Indian J Med Res 157, January 2023, pp 11-21
DOI: 10.4103/ijmr.IJMR_555_20

Quick Response Code:

Scoping Review

Extent of use of artificial intelligence & machine learning protocols in cancer diagnosis: A scoping review

Amit Dang, Dimple Dang & B. N. Vallish

MarksMan Healthcare Communications, Hyderabad, Telangana, India

Received March 3, 2020

***Background & objectives*: Artificial intelligence (AI) and machine learning (ML) have shown promising results in cancer diagnosis in validation tests involving retrospective patient databases. This study was aimed to explore the extent of actual use of AI/ML protocols for diagnosing cancer in prospective settings.**

***Methods*: PubMed was searched for studies reporting usage of AI/ML protocols for cancer diagnosis in prospective (clinical trial/real world) setting with the AI/ML diagnosis aiding clinical decision-making, from inception till May 17, 2021. Data pertaining to the cancer, patients and the AI/ML protocol were extracted. Comparison of AI/ML protocol diagnosis with human diagnosis was recorded. Through a *post hoc* analysis, data from studies describing validation of various AI/ML protocols were extracted.**

***Results*: Only 18/960 initial hits (1.88%) utilized AI/ML protocols for diagnostic decision-making. Most protocols used artificial neural network and deep learning. AI/ML protocols were utilized for cancer screening, pre-operative diagnosis and staging and intra-operative diagnosis of surgical specimens. The reference standard for 17/18 studies was histology. AI/ML protocols were used to diagnose cancers of the colorectum, skin, uterine cervix, oral cavity, ovaries, prostate, lungs and brain. AI/ML protocols were found to improve human diagnosis, and had either similar or better performance than the human diagnosis, especially made by the less experienced clinician. Validation of AI/ML protocols was described by 223 studies of which only four studies were from India. Also there was a huge variation in the number of items used for validation.**

***Interpretation & conclusions*: The findings of this review suggest that a meaningful translation from the validation of AI/ML protocols to their actual usage in cancer diagnosis is lacking. Development of regulatory framework specific for AI/ML usage in healthcare is essential.**

Key words Artificial intelligence - machine learning - neoplasms - diagnosis

Artificial intelligence (AI) is a field of computer science which deals with creation of computers that can perform complex tasks usually associated with intelligent human behaviour[1]. AI can improve the accuracy of diagnosis and therapeutic outcomes of several health conditions[1,2]. The most successful branch of AI is machine learning (ML), which deals with developing machine systems capable of 'learning' to identify hidden patterns within data, and perform specified tasks without further

programming[3]. Deep learning (DL) is a subspecialty of ML in which diverse computational models with multiple, hidden data processing layers perform automated feature extraction and pattern recognition, from larger datasets[4,5]. Some commonly used forms of AI technologies include artificial neural networks (ANNs), convolutional neural networks (CNNs), support vector machines (SVMs) and classification trees[4-6].

Among the various medical fields witnessing a notable positive impact from AI/ML is oncology[3-7]. A large number of AI/ML methods for cancer diagnosis are currently in various stages of development and validation across the world[5]. During validation and testing of most of the promising AI/ML methods, encouraging results have been observed in terms of sensitivity, specificity, and accuracy. Various AI/ML methods have been consistently observed to be non-inferior (and in some cases, superior) to human diagnosis in such studies[6]. Recently, a Google-based AI system was shown to outperform radiologists in interpreting screening mammography images[7]. However, since these validation tests have been conducted in experimental conditions, using retrospective patient databases, these are prone to different types of bias, including selection bias and verification bias[8,9]. The question as to how these protocols would perform under real-world conditions and whether the physicians would routinely adopt these AI/ML systems for clinical decision-making based on their superlative performance in validation tests has not been adequately addressed.

With this background, a scoping review was planned for systematically mapping the research done in this area by reviewing published literature about the extent of actual usage of AI/ML protocols in cancer diagnosis in prospective (clinical trial/real world) settings, such that the diagnosis by the AI/ML protocol aids in clinical decision-making.

Methods

The protocol for this review was drafted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for scoping reviews[10]. The protocol was revised by the research team and the final version of the protocol was registered prospectively with Open Science Framework on January 3, 2020 (*https://osf.io/643uq*).

Search strategy: All published articles were searched for eligibility using the following Population, Intervention, Comparator, and Outcomes approach. 'Population' was all articles which included patients with any type of cancer. There was no restriction on age or gender of the patients. 'Intervention' included papers which had described the actual usage of AI/ML protocol for diagnosis of cancer in such a way that the AI/ML diagnosis resulted in or aided in clinical decision-making. 'Comparator' included all comparators; no restriction was applied on the type of comparator. 'Outcome' included studies describing any outcome which described the application of AI/ML in cancer diagnosis.

Inclusion/exclusion criteria: Only those studies in which patients were prospectively enrolled, either in a clinical trial or a real-world setting, and in which an AI/ML protocol was used to newly diagnose a cancer or for performing staging of a patient already diagnosed with cancer, thereby facilitating clinical decision-making were included. Studies involving retrospective analysis of data, and studies which had utilized any type of database for training, validation and testing of an AI/ML protocol, were excluded.

Studies wherein an AI/ML protocol was used for estimating prognosis, performing therapy (such as robotic surgery) or any other applications apart from diagnosis of cancer were also excluded. Finally, all studies involving non-human participants, reviews, editorials, commentaries and other articles types apart from prospective studies were excluded. The search was restricted to include studies published in English language only, since the study team was proficient in English language.

Literature review: A systematic literature search was performed in MEDLINE/PubMed from their inception till May 17, 2021 using a combination of MeSH terms and Boolean operators. The search strategy was drafted by an in-house expert VBN. The final version of the PubMed search strategy is provided in Supplementary Table I. Using the eligibility criteria detailed above, the titles and abstracts of all retrieved records were separately and independently scanned; full texts of all potentially relevant records were assessed for eligibility. In addition, reference lists of the eligible papers were also hand-searched to identify other eligible records. Grey literature search was performed on Google Scholar using relevant search terms, and the first 200 hits were screened for eligible records. Once all the eligible records were pooled, relevant data from the papers were extracted into a predefined data table after reading through the full texts of the individual

studies. Methodological quality of the included studies was assessed using the QUADAS-2 tool[11], and the overall quality of each paper was assessed based on their completeness in four different domains and the presence of any applicability concerns[11].

