies [58,64,65].

Due to the need to preserve patient information, which can lead to overfitting, it is challenging to get the data in sufficient quantities to have credibility in training and validation in DL. Companies handling this data must adhere to current data protection and privacy laws in both their home countries and the countries of residence of the data subjects. Before exploiting delicate data, such as genetic data, informed agreement from patients must be sought. Patients must be informed about the potential uses of their data, and it must be made sure that everyone would benefit from them. Furthermore, thorough monitoring and validation procedures must be implemented in order to evaluate AI performance across various applications [58,64].

Before DL techniques are used in therapeutic settings, there are significant ethical issues that need to be resolved. The level of supervision needed for doctors must first be decided. Second, the party accountable for DL tools' inaccurate judgments must be identified. Before AI is implemented in real-world settings, it is also necessary to outline legal obligations in the event of a malfunction. In addition, the majority of high-end AI software works in a "black box" testing environment, meaning that users are unaware of the software's fundamental workings. The tester just knows the input/output; the reasoning behind coming to a particular conclusion is still a mystery. Clinicians frequently confront moral conundrums when making predictions without a thorough grasp of the processes underlying them, hence it is imperative to offer greater transparency in AI models by creating techniques that let users examine the details of the input data that affected the result. closer to the truth [58–65].

The main databases used in training ML and DL technologies vary according to the type of cancer. The most used ones are: Breast Cancer dataset (WBCD); Wisconsin Diagnostic Breast Cancer (WDBC); Wisconsin Prognostic Breast Cancer (WPBC) [66]; Digital Database for Screening Mammography (DDSM) [67]; The Mammographic Image Analysis Society (MIAS) [68]; Breast Cancer Digital Repository (BCDR) [69]; The Cancer Imaging Archive (TCIA) Public Access [70] and Lung Image Database Consortium–the LIDC [71]. Breast cancer databases and other databases have been reported up to date for studying cancer, but the information contained in these databases frequently presents some unfavorable issues: a) some are lacking in terms of available features (image-based descriptors, clinical data, etc.); b) others have a limited number of annotated patient cases; c) and/or the database is private and cannot be used as a reference, which makes it difficult to explore and compare performance [69]; the

lack of larger datasets with manual malignancy annotations and diagnostic cancer labels constitutes the main limitation [72].

Other limitations of the databases that can be listed are: the availability of patient-based pathologic diagnoses for only a subset of cases, the inability to perform reader studies because the files do not maintain radiologists identities or a consistent ordering of radiologists marks, the interpretation of CT scans using only transaxial images, the somewhat artificial nature of the lesion categories relative to clinical practice, the interpretation of every case is not performed by the same radiologists, and the design of the manual QA process that focus mostly on the visual identification of objective lesion annotation errors and did not analyzes inconsistencies in the subjectives lesions characteristic ratings, although the benefit of this quality assurance process to the integrity of the Database should not be understated [73].

The critical analysis of meta-analyses that presented complete data

Of the studies selected in this overview, only three studies presented meta-analyses regarding the sensitivity, specificity and diagnostic accuracy of the use of medical radiological images in the detection of cancer lesions, based on artificial intelligence tools [39,42,44].

Critical analysis of the meta-analysis for diagnosing thyroid nodules based on ultrasound imaging through CAD [42] showed that the CAD system had similar sensitivity and negative likelihood ratio compared to experienced radiologists. However, specificity, positive likelihood ratio and DOR were relatively low. These results indicated that there was a clear gap between the CAD system and the radiologist experienced in making the diagnosis of thyroid nodules. Furthermore, successful nodule segmentations were important and influenced the nodule recognition accuracy. Nodule malsegmentation occurred more frequently with benign nodules (n = 11, 18.6%) than with malignant nodules (n = 2, 4.7%) and the difference was statistically significant (P = 0.04). Among nodules with poor segmentation, 54.6% of benign nodules (6/11) were also diagnosed as malignant, while all malignant nodules were diagnosed as malignant. As a result, it is clear that a CAD system's subpar segmentation can raise the false positive rate while having no impact on the false negative rate. The CAD system's sensitivity to thyroid nodules was comparable to that of skilled radiologists. However, compared to an expert radiologist, the CAD system showed worse specificity and DOR. [42].

Meta-analysis for the evaluation of breast lesions with MRI showed that the combined sensitivity and specificity of the experienced radiologist remain comparable with the implementation of CAD. Less experienced residents or radiologists seemed to achieve greater sensitivity with CAD implantation, although not statistically significant. Residents or radiologists with little or no experience obtained greater sensitivity when accompanied by a CAD system for discrimination of breast lesions on MRI. The change in sensitivity after using the CAD was not statistically significant. However, a considerable increase could be observed (72% sensitivity; 95% CI: 62–81% to 89%; 95% CI: 80–94%). This rise could be attributable to the fact that CAD alerts radiologist trainees or less skilled radiologists to more enhanced lesions, which may be helpful when assessing breast lesions with MRI [39].

The performance of experienced radiologists showed a non-significant decrease in specificity from 86% (95% CI: 79–91%) without CAD to 82% (95% CI: 76–87%) with CAD. A clarification for this observation may be that CAD systems are based only on the dynamics of enhancement, without taking into account the morphology of the lesion. As a consequence, the use of CAD could lead to a greater number of enhanced lesions, part of which could be classified as benign based on morphology [39].

In another study using mammograms and breast ultrasound imaging in the evaluation of CAD systems, certain types of CAD offered diagnostic benefits compared to radiologists

diagnosing alone: significantly better ln DOR scores were seen with CADx systems used with mammography and breast ultrasound. This fact can be observed, since the use of CADx tends to increase sensitivity and specificity in mammography (mean increase of 8 and 7% between without and with CADx for sensitivity and specificity, respectively) and breast ultrasound (mean increase of 4 and 8% for sensitivity and specificity, respectively), but adversely affects specificity in lung CT (mean reduction 7%), combined breast ultrasound and mammography systems (mean reduction 12%) and dermatologic imaging (mean reduction 17%). According to evidence, using CADe systems results in a tiny net overall drop in ln DOR as well as a similar-sized gain in sensitivity and loss in specificity [44].

