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

Helbert Eustáquio Cardoso da Silva (helbertcardososilva@gmail.com)  
University of Brasília  
Glaucia Nize Martins Santos  
University of Brasília  
Carla Ruffeil Moreira Mesquita  
University of Brasília  
André Ferreira Leite  
University of Brasília  
Paulo Tadeu de Souza Figueiredo  
University of Brasília  
Cristine Miron Stefani  
University of Brasília  
Nilce de Santos Melo  
University of Brasília  

Research Article  

Keywords: Artificial Intelligence, Neoplasms, Cancer, Algorithm, Overview  

Posted Date: October 24th, 2022  

DOI: https://doi.org/10.21203/rs.3.rs-2184114/v1

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

The aim of this overview article is to discuss the application of artificial intelligence (AI) tools in detecting and diagnosing malignant tumors based on different imaging modalities. The acronym PIRDs was used to create a search strategy. A comprehensive literature search was conducted on indexed databases and grey literature 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. In total, 382 records were retrieved in the databases, 364 after removing duplicates, 32 satisfied the full-text reading criterion, and 9 papers were considered for qualitative synthesis. 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. The Super Vector Machine algorithm 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. The use of AI tools benefitted less experienced radiologists more than experienced specialists on the use of machine learning and radiomic analysis in cancer identification. The combination of a CAD system, machine learning algorithms, and radiomic analysis seemed to be effective and promising in the identification and diagnosis of malignant tumors. However, further longitudinal studies with a longer follow-up duration are required for a better understanding of the clinical application of these artificial intelligence systems.

1 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 (Yu 2018). 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 (Bejnordi 2017; Kourou 2015; Chen 2020).

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 (Yu 2018; Bejnordi 2017). 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 (Kim 2021; Tran 2021; Das 2021; Zarzeczny 2021).

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 (LeCun 2015). Progress in Deep Learning (DL) approaches has shown gains in image-based diagnosis and illness detection in the study of cancer and oncology (Baptista 2021; Esteva 2019). 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 (Miotta 2018).

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 more likely 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 (Hayward 2008); b) reducing interobserver variability (Chlebus 2019); and c) increasing the speed of image analysis and reports (Miller 2018; Matheson 2018). 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 (Zaharchuk 2018), prognosis and treatment of diseases, such as brain tumors (Siuly 2016).

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 (Cucchetti 2007), malignant melanoma and breast cancer (Carrara 2007; Papadopoulos 2005), and colon cancer (Selaru 2002). Image processing, pattern recognition, artificial intelligence, and medical pictures are all combined in Computer-Aided Diagnosis (CAD) 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 (Castellino 2005; Nishikawa 2010). Therefore, CAD systems might potentially decrease the workload of radiologists, leading to fast and accurate diagnoses.

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?

2 Materials And Methods

2.1 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 systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-
analyses, following the PRISMA checklist (http://www.prisma-statement.org/).

In view of the lack of a clear and standardized definition of what a systematic review is, the following criteria were adopted to define a systematic review, which is a review that reports or includes:

1. i) research question;
2. 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;

1. iv) selection (screening) methods;
2. v) critically appraises and reports the quality/risk of bias of the included studies;
3. vi) information about data analysis and synthesis that allows the reproducibility of the results; (Krnic Martinic 2019)

2.2 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), Science Citation Index Expanded, Web of Science, Latin American and Caribbean Health Sciences (LILACS), Excerpta Medica 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 Additional file 1. 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.

2.3 Inclusion and exclusion criteria

PIRDs (Participants, Index test, Reference Test, Diagnosis of Interest and Studies) 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.

2.4 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 (Krnic Martinic 2019; Higgins 2021):

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.

2.5 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 (Azavedo 2012; Tabatabaei 2021; Henriksen 2019).

2.6 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) (Tufanaru 2020). It should be noted that risk of bias assessment instruments classically indicated for systematic reviews such as AMSTAR 2 and ROBIS were not used in this overview because these tools were developed for interventional systematic reviews, and the papers included were diagnostic systematic reviews.

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.

2.7 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.

3 Results

3.1 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. Additional file 2 presents excluded articles and reasons for exclusion. A flowchart of the complete process inclusion is shown in Figure 1.