Data analysis: After the planned data extraction as detailed above, a *post hoc* analysis of all the retrieved records was performed to identify studies which described validation of AI/ML protocol (either using standardized patient databases or prospectively enrolled patients) without their actual usage. Data pertaining to the types of cancer studied, the nature of AI/ML protocol being employed, the year of publication of the study, the country of the first author, location of the study site and the number of patients/lesions/images being used for the validation of the AI/ML protocol were extracted.

The record screening, data extraction and quality assessment were performed independently and separately by two reviewers (AD and DD) and any disagreements were discussed and resolved with the help of a third reviewer.

All data were entered electronically and analyzed in Microsoft Excel. Inter-rater reliability (IRR) and Cohen's kappa were calculated using Microsoft Excel 2019. The cut-offs for the kappa score were defined as ≤0.20=slight agreement, 0.21-0.40=fair agreement, 0.41-0.60=moderate agreement, 0.61-0.80=substantial agreement, 0.81-0.99=near-perfect agreement and 1.00=perfect agreement[12].

Results

Eighteen prospective studies published between 1996 and 2021 were included in this literature review from an initial pool of 960 eligible records[8,9,13-28]. The PRISMA flowchart is depicted in Figure 1. The IRR for selection of papers for review between the two reviewers was substantial, with Cohen's kappa value being 0.649 and 0.627 for article screening and eligibility assessment, respectively. The methodological quality of the included papers assessed as per the QUADAS-2 tool is depicted in Figure 2 and Table I. The first author of the 18 included studies came from 10 different countries. One study recruited patients from five countries; the remaining 17 studies were conducted in 10 different countries. The important characteristics of all included studies are presented in Table II.

All 18 studies had a prospective and observational study design, and eight studies had randomized the patient recruitment. AI/ML protocol was used for screening of cervical cancer in asymptomatic women (2 studies), for pre-operatively differentiating between benign and malignant lesions (14 studies), for pre-operative staging of prostate cancer already diagnosed by other means (1 study) and for intra-operative diagnosis of surgical specimen (1 study). The AI/ML protocol was based on ANN (4 studies), DL with CNN (7 studies), DL architecture (3 studies), SVM (1 study) and multifactorial decision-support expert system (1 study) and two studies did not specify the nature of the AI/ML protocol. The diagnosis provided by the AI/ML protocol before any medical intervention was confirmed in 17/18 studies by biopsy and histology of the lesions, which was the 'reference standard' test; in one study which compared AI-assisted reading of chest computed tomography scans with double reading by radiologists, the reference standard was combined diagnosis by AI-assisted protocol and an expert radiologist[22]. Cancers studied included colorectal cancer (8 studies), skin cancer (3 studies), cervical cancer screening (2 studies) and cancers of oral cavity, ovaries, prostate, lungs and brain (1 study each).

Not all included studies uniformly reported the performance of the AI/ML protocol in terms of sensitivity, specificity, positive predictive value and negative predictive value. The available values are presented in Table III. Fifteen studies compared the diagnostic performance of the AI/ML protocol with human diagnosis; the conclusion of these comparisons varied from 'AI/ML protocol is similar to human diagnosis' (2 studies), 'AI/ML protocol improves human diagnosis' (8 studies) and 'AI/ML protocol is better than human diagnosis' (4 studies). One study[9] observed that the performance of the AI/ML protocol differed with the basic experience and expertise of the user, while the AI/ML protocol improved the performance of the non-expert, the performance of the expert was better than that of the AI/ML protocol. None of the included studies contained information pertaining to the cost-effectiveness of the AI technology.

Post hoc analysis findings: Validation of an AI/ML protocol in cancer diagnosis was described by 223 studies (Supplementary Tables II-V and Supplementary Figure). These studies described 23 different types of cancer, the most frequent cancers being cancers of breast, skin and prostate [63 (28.3%), 24 (10.8%) and 22 (9.9%) studies, respectively]. The vast majority of the validation papers used ANNs

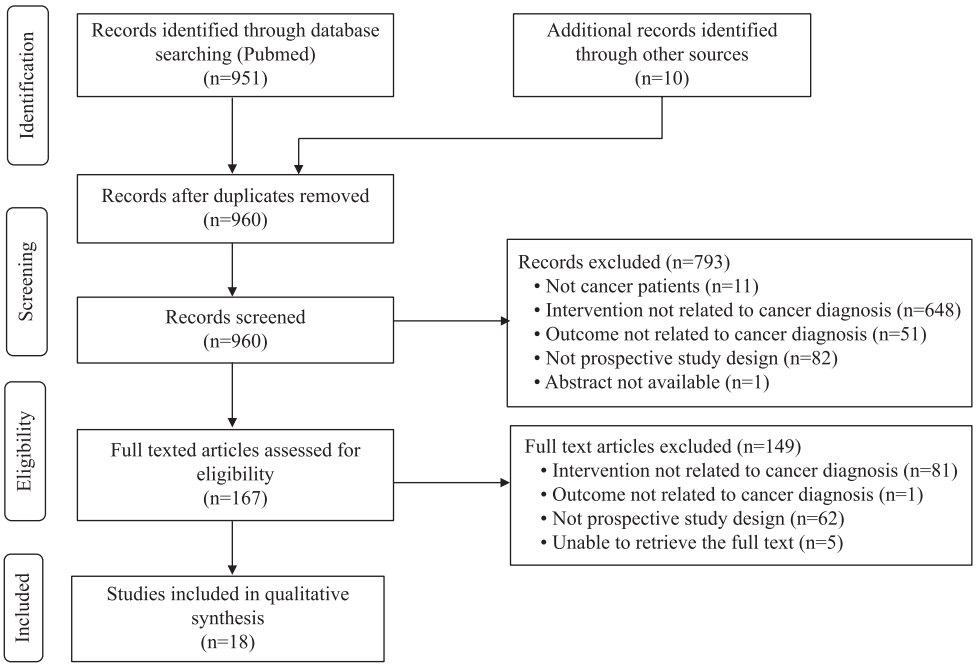


Fig. 1. Study selection process: PRISMA flow diagram. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