It is also noticed that the use of CADx improved the diagnosis. However, the overlapping of the 95% confidence interval (CI) curves suggests that the difference is not significant. The AUC is 0.88 (SD: 0.03) for radiologists alone and 0.92 (SD: 0.03) for the same radiologists using CADx and 0.85 (SD: 0.19) for radiologists alone in studies of detection and 0.84 (SD: 0.19) for those radiologists using CADe [44].

The examined meta-analyses did, however, have several drawbacks. First, all displayed significant variation among trials in terms of sensitivity and specificity. This variability is probably due to both the fundamental variations in the patients who were included in the studies' methodologies. Second, the included studies' sample sizes were somewhat modest [39,42,44]. When conducting the meta-analyses, the authors took into account the possibility of selection [39,42,44], measurement [42], and publication [42,44] bias.

The role of explainable artificial intelligence in DL and ML models

Recent advances in ML have sparked a new wave of applications for AI that provide significant advantages to a variety of fields. Many of these algorithms, however, are unable to articulate to human users why they made certain decisions and took certain actions. Explanations are necessary for users to comprehend, have faith in, and manage these new artificially intelligent partners in the crucial knowledge domains of defense, medical, finance, and law for exemplo [74–76].

New ML methods including SVMs, random forests, probabilistic graphical models, reinforcement learning (RL), and DL neural networks are significantly responsible for the current strong performance of AI. These models exhibit good performance, but they are difficult to understand. In many cases, the most performing methods (such as decision trees) are the least explainable, and the most explainable methods (such as DL) are the least accurate. Explanations might be complete or incomplete. Full explanations are provided by fully interpretable models in a transparent manner. Partially interpretable models shed light on key aspects of their thought process. Contrary to black box or unconstrained models, interpretable models adhere to "interpretability restrictions" that are established according to the domain [77].

Although there may be many different types of users, frequently at various times in the development and use of the system, the Explainable Artificial Intelligence (XAI) assumes that an explanation is provided to an end user who depends on the decisions, recommendations, or actions produced by an AI system. For instance, an intelligence analyst, a judge, an operator, developers or test operators, or policy makers. Each user group could have a particular explanation style that they find to be the most successful in conveying information [77,78].

The effectiveness of an explanation has been evaluated and measured in a number of ways, but there is presently no accepted method of determining if a XAI system is more user-intelligible than a non-XAI system. Task performance may be a more objective indicator of an explanation's efficacy than other of these indicators, such as user satisfaction. It remains an

outstanding research question how to accurately and consistently measure the impact of explanations [79,80].

Before explainability can be achieved in DL models, there are still several open problems and obstacles at the intersection of ML and explanation. First, there is a lack of consensus over the terminology and many definitions used in relation to XAI. Since XAI is still a relatively new field, there isn't yet a set of accepted terms in use [81].

Second, there is a trade-off between accuracy and interpretability [82], i.e., between the thoroughness of this description and the simplicity of the information provided by the system regarding its internal functioning. This is one of the reasons why developing objective measurements for what makes a good explanation is difficult with XAI.

Utilizing findings from experiments in human psychology, sociology, or cognitive sciences to develop objectively compelling explanations is one way to lessen this subjectivity. This would allow programmers to design software for their target audience rather than for themselves, with the evaluation of these models being more concerned with people than with technology [83,84]. A promising approach to solving this problem is to combine the connectionist and symbolic paradigms [85–89]. Connectionist approaches are more exact but opaque on the one hand. Symbolic approaches, on the other hand, are more easily understood while being generally seen as less effective. Additionally, it has been demonstrated that the introduction of counterfactual explanations might aid the user in comprehending a model's conclusion [90–92].

Third, XAI approaches for DL must address the issue of delivering explanations that are understandable to society, decision-makers, and the legal system as a whole. In order to address ambiguities and establish the social right to the (not yet existing) right to explanation under the General Data Protection Regulation of all countries in general, it will be especially important to communicate explanations that require non-technical competence [93].

It is obvious that incorporating this work into explainable AI is not an easy process. These models will need to be improved and expanded from a social science perspective in order to produce good explanatory agents, necessitating strong collaboration between explainable AI researchers and those in philosophy, psychology, cognitive science, and human-computer interaction [83].

The use of uncertainty quantification approaches in medical imaging

In addition to using uncertainty quantification (UQ) approaches for medical image analysis, XAI is also used in decision-making in DL methods. Tools have been created to quantify the predicted uncertainty of a specific DL model (Abdar et al., 2021a). The implementation of a deep learning algorithm for uncertainty quantification in oncology can aid in improving performance while analyzing medical images. As a result, for exemplo, the outcomes of prostate cancer segmentation from ultrasound pictures are enhanced by the addition of uncertainty quantification [94].

Numerous advantages result from improving the application of the uncertainty quantification metric. In a medical setting, it becomes essential to identify questionable samples that require human evaluation in order to avoid silent errors that could result in incorrect diagnosis or treatments. Second, UQ makes it possible to spot the model's flaws, such as uncertain forecasts, which may point to a deficient training set. Inconsistencies in the incoming data might also be shown by a high level of UQ, which is crucial for quality control (QC). Overall, UQ strengthens user confidence in the algorithm and makes it easier for the algorithm and user to communicate. Additionally, UQ is supported by solid theoretical underpinnings and has developed as a clinically expected characteristic of an applied AI system [95]. In this situation, the model's predicted performance alone is insufficient to achieve a high level of acceptability.

In order to encourage human-machine collaboration and eliminate the black-box effect, UQ is essential.

In this context, the collaboration between researchers in medicine and artificial intelligence is one future study area that might be taken into consideration. As a result, the suggested machine learning and deep learning methods can do a better job of forecasting various diseases and cancers. This can be very beneficial for resolving uncertainties [94].

The collecting of medical data to the greatest extent possible is one of the gaps for enhancing the uncertainty metric in choices. The accuracy of the findings generated from medical picture segmentation depends heavily on the use of ground truth data. The sending of inaccurately projected facts to experts also plays a significant part in coping with uncertainty. Therefore, in the field of medical picture segmentation, there is a need for strong cooperation between researchers in medicine and computer science [94].