Included studies were conducted in EUA (Tabatabaei 2021), Netherlands (Dorrius 2011), Italy (Cuocolo 2020), Sweden (Azavedo 2012), China (Xing 2021; Zhao 2019), Indonesia (Nindrea 2018), United Kingdom (Eadie 2012) and Denmark (Henriksen 2019). All included studies were published in English. One RS included descriptive studies (Dorrius 2011), three RS included diagnostic accuracy studies (Cuocolo 2020; Nindrea 2018; Eadie 2012), four RS included prospective or retrospective studies (Tabatabaei 2021; Azavedo 2012; Xing 2021; Zhao 2019) and one RS included clinical trial studies (Henriksen 2019). 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.

Table 1 Summary of descriptive characteristics of included articles (n=09).
<table>
<thead>
<tr>
<th>Author, year, country and design studies</th>
<th>Included Studies</th>
<th>Type of cancer</th>
<th>Index test</th>
<th>Reference test</th>
<th>True positives / N of images</th>
<th>True Negatives / N of images</th>
<th>Sensitivity and Specificity/ odds ratio, Mean±SD, p’ value</th>
<th>Diagnostic accuracy (%) , Mean±SD, p’ value</th>
<th>Cor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Radiologist no CAD, general</td>
<td>CAI infl sen spe the perf rad exp breast dia dia by i ressRad with exp see for sys pers breast eva</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82% (95% CI: 72%–90%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Radiologist with CAD, general</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>89% (95% CI: 83%–93%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Specificity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Radiologist no CAD, general</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>81% (95% CI: 74%–87%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Radiologist with CAD, general</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>81% (95% CI: 76%–85%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nindrea et al, 2018, Indonesia, Diagnostic Accuracy studies</td>
<td>11</td>
<td>Breast cancer</td>
<td>Machine Learning Algorithms</td>
<td>Mammography (MM)</td>
<td>SVM</td>
<td>SVM</td>
<td>Sensitivity</td>
<td>SVM: 99.51%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ANN: 97.3%;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>DT: 95.13%;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NB: 95.99%;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KNM: 95.27%;</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The year of t can m amo bre the Adc the met des forr bas sm ap x</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>wom will the bre will acc inf eas life</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Year</td>
<td>Country</td>
<td>Type of Studies</td>
<td>Cancer Types</td>
<td>CADe; Diagnostic CAD (CADx)</td>
<td>MM;</td>
<td>CADx by Samsung</td>
<td>Sensitivity (SD)</td>
<td>Diagnostic odds ratio (DOR) (SD)</td>
</tr>
<tr>
<td>------------------------</td>
<td>------</td>
<td>---------</td>
<td>----------------</td>
<td>----------------------------------------</td>
<td>-------------------------------</td>
<td>-----</td>
<td>-----------------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Eadie et al, 2012, UK</td>
<td>48</td>
<td></td>
<td>Diagnostic Accuracy studies</td>
<td>Breast cancer, lung cancer, liver cancer, prostate cancer, bone cancer, bowel cancer, skin cancer, neck cancer</td>
<td>CADe; Diagnostic CAD (CADx)</td>
<td>MM;</td>
<td>CADx by Samsung</td>
<td>80.41±1.46</td>
<td>3.63±0.16</td>
</tr>
<tr>
<td>Zhao et al, 2019, CH</td>
<td>5</td>
<td></td>
<td>Prospective or Retrospective studies</td>
<td>Thyroid (nodules) cancer</td>
<td>CADx system</td>
<td>US</td>
<td>CADx system</td>
<td>83.00±14.46</td>
<td>3.44±0.79</td>
</tr>
</tbody>
</table>
Azavedo et al, 2012, Sweden, Prospective or Retrospective studies

Cuocolo et al, 2020, Italy, Diagnostic Accuracy studies
<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Region</th>
<th>Tissue Type</th>
<th>Methodologies</th>
<th>CAD System</th>
<th>MRI</th>
<th>SVM</th>
<th>Sensitivity</th>
<th>AUC</th>
<th>Prostate Zones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xing et al, 2021, China, Retrospective studies</td>
<td>15</td>
<td>Prostate cancer (PCa)</td>
<td></td>
<td>CAD system; ANN; SVM</td>
<td>SVM</td>
<td>SVM</td>
<td>42.76%/608; 41.94%/608; 34.55%/301; 37.54%/301; 34.78%/738; 32.60%/738; 19.41%/1586; 65.15%/1586; 51.95%/256; 32.81%/256; 59.67%/186; 26.34%/186; 32.39%/71; 46.47%/71;</td>
<td>Sensitivity: 0.47 to 1.00</td>
<td>0.89 (95% CI: 0.86–0.91)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peripheral zone (PZ); SVM; Linear Discriminant Analysis (LDA); Radiomic Machine Learning (RML); Non-specific classifier (NSC);</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transitional zone (TZ);</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Central gland (CG);</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tabatabaei et al, 2021, USA Retrospective studies</td>
<td>18</td>
<td>Glioma</td>
<td>DT; KNN; SVM; RF;</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Henriksen EL et al, (2018), Denmark
Clinical trials