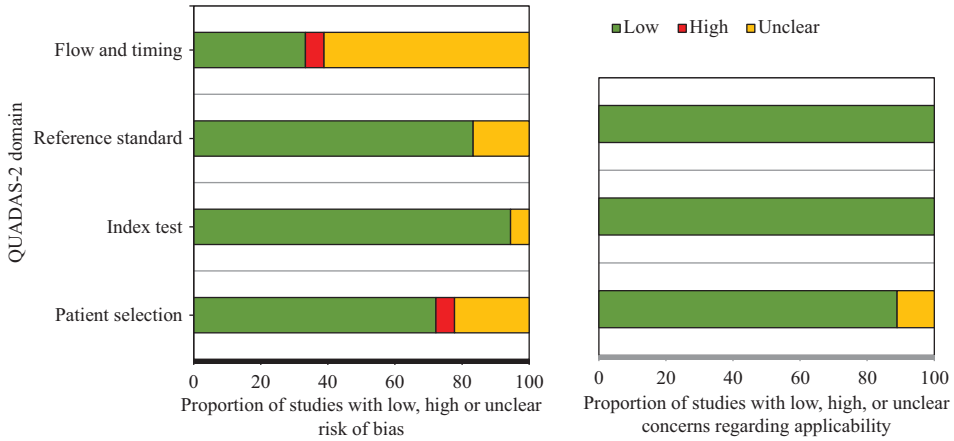


Fig. 2. Methodological quality of included studies as per QUADAS-2 assessment. QUADAS-2, quality assessment of diagnostic accuracy studies

and SVMs for developing the AI/ML protocol. The validation studies were published between 1993 and 2021. An increasing trend with time was observed in the number of publications of validation studies, with 79/223 studies being published after 2018. A huge variation in the number of samples/patients/lesions/images included for validation of the AI/ML protocol was observed, with patient numbers ranging from eight to 84,424, and image/lesion numbers varying from 15 to 1,036,496. The first authors of the 223 included studies came from 35 different countries across the globe, with the top three countries being the USA, China and Germany [72 (32.3%), 26 (11.7%) and 17 (7.6%) studies, respectively]. There were four validation studies from India[29-32], which described the validation of AI/ML protocols based on SVM (2 studies), CNN and a genetic algorithm. All these four studies described the diagnosis of cancers involving liver, brain, breast and oral cavity. The lead authors of all these four papers hailed from engineering and technological institutions.

Discussion

The healthcare market for AI/ML is proliferating rapidly, and is expected to reach USD 6.6 billion by 2021[1]. AI/ML is expected to have an increasingly

Table I. Methodological quality of included studies as per QUADAS-2 assessment

Study	Risk of bias				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow & timing	Patient selection	Index test	Reference standard
Kok *et al*[13], 1996	☺	☺	?	?	☺	☺	☺
Chang *et al*[14], 1999	?	☺	☺	?	☺	☺	☺
Nieminen *et al*[15], 2002	☺	☺	?	?	☺	☺	☺
de Veld *et al*[16], 2004	?	☺	☺	☹	☺	☺	☺
Dreiseitl *et al*[9], 2009	☺	☺	☺	☺	☺	☺	☺
Lucidarme *et al*[17], 2010	?	☺	☺	☺	?	☺	☺
Fink *et al*[18], 2017	☹	?	☺	?	?	☺	☺
Mori *et al*[8], 2018	☺	☺	☺	☺	☺	☺	☺
Walker *et al*[19], 2019	☺	☺	☺	?	☺	☺	☺
Wang *et al*[20], 2019	☺	☺	☺	☺	☺	☺	☺
Su *et al*[21], 2019	☺	☺	☺	?	☺	☺	☺
Li *et al*[22], 2019	?	☺	☺	☺	☺	☺	☺
Hollon *et al*[23], 2020	☺	☺	☺	?	☺	☺	☺
Wang *et al*[24], 2020	☺	☺	?	☺	☺	☺	☺
Repici *et al*[25], 2020	☺	☺	☺	?	☺	☺	☺
Gong *et al*[26], 2020	☺	☺	☺	?	☺	☺	☺
Wang *et al*[27], 2020	☺	☺	☺	?	☺	☺	☺
Liu *et al*[28], 2020	☺	☺	☺	?	☺	☺	☺

☺ low risk; ☹ high risk; ? unclear risk, QUADAS-2, quality assessment of diagnostic accuracy studies - 2

prominent role in all fields of medicine including oncology as a result of developments in incrementally accelerating computational capabilities, availability of open-source software and accumulation of large standardized global patient datasets sourced from medical records and wearable health monitors[3]. The success rates of AI/ML protocols in terms of accuracy of diagnosis have significantly improved as reported in validation studies. The overwhelming success of AI systems notwithstanding, it seems unlikely that AI-based systems will completely replace physicians, at least in the near future[33]. However, as seen in this review, AI has the ability to augment the efficiency of physicians and improve health outcomes[1].

In this review, 15/18 studies compared diagnosis by AI/ML protocol with human diagnosis, and the emerging finding from all of these studies was that the AI/ML protocol is able to improve the human diagnosis, especially that made by the less experienced clinician. Notably, none of the studies reported that an AI/ML protocol performing poorly as compared humans. Shen *et al*[6] reported findings similar to our observation in their systematic review of nine studies pertaining to AI/ML protocols in various fields of medicine, observing that the performance of AI was at par with that of clinicians and exceeded that of clinicians with less experience, especially in image recognition-related fields. This suggests that AI/ML protocols have a potential to significantly improve upon the prevailing diagnostic capabilities.