Big medical data collection may be a significant future direction. More data can significantly enhance the performance of several deep learning techniques. Transfer learning approaches, however, can be a good alternative if huge datasets are not available for training [94]. The majority of the UQ approaches that have been put into practice (81.15%) are based on a sampling protocol and try to produce several predictions for the same query input. The potential of deterministic UQ approaches that only require one step to compute uncertainty should be thoroughly investigated [96].

And finally, while being critical in real-world medical circumstances, the detection of Out-of-distribution (OOD) predictions using uncertainty is a subject of relatively few investigations. In an automated medical picture pipeline, input samples may show a variety of anomalies and artifacts that could interfere with the NN's performance and lead to severely inaccurate predictions. This inspires the creation of feature-based techniques designed specifically for OOD detection. Noting that OOD detection is a very active research area that is not exclusive to the UQ sector, it should be noted that OOD detection is currently not often used for medical picture analysis [96,97].

Limitations of the included systematic reviews and the overview

Regarding the limitations presented in the systematic reviews included in this overview, it was observed: short follow-up time, which leads to an overestimated sensitivity [35] or a loss in the calculation of diagnostic accuracy measures [37]; relatively low number of studies [40]; high heterogeneity can be partly explained by the diversity of methodological aspects, difference between patients, or diversity of techniques used [35,40,42]; presence of selection bias by choice of articles reporting sensitivity and specificity results [44], by use of retrospective studies, vaguely reported sample of patients [35], by use of studies with relatively small samples [41,42]; possible presence of publication bias due to lack of studies with unfavorable data [44]; use of digitized analog radiographs to the detriment of digital images [35,37]; behavior of radiologists in terms of training, conducting clinical tests and surveillance in the analysis [35,44]; relatively small and old technology dataset number[35,39]; presence of measurement bias due to the large difference between the groups studied and the small number of outcomes observed in the included studies.

This overview presented as limitations: there are still few studies that use artificial intelligence, in its various approaches, in the detection of cancer, being limited to some more favorable types of cancer, such as breast cancer, prostate cancer and thyroid cancer. There is considerable heterogeneity in the methodologies of the studies, which makes it difficult to standardize the artificial intelligence technologies used. Finally, the limitation of the type and quality of images makes it difficult or impossible to use artificial intelligence in the detection of certain types of cancer, such as in the case of skin or lung cancer.

Mains research gaps and future ML/DL research directions

The fact that deep learning algorithms demand a lot of data, sophisticated imaging technology, top-tier statisticians, and research funding to produce is one of their key gaps. First of all, because of the research's existing variances in sample size, research design, data source, and imaging collecting criteria, it is challenging to quantify, integrate, and extrapolate the findings in a way that was applicable to all situations. Additionally, researchs might exhibit a significant degree of publication bias, especially when they lack external validity [98].

Furthermore, Most AI models also disregard social and cultural risk variables, and the majority of those that have been developed were built using data from the entire population. To increase the accuracy of current models' predictions and modify these tools to the unique characteristics of the population being examined, combining critical risk factors, including imaging, pathology, demographics, clinical data, smoking status, tumor histology and new and ancient technology is advised [98,99]. Researchers can create predictive models by combining several features [100,101]. The concept of multi-omics [102,103] or "Medomics" [104] is introduced as a result. Therefore, it will be worthwhile to continue to pursue the merger of various domain expertise and multidisciplinary integration.

The need for additional large-scale multicenter prospective researchs is highlighted by the fact that this type of research necessitates big datasets. Future research should concentrate on creating deep learning models from decentralized, nonparametric data [105,106]. When compared to conventional models, these methods directly process the raw data, which reduces variability while enhancing model performance [98].

However, Large datasets on the order of (tens of) thousands of patients from various medical centers are now available for research using digital mammography (DM) and digital breast tomosynthesis (DBT). Rarely does MRI research involve more than 500 individuals, and it often comes from a single center. This certainly benefits AI performance in DM and DBT research, as larger datasets and data from various sources typically result in DL models that perform better and have better generalization. Although there are currently a number of sizable retrospective and multi-reader studies for the evaluation of DL CAD systems for DM and DBT, there are less of them for ultrasound (US) and none that that are known for MRI. Thus, DL research in US and MRI needs to invest in generating larger and more diverse datasets to move from proof-of-concept models to systems ready for large multi-case studies with multiple readers, as is now the case with DM/ DBT. However, this does not mean that all DM/DBT models are sufficiently tested for implementation in clinical practice [59,106].

In this way, sharing data between medicals center is a simple way to prevent small datasets from becoming obsolete and large ones from expanding quickly. Regulated data exchange is unfortunately a major barrier for researchers [105]. Swarm learning, where all participants contribute to both case collection and algorithm development [107,108], or even federated learning, where the data stays local but the algorithm travels [105,109,110], are positioned to solve this issue. Such methods haven't, however, been widely used up until now. The challenge of validating the precise results of DL research in cancer pictures, which is typically not achievable since the (training) data are not can be shared, is solved by the construction of checklists, showing the basic requirements for the transparent reporting AI clinical investigations. Future studies on AI will be able to be more thorough and consistent thanks to these lists, which is necessary before they are applied broadly [59,106].

And Finally, it is crucial to address the related ethical, medicolegal, and regulatory challenges as more AI technologies are developed that have the potential for clinical translation. There are a lot of unsolved questions on the ethical front. What situations must doctors tell their patients they're using AI techniques in their clinical workup? It might be crucial in

scenarios where AI functions as a "black box," in which clinicians act on the output of an AI tool without knowing how the algorithm came to its conclusion. When an AI technology misses a cancer, who is responsible? How much should be under human control? an the DL CAD systems make final decisions? Who is responsible for bad DL decisions? Will radiologists be biased as a result of AI assistance? What are people's perceptions of DL decision tools? Can DL CAD algorithms correctly describe their thought process? Before DL models can be widely used in actual clinical settings, it is evident that there must be discussion of these algorithmic biases, which also raise ethical issues [59,106].