<table>
<thead>
<tr>
<th>13</th>
<th>Breast cancer</th>
<th>CAD system;</th>
<th>MM</th>
<th>-</th>
<th>-</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single Reading (SR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SR vs SR + CAD;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Double Reading (DR)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DR vs SR p CAD;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 Computed 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;

### 3.2 Methodological quality within Studies

None of the studies fulfilled all methodological quality criteria. However, five studies (Dorrius 2011; Cuocolo 2020; Xing 2021; Zhao 2019; Eadie 2012) were considered of “high” methodological quality, three studies (Azavedo 2012; Nindrea 2018; Henriksen 2019) were of “moderate” methodological quality and only one study (Tabatabaei 2021) was considered of “low” methodological quality.
In two studies (Tabatabaei 2021; Eadie 2012), the review question was not considered clearly and explicitly stated. The inclusion criteria was not appropriate for the review question in one study (Tabatabaei 2021), the sources and resources used to search for studies was not adequate in one study (Dorrius 2011), the likelihood of publication bias was not assessed in four studies (Tabatabaei 2021; Azavedo 2012; Nindrea 2018; Henriksen 2019), the recommendations for policies and/or practices supported by the reported data were unclear for a study (Henriksen 2019), and the specific directives for new research were inconclusive for three studies (Tabatabaei 2021; Xing 2021; Henriksen 2019). 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).

<table>
<thead>
<tr>
<th>Study</th>
<th>Methodological quality items assessed</th>
<th>Overall quality²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dorrius (2011)</td>
<td>Q1 Y Q2 Y Q3 N Q4 Y Q5 Y Q6 Y Q7 Y Q8 Y Q9 Y Q10 Y</td>
<td>High</td>
</tr>
<tr>
<td>Nindrea (2018)</td>
<td>Q1 Y Q2 Y Q3 Y Q4 Y Q5 Y Q6 N Q7 Y Q8 N Q9 Y</td>
<td>Moderate</td>
</tr>
<tr>
<td>Eadie (2012)</td>
<td>Q1 Y Q2 Y Q3 Y Q4 Y Q5 Y Q6 N Q7 Y Q8 Y Q9 Y</td>
<td>High</td>
</tr>
<tr>
<td>Zhao (2019)</td>
<td>Q1 Y Q2 Y Q3 N Q4 Y Q5 Y Q6 Y Q7 Y Q8 Y Q9 Y</td>
<td>High</td>
</tr>
<tr>
<td>Henriksen (2019)</td>
<td>Q1 Y Q2 Y Q3 N Q4 Y Q5 Y Q6 Y Q7 N Q8 Y Q9 Y</td>
<td>Moderate</td>
</tr>
<tr>
<td>Azavedo (2012)</td>
<td>Q1 Y Q2 Y Q3 Y Q4 Y Q5 Y Q6 Y Q7 Y Q8 Y Q9 Y</td>
<td>Moderate</td>
</tr>
<tr>
<td>Cuocolo (2020)</td>
<td>Q1 Y Q2 Y Q3 Y Q4 Y Q5 Y Q6 Y Q7 Y Q8 Y Q9 Y</td>
<td>High</td>
</tr>
<tr>
<td>Xing (2021)</td>
<td>Q1 Y Q2 Y Q3 Y Q4 Y Q5 Y Q6 Y Q7 Y Q8 Y Q9 Y</td>
<td>High</td>
</tr>
<tr>
<td>Tabatabaei (2021)</td>
<td>Q1 N Q2 U Q3 Y Q4 Y Q5 U Q6 U Q7 N Q8 Y Q9 U</td>
<td>Low</td>
</tr>
</tbody>
</table>

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?

²Low 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

### 3.3 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. (Dorrius 2011).

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), Naïve 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 (Nindrea 2018).

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 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 CAd. The review showed that CADx may offer...
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 (Zhao 2019).

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 (Henriksen 2019).