In the present study, although there was an initial pool of 960 studies, only 18 studies (1.88%) which actually used an AI/ML protocol for diagnosis were eligible for inclusion. Most of the excluded studies focused on developing, testing and validating an AI/ML protocol in cancer diagnosis, whereas, our search strategy focused on the step after validation *i.e.*, the actual usage of the protocol. The low number of studies that were found to be eligible for review is an indirect indicator of the current trend of AI/ML in cancer diagnosis. Despite there being an impressive volume of research being carried out in this field, the translation of the research findings into actual clinical use still remains unsatisfactory. Possible reasons for

Table II. Characteristics of included studies

Study	Type of study	1st author country	Study site (Country)	Cancer studied	Type of lesions studied	AI/ML protocol	Name of the device/ technology	No of patients	Male, n (%)	Female, n (%)	Mean Age (years)	No of lesions studied
Kok *et al*[13], 1996	Prospective, observational	Netherlands	Netherlands	Cervical Cancer (Screening)	Cervical smear	ANN-based DS tool	PAPNET	91,294	0	91,294 (100)	NA	91,294
Chang *et al*[14], 1999	Prospective, observational	Taiwan	Taiwan	Prostate cancer	Multiple parameters	Multifactorial DS system	PCES	43	43 (100)	0	67	43
Nieminen *et al*[15], 2002	Randomized, Prospective, observational	Finland	Finland	Cervical Cancer (Screening)	Cervical smear	ANN-based DS tool	PAPNET	108,686	0	108,686 (100)	44±10.3	108,686
de Veld *et al*[16], 2004	Prospective, observational	Netherlands	Netherlands	Cancer of Oral Cavity	Oral mucosal lesion	PCA; ANN	Autofluorescence spectroscopy	155	NA	NA	57±1	176
Dreiseitl *et al*[9], 2009	Prospective, observational	Austria	Austria	Skin cancer	PSL	ANN-based DS tool	MoleMax II instrument with added decision support system	458	NA	NA	NA	3,021
Lucidarme *et al*[17], 2010	Prospective, observational	France	Multiple*	Ovarian cancer	TVS image of ovary	Not specified	OVHS	264	0	264 (100)	57 (Median)	375
Fink *et al*[18], 2017	Prospective, observational	Germany	Germany	Skin cancer	PSL	Not specified	MelaFind device	111	59 (53.2)	52 (46.8)	45±17.3	346
Mori *et al*[8], 2018	Prospective, observational	Japan	Japan	Colorectal cancer	Colorectal Polyps	Machine learning, SVM	Real-time automatic polyp detection system	325	235 (72.3)	90 (27.7)	67 (Median)	466
Walker *et al*[19], 2019	Prospective, observational	USA	Israel	Skin cancer	PSL	CNN, Deep learning	NA	63	34 (54.0)	29 (46.0)	50.4±14.9	63
Wang *et al*[20], 2019	Randomized, Prospective, observational	China	China	Colorectal cancer	Colorectal Polyps	Deep learning architecture	Real-time automatic polyp detection system	1,058	512 (48.4)	546 (51.6)	49.9±13.8	767
Su *et al*[21], 2019	Randomized, Prospective, observational	China	China	Colorectal cancer	Colorectal polyps	CNN, Deep learning	AQCS-aided colonoscopy	623	307 (49.3)	316 (50.7)	NA	442
Li *et al*[22], 2019	Prospective, observational	China	China	Lung cancer	Lung nodules	CNN, Deep learning	DL-CAD	346	221 (63.9)	125 (36.1)	51.0±10.2	1916
Hollon *et al*[23], 2020	Prospective, observational	USA	USA	Brain cancer	Intraoperative surgical specimen	CNN, Deep learning	NA	278	NA	NA	NA	278
Wang *et al*[24], 2020	Randomized, Prospective, observational	China	China	Colorectal cancer	Colorectal polyps	Deep learning	CADe colonoscopy system	369	179 (48.5)	190 (51.5)	NA	811

Contd...

Study, Year	Type of study	1st author country	Study site (Country)	Cancer studied	Type of lesions studied	AI/ML protocol	Name of the device/ technology	No of patients	Male, n (%)	Female, n (%)	Mean Age (years)	No of lesions studied
Repici *et al*[25], 2020	Randomized, Prospective, observational	Italy	Italy	Colorectal cancer	Colorectal polyps	CNN, Deep learning	CADe colonoscopy system	685	337 (49.2)	348 (50.8)	61.3±10.2	493
Gong *et al*[26], 2020	Randomized, Prospective, observational	China	China	Colorectal cancer	Colorectal polyps	CNN, Deep learning	ENDOANGEL-assisted colonoscopy	704	345 (49.0)	359 (51.0)	NA	369
Wang *et al*[27], 2020	Randomized, Prospective, observational	China	China	Colorectal cancer	Colorectal polyps	Deep learning	CADe colonoscopy system	962	495 (51.5)	467 (48.5)	NA	809
Liu *et al*[28], 2020	Randomized, Prospective, observational	China	China	Colorectal cancer	Colorectal polyps	CNN, Deep learning	CADe colonoscopy system	1026	551 (53.7)	475 (46.3)	NA	734

*Patients were recruited in five countries: France, Sweden, Italy, Germany, and Israel. AI/ML: artificial intelligence/machine learning; ANN: artificial neural network; AQCS: automatic quality control system; CADe: computer-aided detection; CNN: convoluted neural network; DL-CAD: Deep-learning based computer-aided diagnosis; DS: decision support; OVHS: ovarian histoscanning; PCA: principal component analysis; PCES: prostate cancer expert system; PSL: pigmented skin lesion; SVM: support vector machine; TVS: transvaginal scan. ENDOANGEL is a proprietary name

this observation may include a disconnect between the developers of the AI/ML technology and the eventual users of the same, mistrust among the clinicians pertaining to the AI/ML technology owing to fear of being replaced, and lack of credible evidence for the success of AI/ML in the real world, as opposed to demonstrating success using retrospective datasets of patient images. This finding also highlights a need on part of the innovators to explore and develop methods to incorporate validated AI/ML technology into routine practice. Experts in AI/ML protocol development and clinicians need to collaborate to get the innovations in AI/ML forward and closer to the patient so that there is a better translation of the advancements in various fields and improved patient outcome[3]. The need for maintaining adequate quality control in AI/ML-based diagnosis is also essential to ensure that patient care is not compromised: this is another area towards which collaboration is required between developers and users[34]. Further, incorporation of the AI/ML protocol into routine diagnosis brings in complexities such as issues pertaining to storage of images, workforce training, data security and privacy issues, and legal aspects, all of which may enhance the cost of diagnosis, which in turn may be passed on to the patient. Thus, a cost-effectiveness study of AI/ML technologies is essential before adoption of the same into routine diagnosis.