Conclusion

This overview gathered evidence from systematic reviews that evaluated the use of AI tools in the detection and diagnosis of malignant tumors based on radiographic images. The detection and diagnosis of malignant tumors with the help of AI seems to be feasible and accurate with the use of different technologies, such as CAD systems, machine learning algorithms and radiomic analysis when compared with the traditional model. ML algorithms performed better when compared to DL methods. However, these systems yielded better performance in some specific types of tumors such as cancer breast cancer, prostate cancer and thyroid nodules. Although there are limitations regarding the generalization for all types of cancer, these AI tools might aid professionals, serving as an auxiliary and teaching tool, especially for less trained professionals. Therefore, further standardized and longitudinal studies should be performed by using AI algorithms for detecting malignant lesions on different imaging modalities, by using larger datasets. These future perspectives will enable a better understanding of AI use in clinical oncologic practice.

Supporting information

S1 Checklist. PRISMA 2009 checklist.
(DOCX)

S1 Table. Database search strategy.
(DOCX)

S2 Table. Excluded articles and reasons for exclusion (n = 23). Legend—**1—**Studies evaluating diagnosis of areas other than medicine and dentistry (Physiotherapist, Nutritionist, Nurse, Caregivers etc.); 2 –Patients with a confirmed diagnosis of cancer; 3—Systematic Reviews not evaluating the diagnostic accuracy Artificial intelligence, Machine learning, Deep learning and Convolutional Neural Networks; 4—Systematic Reviews with Artificial intelligence use for other diseases diagnosis (Diabetes, Hypertension, etc); 5—Systematic reviews in which AI was not compared to a reference test; 6—Systematic reviews evaluating other technologies for early detection or cancer diagnosis (spectrometry, biomarkers, autofluorescence, Multispectral widefield optical imaging, optical instruments, robotic equipment etc.); 7—literature reviews, integrative reviews, narrative reviews, overviews; 8—Editorials⁄Letters; 9—Conferences, Summaries, abstracts and posters; 10—In vitro studies; 11—Studies of animal models; 12—Thesis and Dissertations and book chapters; 13—Pipelines, guidelines and research protocols; 14—Review papers that do not follow the inclusion criteria adopted for the definition of Systematic Reviews; 15—Primary studies of any type; 16—No full paper avaliable.
(DOC)

S3 Table. Overlaping (n = 09).
(DOCX)

Author Contributions

Conceptualization: Helbert Eustáquio Cardoso da Silva, Glaucia Nize Martins Santos, Cristine Miron Stefani, Nilce Santos de Melo.

Data curation: Helbert Eustáquio Cardoso da Silva, Glaucia Nize Martins Santos, André Ferreira Leite, Carla Ruffeil Moreira Mesquita, Paulo Tadeu de Souza Figueiredo.

Formal analysis: André Ferreira Leite, Carla Ruffeil Moreira Mesquita, Paulo Tadeu de Souza Figueiredo.

Investigation: André Ferreira Leite, Carla Ruffeil Moreira Mesquita, Paulo Tadeu de Souza Figueiredo.

Methodology: Helbert Eustáquio Cardoso da Silva, Glaucia Nize Martins Santos, Cristine Miron Stefani, Nilce Santos de Melo.

Project administration: Cristine Miron Stefani, Nilce Santos de Melo.

Supervision: Cristine Miron Stefani, Nilce Santos de Melo.

Validation: Cristine Miron Stefani, Nilce Santos de Melo.

Writing – original draft: Helbert Eustáquio Cardoso da Silva, Glaucia Nize Martins Santos, Cristine Miron Stefani, Nilce Santos de Melo.

Writing – review & editing: Helbert Eustáquio Cardoso da Silva, Glaucia Nize Martins Santos, André Ferreira Leite, Carla Ruffeil Moreira Mesquita, Paulo Tadeu de Souza Figueiredo, Cristine Miron Stefani, Nilce Santos de Melo.

References

1. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018; 2(10): 719–731. https://doi.org/10.1038/s41551-018-0305-z PMID: 31015651
2. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017; 318(22): 2199–2210. https://doi.org/10.1001/jama.2017.14585 PMID: 29234806
3. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2014; 31(13): 8–17. https://doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696
4. Chen M, Decary M. "Artificial intelligence in healthcare: An essential guide for health leaders," in Healthcare management forum. 2020; 33(1): 10–18. https://doi.org/10.1177/0840470419873123 PMID: 31550922
5. Kim K, Kim S, Han K, Bae H, Shin J, Lim JS. Diagnostic Performance of Deep Learning-Based Lesion Detection Algorithm in CT for Detecting Hepatic Metastasis from Colorectal Cancer. Korean J Radiol. 2021; 22(6): 912–921. https://doi.org/10.3348/kjr.2020.0447 PMID: 33686820
6. Tran WT, Sadeghi-Naini A, Lu FI, Gandhi S, Meti N, Brackstone M, et al. Computational Radiology in Breast Cancer Screening and Diagnosis Using Artificial Intelligence. Can Assoc Radiol J. 2021; 72(1): 98–108. https://doi.org/10.1177/0846537120949974 PMID: 32865001
7. Das K, Cockerell CJ, Patil A, Pietkiewicz P, Giulini M, Grabbe S, et al. Machine Learning and Its Application in Skin Cancer. Int J Environ Res Public Health. 2021; 18(24): 13409. https://doi.org/10.3390/ijerph182413409 PMID: 34949015
8. Zarzeczny A, Babyn P, Adams SJ, Longo J. Artificial intelligence-based imaging analytics and lung cancer diagnostics: Considerations for health system leaders. Healthc Manage Forum. 2021; 34(3): 169–174. https://doi.org/10.1177/0840470420975062 PMID: 33297774
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553): 436–444. https://doi.org/10.1038/nature14539 PMID: 26017442
10. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. Brief Bioinform. 2021; 22(1): 360–379. https://doi.org/10.1093/bib/bbz171 PMID: 31950132

11. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019; 25(1): 24–29. https://doi.org/10.1038/s41591-018-0316-z PMID: 30617335

12. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings Bioinf. 2018; 19(6): 1236–1246. https://doi.org/10.1093/bib/bbx044 PMID: 28481991

13. Hayward RM, Patronas N, Baker EH, Vézina G, Albert PS, Warren KE. Inter-observer variability in the measurement of diffuse intrinsic pontine gliomas. J Neurooncol. 2008; 90(1): 57–61. https://doi.org/10.1007/s11060-008-9631-4 PMID: 18587536

14. Chlebus G, Meine H, Thoduka S, Abolmaali N, van Ginneken B, Hahn HK, et al. Reducing interobserver variability and interaction time of MR liver volumetry by combining automatic CNN-based liver segmentation and manual corrections. PLoS ONE. 2019; 14(5): e0217228. https://doi.org/10.1371/journal.pone.0217228 PMID: 31107915

15. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med. 2018; 131(2): 129–133. https://doi.org/10.1016/j.amjmed.2017.10.035 PMID: 29126825

16. Matheson R. Faster analysis of medical images. MIT News. 2018. Available from: http://news.mit.edu/2018/faster-analysis-of-medical-images-0618.

17. Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP. Deep Learning in Neuroradiology. AJNR Am J Neuroradiol. 2018; 39(10): 1776–1784. https://doi.org/10.3174/ajnr.A5543 PMID: 29419402

18. Siuly S, Zhang Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis. Data Sci Eng. 2016; 1(2): 54–64. https://doi.org/10.1007/s41019-016-0011-3

19. Cucchetti A, Vivarelli M, Heaton ND, Phillips S, Piscaglia F, Bolondi L, et al. Artificial neural network is superior to MELD in predicting mortality of patients with end-stage liver disease. Gut. 2007; 56(2): 253–258. https://doi.org/10.1136/gut.2005.084434 PMID: 16809421

20. Carrara M, Bono A, Bartoli C, Colombo A, Lualdi M, Moglia D, et al. Multispectral imaging and artificial neural network: mimicking the management decision of the clinician facing pigmented skin lesions. Phys Med Biol. 2007; 52(9): 2599–2613. https://doi.org/10.1088/0031-9155/52/9/018 PMID: 17440255

21. Papadopoulos A, Fotiadis DI, Likas A. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines. Artif Intellig Med. 2005; 34(2): 141–150. https://doi.org/10.1016/j.artmed.2004.10.001 PMID: 15894178

22. Selaru FM, Xu Y, Yin J, Zou T, Liu TC, Mori Y, et al. Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. Gastroenterology 2002; 122(3): 606–613. https://doi.org/10.1053/gast.2002.31904 PMID: 11874992

23. Castellino RA. Computer aided detection (CAD): an overview. Cancer Imaging. 2005; 5(1):17–19. https://doi.org/10.1102/1470-7330.2005.0018 PMID: 16154813

24. Nishikawa RM. Computer-aided Detection and Diagnosis In: Bick U, Diekmann F Editors. Digital Mammography. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. pp. 85–106. https://doi.org/10.1007/978-3-540-78450-0

25. Nishikawa RM, Gur D. CADe for early detection of breast cancer-current status and why we need to continue to explore new approaches. Acad Radiol. 2014; 21(10): 1320–1321. https://doi.org/10.1016/j.acra.2014.05.018 PMID: 25086951

26. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018; 2(1): 1–8. https://doi.org/10.1186/s41747-018-0068-z PMID: 30426318

27. Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017; 14(12): 749–762. https://doi.org/10.1038/nrclinonc.2017.141 PMID: 28975929

28. Chan H-P, Samala RK, Hadjiiski LM. CAD And AI for breast cancer–recent development and challenges. Br J Radiol 2019; 93(1108): 20190580. https://doi.org/10.1259/bjr.20190580 PMID: 31742424.

29. Jones MA, Faiz R, Qiu Y, Zheng B. Improving mammography lesion classification by optimal fusion of handcrafted and deep transfer learning features. Phys Med Biol 2022; 67(5): 054001. https://doi.org/10.1088/1361-6560/ac5297 PMID: 35130517

30. Danala G, Maryada SK, Islam W, Faiz R, Jones M, Qiu Y, et al. Comparison of computer-aided diagnosis schemes optimized using radiomics and deep transfer learning methods. Bioengineering (Basel) 2022; 9(6): 256. https://doi.org/10.3390/bioengineering9060256

31. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 2021; 13(1): 152. https://doi.org/10.1186/s13073-021-00968-x PMID: 34579788

32. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intelligence 2021; 3(3): 199–217. https://doi.org/10.1038/s42256-021-00307-0

33. Krnic Martinic M, Pieper D, Glatt A, Puljak L. Definition of a systematic review used in overviews of systematic reviews, meta-epidemiological studies and textbooks. BMC Med Res Methodol. 2019; 19(1): 203. https://doi.org/10.1186/s12874-019-0855-0 PMID: 31684874

34. Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

35. Azavedo E, Zackrisson S, Mejàre I, Heibert Arnlind M. Is single reading with computer-aided detection (CAD) as good as double reading in mammography screening? A systematic review. BMC Med Imaging. 2012; 12(1): 22. https://doi.org/10.1186/1471-2342-12-22 PMID: 22827803

36. Tabatabaei M, Razaei A, Sarrami AH, Saadatpour Z, Singhal A, Sotoudeh H. Current Status and Quality of Machine Learning-Based Radiomics Studies for Glioma Grading: A Systematic Review. Oncology. 2021; 99(7): 433–443. https://doi.org/10.1159/000515597 PMID: 33849021

37. Henriksen EL, Carlsen JF, Vejborg IM, Nielsen MB, Lauridsen CA. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. Acta Radiol. 2019; 60(1): 13–18. https://doi.org/10.1177/0284185118770917 PMID: 29665706

38. Tufanaru C, Munn Z, Aromataris E, Campbell J, Hopp L (2020). Chapter 3: Systematic reviews of effectiveness. In: Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI. Available from: https://synthesismanual.jbi.global.