Researchers from Lund University, Skene 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 (Azavedo 2012).

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 (Cuocolo 2020).

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 (Xing 2021).

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 (Tabatabaei 2021).

### 3.4 Certainty of the evidence in the systematic reviews included

Only two articles (Azavedo 2012; Xing 2021) 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 (Xing 2021) 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 (Azavedo 2012) evaluated only one study regarding the quality 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.

### 3.5 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. Additional file 3 provides more details on the overlap and features of the primary studies.

4 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 oncological field. The selected studies demonstrated the use of computer-assisted detection (CAD) (Dorrius 2011; Azavedo 2012; Xing 2021; Zhao 2019; Eadie 2012; Henriksen 2019), machine learning algorithms (Cuocolo 2020; Xing 2021; Nindrea 2018) and radiomic analysis (Tabatabaei 2021) 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 (Xing 2021; Nindrea 2018). In four studies, CAD systems demonstrated some benefit in helping to detect cancer (Dorrius 2011; Xing 2021; Zhao 2019; Eadie 2012). 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 (Eadie 2012). 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 (Tabatabaei 2021; Cuocolo 2020).

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 specificity 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 (Xing 2021). Some papers corroborate the findings that radiologists benefit most from the use of CAD systems in the detection of prostate cancer lesions (Winkel 2021; Fei 2017; Hussain 2018; Gaur 2018).

However, in other study, less experienced radiologists benefited more from the use of artificial intelligence than experienced professionals (Dorrius 2011). 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 (Dorrius 2011). 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. (Singh 2011; Peters 2021; Watanabe 2021; Giannini 2021). 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 (Zhao 2019).

Two studies (Azavedo 2012; Henriksen 2019) found no significant evidence regarding sensitivity, specificity, and diagnostic accuracy, between single-reading or double-reading mammmography compared with single-reading plus CAD or double-reading plus CAD. The use of CADx 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 (Eadie 2012).

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 (Gilbert 2006). However, for detecting pulmonary nodules, the performance of a CAD system was comparable to a second opinion reading (Wormanns 2004).

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 (Freeman 2021).

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 (Tabatabaei 2021).

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 (Chen 2021).

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
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 (Chen 2021; Majumder 2021).

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 (Chen 2021; Majumder 2021; Shimizu 2020).

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 (Azavedo 2012) or a loss in the calculation of diagnostic accuracy measures (Henriksen 2019); relatively low number of studies (Cuocolo 2020); high heterogeneity can be partly explained by the diversity of methodological aspects, difference between patients, or diversity of techniques used (Tabatabaei 2021; Cuocolo 2020; Zhao 2019); presence of selection bias by choice of articles reporting sensitivity and specificity results (Eadie 2012), by use of retrospective studies, vaguely reported sample of patients (Tabatabaei 2021), by use of studies with relatively small samples (Xing 2021; Zhao 2019); presence of publication bias due to lack of studies with unfavorable data (Eadie 2012), by use of digitized analog radiographs to the detriment of digital images (Azavedo 2012; Henriksen 2019); behavior of radiologists in terms of training, conducting clinical tests and surveillance in the analysis. (Azavedo 2012; Eadie 2012); relatively small and old technology dataset number (Tabatabaei 2021; Nindrea 2018); presence of measurement bias due to the large difference between the groups studied and the small number of outcomes observed in the included studies.

There are still few that use AI tools for detecting malignant tumors on different imaging modalities, being limited to some more favorable types of cancer, such as breast cancer, prostate cancer and thyroid cancer. Furthermore, their methodologies are heterogeneous regarding the type of technology applied. 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.

5 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. 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.

Declarations

Funding: No funding was received to assist with the preparation of this manuscript.

Conflicts of interest/Competing interests: The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval: Not applicable Consent to participate: Not applicable Consent for publication: Not applicable

Availability of data and material: Not applicable Code availability: Not applicable

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.

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

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

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


44. Xing X, Zhao X, Wei H, Li Y (2021) Diagnostic accuracy of different pulmonary diagnostic systems for prostate cancer based on magnetic resonance imaging: A systematic review with diagnostic meta-analysis. Med (Baltim) 100:e23817. https://doi.org/10.1097/MD.0000000000023817


Figures
PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources

*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).

**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.


**Supplementary Files**

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

- Additionalfile1Databasesearchstrategy.docx
- Additionalfile2Excludedarticles.doc
- Additionalfile3OVERLAPPING.docx