Most studies included in this review had developed AI/ML protocols based on ANN and DL technologies similar to a previous systematic review conducted on the comparison of AI in diagnosis of diseases from across specializations in which CNN (a form of DL), was the most frequent technology used[6]. Further, one study published in 1999 reported the usage of an expert system which is an older generation AI technology[14]; two studies did not specify which AI/ML technology was used in the device being tested[17,18], possibly for commercial and patent reasons. On the other hand, one study had specified that the software used for AI/ML protocol development was available for free from the authors[9].

While most of the diagnostic protocols observed in the present study provided a near real-time diagnosis, the time required for alternative diagnosis (in the form of biopsy and histology) is considerably longer. This feature of AI/ML diagnostic protocols might improve patient outcomes in different ways, for example, facilitate better surgical decision-making in brain tumours[23], identify which polyps observed during colonoscopy may be safe to 'diagnose and leave'[8],

Table III. Diagnostic performance of AI/ML protocols in the included studies

Study	Sensitivity of the AI/ML protocol (%)	Specificity of the AI/ML protocol (%)	Accuracy of the AI/ML protocol (%)	PPV of the AI/ML protocol (%)	NPV of the AI/ML protocol (%)	Performance of AI/ML diagnosis as compared to human diagnosis
Kok *et al*[13], 1996	NA	NA	NA	NA	NA	AI is similar to human diagnosis
Chang *et al*[14], 1999	92	84	88.4	NA	NA	AI improves human diagnosis
Nieminen *et al*[15], 2002	NA	92.5	NA	55	NA	AI is similar to human diagnosis
de Veld *et al*[16], 2004	NA	NA	NA	NA	NA	Comparison not performed
Dreiseitl *et al*[9], 2009	72	82	NA	NA	NA	Depends on the user's background
Lucidarme *et al*[17], 2010	98	88	NA	NA	NA	AI improves human diagnosis
Fink *et al*[18], 2017	100	68.5	2.3	2.80	100	Comparison not performed
Mori *et al*[8], 2018	NA	NA	98.1	NA	93.7 to 96.5	AI is better than human diagnosis
Walker *et al*[19], 2019	86 (system B); 91 (system A)[1]	69 (system B)[1]	NA	88.9	88.9	Comparison not performed
Wang *et al*[20], 2019	NA	NA	NA	NA	NA	AI improves human diagnosis
Su *et al*[21], 2019	NA	NA	NA	NA	NA	AI improves human diagnosis
Li *et al*[22], 2019	86.2	NA	NA	57.0	NA	AI is better than human diagnosis
Hollon *et al*[23], 2020	NA	NA	94.6	NA	NA	AI is better than human diagnosis
Wang *et al*[24], 2020	NA	NA	NA	NA	NA	AI is better than human diagnosis
Repici *et al*[25], 2020	NA	NA	NA	NA	NA	AI improves human diagnosis
Gong *et al*[26], 2020	NA	NA	NA	NA	NA	AI improves human diagnosis
Wang *et al*[27], 2020	NA	NA	NA	NA	NA	AI improves human diagnosis
Liu *et al*[28], 2020	NA	NA	NA	NA	NA	AI improves human diagnosis

[1]System A is a deep learning classifier whose outputs from image processing of pigmented skin lesions were converted into sound waves, which were once again classified by system B. PPV: positive predictive value; NPV: negative predictive value

identify which pigmented skin lesions are to be excised and which are safer to be left behind[18,19] and ease the workload of human cytotechnologists[13,15]. Shortage of qualified medical personnel is a globally observed phenomenon, and it is possible that promoting research and uptake of AI/ML in diagnostics might contribute towards solving this problem in a novel way by reducing the workload on medical personnel.

The *post hoc* analysis in our study brought out some interesting findings. First, although a large number of validation studies are published, the number of published studies reporting the actual usage of AI/ML protocols is relatively small (223 *vs.* 18). Second, despite breast cancer being the most frequently studied cancer in the validation studies (63 studies, 28.3%), none of the 18 included studies in the main review

described the actual usage of any AI/ML protocol in breast cancer diagnosis. It might be possible that many instances of actual usage are presented in conferences and workshops but are not published in peer-reviewed journals indexed in PubMed which was the source of literature in the current study. Further, the contribution of publication bias, wherein only positive results of actual usage are published, should also be considered. A stronger collaboration between AI/ML scientists (who develop the AI/ML protocols) and clinicians and diagnosticians (who actually use these protocols for diagnosis) might help in reducing this gap. There is also a possibility of reluctance on the part of clinicians to accept technological solutions for diagnosis due to the fear of 'replacement' by technology, which needs to be mitigated appropriately by all relevant stakeholders. The notion that 'technology is superior to human diagnosis' has often alienated clinicians; we believe the optimal approach for improving collaboration between these two stakeholders would be that 'technology will augment human diagnosis'[1]. Third, the dominance of validation studies published from the USA and China indicates the trend of research importance given to the field of AI/ML in these countries. Notably, four validation studies are reported from India[29-32], but it does not appear likely that any of these studies were followed up with publications of actual usage of the AI/ML protocols. In this background, considering that healthcare needs are often region specific, collaborations between AI/ML specialists from India with clinicians and diagnosticians practicing in India are essential to develop solutions that are meaningful and impactful for use in India.