39. Dorrius MD, Jansen-van der Weide MC, van Ooijen PM, Pijnappel RM, Oudkerk M. Computer-aided detection in breast MRI: a systematic review and meta-analysis. Eur Radiol. 2011; 21(8): 1600–1608. https://doi.org/10.1007/s00330-011-2091-9 PMID: 21404134

40. Cuocolo R, Cipullo MB, Stanzione A, Romeo V, Green R, Cantoni V, et al. Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. Eur Radiol. 2020; 30 (12): 6877–6887. https://doi.org/10.1007/s00330-020-07027-w PMID: 32607629

41. Xing X, Zhao X, Wei H, Li Y. Diagnostic accuracy of different computer-aided diagnostic systems for prostate cancer based on magnetic resonance imaging: A systematic review with diagnostic meta-analysis. Medicine (Baltimore). 2021; 100(3): e23817. https://doi.org/10.1097/MD.0000000000023817 PMID: 33545946

42. Zhao WJ, Fu LR, Huang ZM, Zhu JQ, Ma BY. Effectiveness evaluation of computer-aided diagnosis system for the diagnosis of thyroid nodules on ultrasound: A systematic review and meta-analysis. Medicine (Baltimore). 2019; 98(32): e16379. https://doi.org/10.1097/MD.0000000000016379 PMID: 31393347

43. Nindrea RD, Aryandono T, Lazuardi L, Dwiprahasto I. Diagnostic Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. Asian Pac J Cancer Prev. 2018; 19(7): 1747–1752. https://doi.org/10.22034/APJCP.2018.19.7.1747 PMID: 30049182

44. Eadie LH, Taylor P, Gibson AP. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. Eur J Radiol. 2012; 81(1): e70–76. https://doi.org/10.1016/j.ejrad.2011.01.098 PMID: 21345631

45. Winkel DJ, Tong A, Lou B, Kamen A, Comaniciu D, Disselhorst JA, et al. A Novel Deep Learning Based Computer-Aided Diagnosis System Improves the Accuracy and Efficiency of Radiologists in Reading Biparametric Magnetic Resonance Images of the Prostate: Results of a Multireader, Multicase Study. Invest Radiol. 2021; 56(10): 605–613. https://doi.org/10.1097/RLI.0000000000000780 PMID: 33787537

46. Fei B. Computer-aided diagnosis of prostate cancer with MRI. Curr Opin Biomed Eng. 2017; 3: 20–27. https://doi.org/10.1016/j.cobme.2017.09.009 PMID: 29732440

47. Hussain L, Ahmed A, Saeed S, Rathore S, Awan IA, Shah SA, et al. Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. Cancer Biomark. 2018; 21(2): 393–413. https://doi.org/10.3233/CBM-170643 PMID: 29226857

48. Gaur S, Lay N, Harmon SA, Doddakashi S, Mehralivand S, Argun B, et al (2018) Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? a multi-center, multi-reader investigation. Oncotarget. 2018; 9(73): 33804–33817. https://doi.org/10.18632/oncotarget.26100 PMID: 30333911

49. Singh S, Maxwell J, Baker JA, Nicholas JL, Lo JY. Computer-aided classification of breast masses: performance and interobserver variability of expert radiologists versus residents. Radiology. 2011; 258(1): 73–80. https://doi.org/10.1148/radiol.10081308 PMID: 20971779

50. Peters AA, Decasper A, Munz J, Klaus J, Loebelenz LI, Hoffner MKM, et al. Performance of an AI based CAD system in solid lung nodule detection on chest phantom radiographs compared to radiology residents and fellow radiologists. J Thorac Dis.2021; 13(5): 2728–2737. https://doi.org/10.21037/jtd-20-3522 PMID: 34164165

51. Watanabe Y, Tanaka T, Nishida A, Takahashi H, Fujiwara M, Fujiwara T, et al. Improvement of the diagnostic accuracy for intracranial haemorrhage using deep learning-based computer-assisted detection. Neuroradiology. 2021; 63(5): 713–720. https://doi.org/10.1007/s00234-020-02566-x PMID: 33025044

52. Giannini V, Mazzetti S, Cappello G, Doronzio VM, Vassallo L, Russo F, et al. Computer-Aided Diagnosis Improves the Detection of Clinically Significant Prostate Cancer on Multiparametric-MRI: A Multi-Observer Performance Study Involving Inexperienced Readers. Diagnostics (Basel) 2021; 11(6): 973. https://doi.org/10.3390/diagnostics11060973 PMID: 34071215

53. Gilbert FJ, Astley SM, McGee MA, Gillan MG, Boggis CR, Griffiths PM, et al. Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. Radiology. 2006; 241(1): 47–53. https://doi.org/10.1148/radiol.2411051092 PMID: 16990670

54. Wormanns D, Beyer F, Diederich S, Ludwig K, Heindel W. Diagnostic performance of a commercially available computer-aided diagnosis system for automatic detection of pulmonary nodules: comparison with single and double reading. Rofo. 2004; 176(7): 953–958. https://doi.org/10.1055/s-2004-813251 PMID: 15237336

55. Karssemeijer N, Otten JD, Verbeek AL, Groenewoud JH, de Koning HJ, Hendriks JH, et al. Computer-aided detection versus independent double reading of masses on mammograms. Radiology. 2003; 227(1): 192–200. https://doi.org/10.1148/radiol.2271011962 PMID: 12616008

56. Ciatto S, Ambrogetti D, Bonardi R, Brancato B, Catarzi S, Risso G, et al. Comparison of two commercial systems for computer-assisted detection (CAD) as an aid to interpreting screening mammograms. Radiol Med. 2004; 107(5–6): 480–488. PMID: 15195010

57. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. BMJ. 2021; 374: n1872. https://doi.org/10.1136/bmj.n1872 PMID: 34470740

58. Chen ZH, Lin L, Wu CF, Li CF, Xu RH, Sun Y. Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine. Cancer Commun (Lond). 2021; 41(11): 1100–1115. https://doi.org/10.1002/cac2.12215 PMID: 34613667

59. Balkenende L, Teuwen J, Mann RM. Application of Deep Learning in Breast Cancer Imaging. Semin Nucl Med. 2022; 52(5): 584–596. https://doi.org/10.1053/j.semnuclmed.2022.02.003 PMID: 35339259

60. Chen J, You H, Li K. A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images. Comput Methods Programs Biomed. 2020; 185: 105329. https://doi.org/10.1016/j.cmpb.2020.105329 PMID: 31955006

61. Din NMU, Dar RA, Rasool M, Assad A. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. Comput Biol Med. 2022; 149: 106073. https://doi.org/10.1016/j.compbiomed.2022.106073 PMID: 36103745

62. Grieve P. Deep Learning vs. Machine Learning: What's the Difference? 2020. Avaliable in: https://www.zendesk.com/blog/machine-learning-and-deep-learning.