Finally, we observed that there is a huge variation in the number of items (patients/lesions/images) used for validation of AI/ML protocols. This indicates that there is lack of proper regulatory framework for conducting these validation studies and a lack of regulatory standards to assess the safety and efficacy of AI systems has been highlighted in the past as well[35,36]. In this background, the results of validation studies claiming that an AI/ML protocol's 'incredible success surpassing human capabilities' should be viewed with caution. Realising the need for framing guidelines specific for the usage of AI/ML in healthcare, the USFDA published a white paper during April 2019 describing the USFDA's position for premarket review of software protocols based on AI/ML and feedbacks were invited from concerned stakeholders[37]. The proposed regulatory framework is expected to enable the USFDA and potential device manufacturers to evaluate and monitor a software product from its development to post-market performance[37].

During the course of this review, we found out that despite promising performance in validation tests as well as in prospective settings, and despite being approved for commercial usage by the USFDA[18], the MelaFind device which was a non-invasive tool for melanoma diagnosis was discontinued in 2017 because of poor commercial performance[38]. This incident highlights the importance of performing a proper economic evaluation in addition to efficacy and accuracy testing, to prevent a similar fate for other promising AI/ML protocols[20]. In addition, payers and governments should identify and incentivize the development and marketing of such promising AI/ML protocols with a view of improving patient outcomes.

Our study was not without limitations. Firstly, our search remained restricted to only PubMed. Since this study was not funded by any source, the usage of Embase (which is a paid resource) was not possible. Secondly, all non-English articles were exlcuded. Thus, relevant studies published in other languages and studies not indexed in PubMed might have been missed. We also excluded conference abstracts from our search according to our pre-specified inclusion and exclusion criteria; by doing so, we might have missed unpublished studies presented in conferences that may have reported actual usage of AI/ML in routine practice.

To conclude, this scoping review found that only a handful of studies report the actual usage of AI/ML protocols in a prospective manner for cancer diagnosis. In most of the studies, the AI/ML protocol performed at par or even better than humans, which indicates that the actual usage of AI/ML protocols may enhance the accuracy of human diagnosis resulting in better patient outcomes. Most of the existing protocols depend upon ANN and DL. While many studies describe validation and testing of AI/ML protocols, further research is required to identify methods to actually incorporate these novel technologies in routine clinical practice. Regulatory frameworks specific for AI/ML protocols in medical usage should be developed and implemented to properly evaluate the development and performance of devices utilizing such protocols for healthcare delivery.

***Financial support & sponsorship*:** None.

***Conflicts of Interest*:** None.

References

1. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* 2019; *7* : e7702.
2. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. *Prog Retin Eye Res* 2018; *67* : 1-29.
3. Azuaje F. Artificial intelligence for precision oncology: Beyond patient stratification. *NPJ Precis Oncol* 2019; *3* : 6.
4. Zhang Z. A gentle introduction to artificial neural networks. *Ann Transl Med* 2016; *4* : 370.
5. Kann BH, Thompson R, Thomas CR Jr., Dicker A, Aneja S. Artificial intelligence in oncology: Current applications and future directions. *Oncology (Williston Park)* 2019; *33* : 46-53.
6. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, *et al.* Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Med Inform* 2019; *7* : e10010.
7. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, *et al.* International evaluation of an AI system for breast cancer screening. *Nature* 2020; *577* : 89-94.
8. Mori Y, Kudo SE, Misawa M, Saito Y, Ikematsu H, Hotta K, *et al.* Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: A prospective study. *Ann Intern Med* 2018; *169* : 357-66.
9. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: Evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. *Melanoma Res* 2009; *19* : 180-4.
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, *et al.* PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann Intern Med* 2018; *169* : 467-73.
11. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, *et al.* QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; *155* : 529-36.
12. Dang A, Chidirala S, Veeranki P, Vallish BN. A critical overview of systematic reviews of chemotherapy for advanced and locally advanced pancreatic cancer using both AMSTAR2 and ROBIS as quality assessment tools. *Rev Recent Clin Trials* 2021; *16* : 180-92.
13. Kok MR, Boon ME. Consequences of neural network technology for cervical screening: Increase in diagnostic consistency and positive scores. *Cancer* 1996; *78* : 112-7.
14. Chang PL, Li YC, Wang TM, Huang ST, Hsieh ML, Tsui KH. Evaluation of a decision-support system for preoperative staging of prostate cancer. *Med Decis Making* 1999; *19* : 419-27.
15. Nieminen P, Hakama M, Viikki M, Tarkkanen J, Anttila A. Prospective and randomised public-health trial on neural network-assisted screening for cervical cancer in Finland: Results of the first year. *Int J Cancer* 2003; *103* : 422-6.
16. de Veld DC, Skurichina M, Witjes MJ, Duin RP, Sterenborg HJ, Roodenburg JL. Clinical study for classification of benign, dysplastic, and malignant oral lesions using autofluorescence spectroscopy. *J Biomed Opt* 2004; *9* : 940-50.
17. Lucidarme O, Akakpo JP, Granberg S, Sideri M, Levavi H, Schneider A, *et al.* A new computer-aided diagnostic tool for non-invasive characterisation of malignant ovarian masses: Results of a multicentre validation study. *Eur Radiol* 2010; *20* : 1822-30.
18. Fink C, Jaeger C, Jaeger K, Haenssle HA. Diagnostic performance of the MelaFind device in a real-life clinical setting. *J Dtsch Dermatol Ges* 2017; *15* : 414-9.
19. Walker BN, Rehg JM, Kalra A, Winters RM, Drews P, Dascalu J, *et al.* Dermoscopy diagnosis of cancerous lesions utilizing dual deep learning algorithms via visual and audio (sonification) outputs: Laboratory and prospective observational studies. *EBioMedicine* 2019; *40* : 176-83.
20. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, *et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut* 2019; *68* : 1813-9.
21. Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, *et al.* Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: A prospective randomized controlled study (with videos). *Gastrointest Endosc* 2020; *91* : 415-24.
22. Li L, Liu Z, Huang H, Lin M, Luo D. Evaluating the performance of a deep learning-based computer-aided diagnosis (DL-CAD) system for detecting and characterizing lung nodules: Comparison with the performance of double reading by radiologists. *Thorac Cancer* 2019; *10* : 183-92.
23. Hollon TC, Pandian B, Adapa AR, Urias E, Save AV, Khalsa SSS, *et al.* Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med* 2020; *26* : 52-8.
24. Wang P, Liu P, Glissen Brown JR, Berzin TM, Zhou G, Lei S, *et al.* Lower adenoma miss rate of computer-aided detection-assisted colonoscopy vs. routine white-light colonoscopy in a prospective tandem study. *Gastroenterology* 2020; *159* : 1252-61.e5.
25. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, *et al.* Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. *Gastroenterology* 2020; *159* : 512-20.e7.
26. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, *et al.* Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): A randomised controlled study. *Lancet Gastroenterol Hepatol* 2020; *5* : 352-61.
27. Wang P, Liu X, Berzin TM, Glissen Brown JR, Liu P, Zhou C, *et al.* Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): A double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020; *5* : 343-51.

28. Liu WN, Zhang YY, Bian XQ, Wang LJ, Yang Q, Zhang XD, *et al.* Study on detection rate of polyps and adenomas in artificial-intelligence-aided colonoscopy. *Saudi J Gastroenterol* 2020; *26* : 13-9.
29. Krishnan MM, Acharya UR, Chakraborty C, Ray AK. Automated diagnosis of oral cancer using higher order spectra features and local binary pattern: A comparative study. *Technol Cancer Res Treat* 2011; *10* : 443-55.
30. Virmani J, Kumar V, Kalra N, Khandelwal N. Characterization of primary and secondary malignant liver lesions from B-mode ultrasound. *J Digit Imaging* 2013; *26* : 1058-70.
31. Bahadure NB, Ray AK, Thethi HP. Comparative approach of MRI-based brain tumor segmentation and classification using genetic algorithm. *J Digit Imaging* 2018; *31* : 477-89.
32. Saikia AR, Bora K, Mahanta LB, Das AK. Comparative assessment of CNN architectures for classification of breast FNAC images. *Tissue Cell* 2019; *57* : 8-14.
33. Karches KE. Against the iDoctor: Why artificial intelligence should not replace physician judgment. *Theor Med Bioeth* 2018; *39* : 91-110.
34. Mahadevaiah G, Rv P, Bermejo I, Jaffray D, Dekker A, Wee L. Artificial intelligence-based clinical decision support in modern medical physics: Selection, acceptance, commissioning, and quality assurance. *Med Phys* 2020; *47* : e228-35.
35. Johner Institute. *Artificial Institute in Medicine [Monograph on the internet]*. Available from: *https://www.johner-institute.com/articles/software-iec-62304/artificial-intelligence/*, updated on February 2, 2019; accessed on May 28, 2021.
36. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc Neurol* 2017; *2* : 230-43.
37. US Food and Drug Administration. *Artificial intelligence and machine learning in software as a medical device*. Available from: *https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device*, accessed on May 28, 2021.
38. Seeking Alpha. *STRATA Skin Sciences, Inc. quarterly report*. Available from: *https://seekingalpha.com/filing/3550049,* accessed on May 28, 2021.

For correspondence: Dr Amit Dang, MarksMan Healthcare Communications, J1309, Amethyst Tower, PBEL City, Peeramcheruvu Village, Rajendra Nagar Mandal, Hyderabad 500 091, Telangana, India
e-mail: amit.d@marksmanhealthcare.com

Supplementary Table I. Search strategy

Search	Query	Results	Remarks
#1	Search: ((("artificial intelligence"[MeSH Terms]) OR ("machine learning"[MeSH Terms])) OR (artificial intelligence [Title/Abstract])) OR (machine learning[Title/Abstract])	144,127	All types of articles dealing with artificial intelligence and/or machine learning
#2	Search: ("neoplasms"[MeSH Major Topic]) AND ("diagnosis"[MeSH Major Topic])	352,175	All types of articles dealing with any type of diagnosis of any type of cancer
#3	#1 AND #2	5,689	All types of articles dealing with AI/ML AND cancer diagnosis
#4	Search: ("adaptive clinical trial"[Publication Type] OR "clinical study"[Publication Type] OR "clinical trial"[Publication Type] OR "clinical trial, phase i"[Publication Type] OR "clinical trial, phase ii"[Publication Type] OR "clinical trial, phase iv"[Publication Type] OR "clinical trial, phase iii"[Publication Type] OR "comparative study"[Publication Type] OR "controlled clinical trial"[Publication Type] OR "equivalence trial"[Publication Type] OR "multicenter study"[Publication Type] OR "observational study"[Publication Type] OR "pragmatic clinical trial"[Publication Type] OR "randomized controlled trial"[Publication Type])	2,797,020	All clinical trials and related articles as on date
#5	#3 AND #4	983	Studies dealing with AI/ML AND cancer diagnosis in clinical trial and related settings
	Filters: English	951	Studies dealing with AI/ML AND cancer diagnosis in clinical trial and related settings, reported in English Language

Date of execution: 17 May 2021. AI/ML, artificial intelligence/machine learning

Supplementary Table II. Type of cancers in which artificial intelligence/machine learning protocols were validated