63. MathWorks. What Is Deep Learning? Avaliable in: https://www.mathworks.com/discovery/deep-learning.html.

64. Majumder A, Sen D. Artificial intelligence in cancer diagnostics and therapy: current perspectives. Indian J Cancer. 2021; 58(4): 481–492. https://doi.org/10.4103/ijc.IJC_399_20 PMID: 34975094

65. Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer Sci. 2020; 111(5): 1452–1460. https://doi.org/10.1111/cas.14377 PMID: 32133724

66. Ibrahim S, Nazir S, Velastin SA. Feature Selection Using Correlation Analysis and Principal Component Analysis for Accurate Breast Cancer Diagnosis. J Imaging. 2021; 7(11): 225. https://doi.org/10.3390/jimaging7110225 PMID: 34821856

67. University of South Florida. Digital Mammography. DDSM: Digital Database for Screening Mammography. Avaliable in: http://www.eng.usf.edu/cvprg/mammography/database.html.

68. Mammographic Image Analysis Homepage. Databases. Avaliable in: https://www.mammoimage.org/databases/.

69. Breast Cancer Digital Repository (BCDR). Avaliable in: https://bcdr.eu/information/about.

70. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013; 26(6): 1045–1057. https://doi.org/10.1007/s10278-013-9622-7 PMID: 23884657

71. McNitt-Gray MF, Armato SG 3rd, Meyer CR, Reeves AP, McLennan G, Pais RC, et al. The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation. Acad Radiol. 2007; 14(12): 1464–1474. https://doi.org/10.1016/j.acra.2007.07.021 PMID: 18035276

72. Bonavita I, Rafael-Palou X, Ceresa M, Piella G, Ribas V, González Ballester MA. Integration of convolutional neural networks for pulmonary nodule malignancy assessment in a lung cancer classification pipeline. Comput Methods Programs Biomed. 2020; 185: 105172. https://doi.org/10.1016/j.cmpb.2019.105172 PMID: 31710985

73. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011; 38(2): 915–931. https://doi.org/10.1118/1.3528204 PMID: 21452728

74. Samek W, Montavon G, Vedaldi A, Hansen LK, Muller KR. (Eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science. v. 11700. 2019. https://doi.org/10.1007/978-3-030-28954-6

75. Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, et al. Explainable and Interpretable Models in Computer Vision and Machine Learning. The Springer Series on Challenges in Machine Learning. 2018 https://doi.org/10.1007/978-3-319-98131-4

76. Biran O, Cotton C. Explanation and justification in machine learning: A survey. In IJCAI-17 workshop on explainable AI (XAI) 2017; 8(1): 8–13. Avaliable in: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.

77. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI-Explainable artificial intelligence. Science Robotics, 2019; 4(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120 PMID: 33137719

78. Kulesza T, Burnett M, Wong WK, Stumpf S. Principles of Explanatory Debugging to personalize interactive machine learning. In: Brdiczka O. & Chau P (Eds.), Proceedings of the 20th International Conference on Intelligent User Interfaces. New York, USA; 2015. pp. 126–137. https://doi.org/10.1145/2678025.2701399

79. Clark HH, Brennan SE. Grounding in communication. In Resnick L. B., Levine J. M., & Teasley S. D. (Eds.), Perspectives on socially shared cognition. American Psychological Association; 1991. pp. 127–149. https://doi.org/10.1037/10096-006

80. Wang D, Yang Q, Abdul A, Lim B Y. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems. 2019. pp. 1–15. https://doi.org/10.1145/3290605.3300831

81. Arrieta AB, D´ıaz-RodrÍguez N, Ser JD, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion 2020; 58:(C) 82–115 https://doi.org/10.1016/j.inffus.2019.12.012

82. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). Turin, Italy; 2018. pp. 80–89. https://doi.org/10.1109/DSAA.2018.00018

83. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence, 2019; 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007

84. Miller T, Howe P, Sonenberg L. Explainable AI: beware of inmates running the asylum. or: how i learnt to stop worrying and love the social and behavioural sciences. In: Aha DW, Dar-rell T, Pazzani M, Reid D, Sammut C, Stone P (eds) Proceedings of the IJCAI 2017 workshop on explainable artificial intelligence (XAI). IJCAI, Santa Clara County, CA; 2017. pp. 36–42.

85. Razavian AS, Azizpour H, Sullivan J, Carlsson S. IEEE 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)—Columbus, OH, USA. 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops—CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. ArXiv: 1403.6382v3 [Preprint] 2014 [Submitted on 23 Mar 2014 (v1), last revised 12 May 2014 (this version, v3); cited 2023 february 17] Available from: https://arxiv.org/abs/1403.6382 https://doi.org/10.1109/cvprw.2014.131 arXiv:1403.6382.

86. Du S, Guo H, Simpson A. Self-driving car steering angle prediction based on image recognition. ArXiv: 1912.05440v1 [Preprint] 2019 [Submitted on 11 Dec 2019; cited 2023 february 17] Available from: https://arxiv.org/abs/1912.05440 https://doi.org/10.48550/arXiv.1912.05440

87. Garnelo M, Arulkumaran K, Shanahan M. Towards deep symbolic reinforcement learning. ArXiv: 1609.05518v2 [Preprint] 2016 [Submitted on 18 Sep 2016 (v1), last revised 1 Oct 2016 (this version,

v2); cited 2023 february 17] Available from: https://arxiv.org/abs/1609.05518 https://doi.org/10.48550/arXiv.1609.05518

88. Garcez ADA, Gori M, Lamb LC, Serafini L, Spranger M, Tran SN. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. ArXiv: 1905.06088v1 [Preprint] 2019 [Submitted on 15 May 2019; cited 2023 february 17] Available from: https://arxiv.org/abs/1905.06088 https://doi.org/10.48550/arXiv.1905.06088