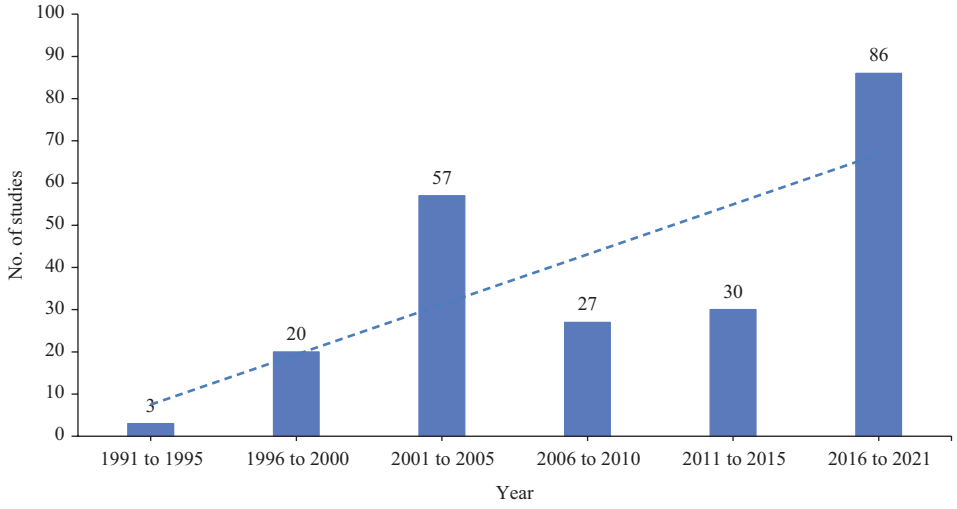
Type of cancer	Number of studies (%)	Type of cancer	Number of studies (%)
Breast	63 (28.3)	Oral	4 (1.8)
Dermatological	24 (10.8)	Bone	3 (1.3)
Prostate	22 9.9)	Pancreas	3 (1.3)
Lung cancer	19 (8.8)	Uterus	3 (1.3)
Brain	18 (8.1)	Mesothelioma	2 (0.9)
Liver	12 (5.4)	Cardiac	1 (0.4)
Haematological	11 (4.9)	Ovarian	1 (0.4)
Cervical cancer	9 (4.0)	Parathyroid	1 (0.4)
Head and neck	7 (3.1)	Renal	1 (0.4)
Colorectal cancer	6 (2.7)	Soft tissue tumours	1 (0.4)
GIT	6 (2.7)	Spine	1 (0.4)
Thyroid	5 (2.2)	Total	223 (100)

GIT, gastrointestinal tract

Supplementary Table III. Nature of artificial intelligence/machine learning protocols used during validation studies

AI/ML protocol	Number of studies	AI/ML Protocol	Number of studies	AI/ML protocol	Number of studies
ANN	74	NLP	2	Expert system	2
SVM	49	RBF	2	PCA	1
CNN	35	BoVW	1	PLSDA	1
KNN	16	BRNC	1	QDA	1
RF	14	CART	1	RF	1
LRA	12	CPDF	1	RVM	1
NB	9	DSS	1	SDA	1
DL	6	FCN	1	LRBC	1
DT	6	FDF	1	Genetic algorithm	1
LDA	6	FLD	1	MLSAM	1
BN	3	FNN	1	Mean shift clustering	1
PNN	3	KFD	1	NLDR	1
AdaBoost	2	LA	1	Seeded atlas deformation	1
CBRA	2	LR	1	Likelihood ratio-based classifier	1
DNN	2	MLP	1	FMWTA	1
FCM	2	MSA	1	Not specified	21
				Total	296

*Since 20/223 studies used more than 1 AI/ML protocols, the total number of protocols is more than the total number of studies. ANN, artificial neural network; AdaBoost, adaptive boosting; BN, Bayesian network; BoVW, bags of visual words; CART, classification and regression tree; CBRA, case-based reasoning algorithm; CNN, convoluted neural network; CPDF, compound probability density function; DL, deep learning; DNN, deep neural network; DSS: decision support system; DT, decision tree; FCM, fuzzy c-means clustering; FCN, fully convolutional network; FDF, Fractal dimension feature; FLD, Fisher linear discrimination; FMWTA, fuzzy merging and wall-thickening analysis; FNN, fuzzy neural network; KFD, kernel fisher discriminant; KNN, k-nearest neighbour; LDA, linear discriminant analysis; LRA, logistic regression analysis; LRBC, likelihood ratio-based classifier; MLP, multilayer perceptron network; MLSAM, maximum likelihood and spectral angle mapper; MSA, mean shift algorithm; NB, Naïve Bayes; NLDR, non-linear dimensionality reduction; NLP, natural language processing; PCA, principal component analysis; PLSDA, partial least square discriminant analysis; PNN, probabilistic neural network; QDA, quadratic discriminant analysis; RBF, radial basis function; RF, random forest; RVM, relevance vector machine; SDA, stepwise discriminant analysis; SVM, support vector machine; AI/ML, artificial intelligence/machine learning; BRNC, Bayesian regularization neural classifier; LA, least absolute; LR, logistic regression



Supplementary Figure. Yearly trend of publication of studies validating AI/ML protocols for cancer diagnosis.

Supplementary Table IV. Country of first author of studies validating artificial intelligence/machine learning protocols in cancer diagnosis

Country	Number of studies (%)	Country	Number of studies (%)
USA	72 (32.3)	Austria	2 (0.9)
China	26 (11.7)	Denmark	2 (0.9)
Germany	17 (7.6)	Singapore	2 (0.9)
Japan	12 (5.4)	Spain	2 (0.9)
South Korea	12 (5.4)	Belgium	1 (0.4)
Italy	10 (4.5)	Brazil	1 (0.4)
UK	9 (4.0)	Colombia	1 (0.4)
Taiwan	7 (3.1)	Czech Republic	1 (0.4)
Canada	5 (2.2)	Egypt	1 (0.4)
India	4 (1.8)	Israel	1 (0.4)
Greece	4 (1.8)	Malaysia	1 (0.4)
The Netherlands	4 (1.8)	NA	1 (0.4)
Poland	4 (1.8)	Portugal	1 (0.4)
Switzerland	4 (1.8)	Serbia	1 (0.4)
Australia	3 (1.3)	Syria	1 (0.4)
Romania	3 (1.3)	Tunisia	1 (0.4)
Sweden	3 (1.3)	Uruguay	1 (0.4)
Turkey	3 (1.3)	Total	223 (100)

Supplementary Table V. Number of items used for validation of artificial intelligence/machine learning protocols

Parameter	Patients	Lesions/images
Range	8-84,424	15-1,036,496
Median	121	487
Mean±SD	1136.74±7276.90	11,478.32±96,664.63

SD, standard deviation