89. Li Y, Yosinski J, Clune J, Lipson H, Hopcroft J. Convergent learning: Do different neural networks learn the same representations? ArXiv: 1511.07543v3 [Preprint] 2015 [Submitted on 24 Nov 2015 (v1), last revised 28 Feb 2016 (this version, v3); cited 2023 february 17] Available from: https://arxiv.org/abs/1511.07543 https://doi.org/10.48550/arXiv.1511.07543

90. Goudet O, Kalainathan D, Caillou P, Guyon I, Lopez-Paz D, Sebag M. Learning functional causal models with generative neural networks, In: Escalante HJ, Guyon I, Escalera S editors. Explainable and Interpretable Models in Computer Vision and Machine Learning. Springer; 2018. pp. 39–80. https://doi.org/10.1007/978-3-319-98131-4

91. Lopez-Paz D, Nishihara R, Chintala S, Scholkopf B, Bottou L. Discovering causal signals in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. arXiv:1605.08179v2 [Preprint] 2017 [Submitted on 26 May 2016 (v1), last revised 31 Oct 2017 (this version, v2); cited 2023 february 17] pp. 6979–6987. Available from: https://arxiv.org/abs/1605.08179 https://doi.org/10.48550/arXiv.1605.08179

92. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? ArXiv: 1411.1792v1 [Preprint] 2014 [Submitted on 6 Nov 2014; cited 2023 february 17] Available from: https://arxiv.org/abs/1411.1792 https://doi.org/10.48550/arXiv.1411.1792

93. Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 2017; 7(2): 76–99. https://doi.org/10.1093/idpl/ipx005

94. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion 2021; 76: 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

95. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In Machine learning for healthcare conference. ArXiv: 1905.05134v2 [Preprint] 2019 [Submitted on 13 May 2019 (v1); last revised 7 Aug 2019 (this version, v2); cited 2023 february 17]. Available from: https://arxiv.org/abs/1905.05134 https://doi.org/10.48550/arxiv.1905.05134

96. Lambert B, Forbes F, Tucholka A, Doyle S, Dehaene H, Dojat M. Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. ArXiv: 2210.03736v1 [Preprint] 2022 [Submitted on 5 Oct 2022; cited 2023 february 17]. Available from: https://arxiv.org/abs/2210.03736 https://doi.org/10.48550/arXiv.2210.03736

97. Bulusu S, Kailkhura B, Li B, Varshney PK, Song D. Anomalous Example Detection in Deep Learning: A Survey. ArXiv: 2003.06979v2 [Preprint] 2020 [cited 2023 february 17]. Available from: https://arxiv.org/abs/2003.06979 https://doi.org/10.48550/arXiv.2003.06979

98. Wu Z, Wang F, Cao W, Qin C, Dong X, Yang Z, et al. Lung cancer risk prediction models based on pulmonary nodules: A systematic review. Thorac Cancer. 2022; 13(5): 664–677. https://doi.org/10.1111/1759-7714.14333 PMID: 35137543

99. Chiu HY, Chao HS, Chen YM. Application of Artificial Intelligence in Lung Cancer. Cancers (Basel). 2022; 14(6): 1370. https://doi.org/10.3390/cancers14061370 PMID: 35326521

100. Wang DD, Zhou W, Yan H, Wong M, Lee V. Personalized prediction of EGFR mutation-induced drug resistance in lung cancer. Sci Rep. 2013; 3: 2855. https://doi.org/10.1038/srep02855 PMID: 24092472

101. Giang TT, Nguyen TP, Tran DH. Stratifying patients using fast multiple kernel learning framework: case studies of Alzheimer's disease and cancers. BMC Med Inform Decis Mak. 2020; 20(1): 108. https://doi.org/10.1186/s12911-020-01140-y PMID: 32546157

102. Gao Y, Zhou R, Lyu Q. Multiomics and machine learning in lung cancer prognosis. J Thorac Dis. 2020; 12(8): 4531–4535. https://doi.org/10.21037/jtd-2019-itm-013 PMID: 32944369

103. Wissel D.; Rowson D.; Boeva V. Hierarchical autoencoder-based integration improves performance in multi-omics cancer survival models through soft modality selection. BioRxiv [Preprint] 2022 BioRxiv: 2021.09.16.460589 [cited 2023 february 17]. Available from: https://www.biorxiv.org/content/10.1101/2021.09.16.460589v3.full.pdf https://doi.org/10.1101/2021.09.16.460589

104. Wu G, Jochems A, Refaee T, Ibrahim A, Yan C, Sanduleanu S, et al. Structural and functional radiomics for lung cancer. Eur J Nucl Med Mol Imaging. 2021; 48(12): 3961–3974. https://doi.org/10.1007/s00259-021-05242-1 PMID: 33693966

105. Warnat-Herresthal S, Schultze H, Shastry KL, et al: Swarm Learning for decentralized and confidential clinical machine learning. Nature 2021; 594 (7862): 265–270. https://doi.org/10.1038/s41586-021-03583-3 PMID: 34040261

106. Bhowmik A, Eskreis-Winkler S. Deep learning in breast imaging. BJR Open. 2022; 4(1): 20210060. https://doi.org/10.1259/bjro.20210060 PMID: 36105427

107. Huang B, Sollee J, Luo YH, Reddy A, Zhong Z, Wu J, et al. Prediction of lung malignancy progression and survival with machine learning based on pre-treatment FDG-PET/CT. EBioMedicine. 2022; 82: 104127. https://doi.org/10.1016/j.ebiom.2022.104127 PMID: 35810561

108. Rieke N, Hancox J, Li W, et al: The future of digital health with feder-ated learning. npj Digit. Med. 2020; 3(1): 1–7. https://doi.org/10.1038/s41746-020-00323-1 PMID: 33015372

109. Jochems A, Deist TM, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital—A real life proof of concept. Radiother Oncol. 2016; 121(3): 459–467. https://doi.org/10.1016/j.radonc.2016.10.002 PMID: 28029405

110. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries. Int J Radiat Oncol Biol Phys. 2017; 99(2): 344–352. https://doi.org/10.1016/j.ijrobp.2017.04.021 PMID: 28871984