BMJ Health & Care Informatics

Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review

Daraje kaba Gurmessa [1,2] Worku Jimma[1]

To cite: Gurmessa Dkaba, Jimma W. Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review. *BMJ Health Care Inform* 2024;**31**:e100954. doi:10.1136/bmjhci-2023-100954

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2023-100954).

Received 06 November 2023
Accepted 21 January 2024

[1]Department of Information Science, Jimma Institute of Technology, Jimma University, Jimma, Oromia, Ethiopia
[2]Computer Science, Mattu University, Mattu, Oromīya, Ethiopia

Correspondence to
Daraje kaba Gurmessa; darajekaba2020@gmail.com

ABSTRACT

Background Breast cancer is the most common disease in women. Recently, explainable artificial intelligence (XAI) approaches have been dedicated to investigate breast cancer. An overwhelming study has been done on XAI for breast cancer. Therefore, this study aims to review an XAI for breast cancer diagnosis from mammography and ultrasound (US) images. We investigated how XAI methods for breast cancer diagnosis have been evaluated, the existing ethical challenges, research gaps, the XAI used and the relation between the accuracy and explainability of algorithms.

Methods In this work, Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist and diagram were used. Peer-reviewed articles and conference proceedings from PubMed, IEEE Explore, ScienceDirect, Scopus and Google Scholar databases were searched. There is no stated date limit to filter the papers. The papers were searched on 19 September 2023, using various combinations of the search terms 'breast cancer', 'explainable', 'interpretable', 'machine learning', 'artificial intelligence' and 'XAI'. Rayyan online platform detected duplicates, inclusion and exclusion of papers.

Results This study identified 14 primary studies employing XAI for breast cancer diagnosis from mammography and US images. Out of the selected 14 studies, only 1 research evaluated humans' confidence in using the XAI system—additionally, 92.86% of identified papers identified dataset and dataset-related issues as research gaps and future direction. The result showed that further research and evaluation are needed to determine the most effective XAI method for breast cancer.

Conclusion XAI is not conceded to increase users' and doctors' trust in the system. For the real-world application, effective and systematic evaluation of its trustworthiness in this scenario is lacking.

PROSPERO registration number CRD42023458665.

INTRODUCTION

Breast cancer is the first and most common type of cancer in women.[1 2] Anatomically, the breast consists of healthy blood vessels, connective tissue, ductal lobules and lymph nodes.[3] Breast cancer is a problem with abnormal growth of the breast cells. By 2040, the burden of breast cancer is predicted to increase to over three million new cases and one million deaths every year because of population growth and ageing alone.[2]

Breast cancer is highly treatable if identified at an early stage, and hence, early detection is crucial to save lives. Among the methods of breast cancer detection, the most popular are ultrasound (US),[4] mammography[5] and MRI. However, traditional computer-aided design systems generally depend on manually created features and experience of the physiologist, therefore weakening the overall performance of breast cancer identification. Therefore, artificial intelligence (AI) methods like machine learning and deep learning-based techniques have emerged for breast cancer diagnosis with high accuracy. Additionally, improved breast cancer classification by combining graph convolutional network and convolutional neural network[6] and abnormal breast identification by a nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling are used to support patients and doctors' decisions.[7] However, the algorithms lack ethical AI, right of explanation and trustworthy AI. These concepts are considered critical issues by high-level political and technical bodies (eg, G20, EU expert groups, Association of Computing Machinery in the USA).[8 9]

Additionally, AI algorithms like machine learning and deep learning are vulnerable to bad stuff (bad decisions, bad medical diagnosis and bad prediction) is the most common drawback of AI algorithms today. They are also black box for predictive interpretation.

To overcome this issue, the science of explainable AI (XAI) has grown exponentially with its successful application in breast cancer diagnosis. However, it still requires a comprehensive review of existing studies to help researchers and practitioners gain insight and understanding of the field. Therefore, his systematic review is conducted.

Table 1 Search term combination

			AND (&&)			OR (||) or AND (&&)	
OR (||)	Term 1		Term 2	Term 3	Term 4	Term 5	Term 6
	Explainable	AND (&&)	Machine Learning	Breast	Cancer	Mammography	Ultrasound
			Artificial Intelligence				
	Interpretable		Deep learning				
			AI				
	XAI						

XAI is the extent to which people can easily understand the model. It has received much attention over the past few years. The purpose of a model explanation is to clarify why the model makes a certain prediction, to increase confidence in the model's predictions[10] and to describe exactly how a machine learning model achieves its properties.[11] Therefore, using machine learning explanations can increase the transparency, interpretability, fairness, robustness, privacy, trust and reliability of machine learning models. Recently, various methods have been proposed and used to improve the interpretation of machine learning models.

There are different taxonomies for machine learning explainability. An interactive explanation allows consumers to drill down or ask for different types of explanations until they are satisfied, while a static explanation refers to one that does not change in response to feedback from the consumer.[12] A local explanation is for a single prediction, whereas a global explanation describes the behaviour of the entire model. A directly interpretable model is one that by its intrinsic transparent nature is understandable by most consumers, whereas a post hoc explanation involves an auxiliary method to explain a model after it has been trained.[13] Self-explaining may not necessarily be a directly interpretable model. By itself, it generates local explanations. A surrogate model is usually a directly interpretable model that approximates a more complex model, while visualisation of a model may focus on parts of it and is not itself a full-fledged model.

No single method is always the best for interpreting machine learning.[12] For this reason, it is necessary to have the skills and equipment to fill the gap from research to practice. To do so, XAI toolkits like AIX360,[12] Alibi,[14] Skater,[15] H2O,[16 17] InterpretML,[18 19] EthicalML-XAI,[19 20] DALEX,[21 22] tf-explain,[23] Investigate.[24] Most interpretations and explanations are post hoc (local interpretable model-agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP). LIME and SHAP are broadly used explanation types for machine learning models from physical examination datasets. But these made explanations with limited meaning as they lacked fidelity and transparency. However, deep learning and ensemble gradients are preferable in performance for image processing and computer vision. This research is processing mammography and US images. Therefore, deep learning is recommended for breast cancer image processing.

Ensemble gradients are used to interpret deep neural networks,[11] GradientSHAP is a sample interpretation algorithm that approximates SHAP values.[25] Occlusion methods are most useful in situations such as image processing. Biological nurturing(BN) is ideal for clinical decision-making and, in general, for all assessments and studies involving multiple interventions and orientations. The oriented, modified integrated gradient (OMIG) interpretability method is inspired by the integrated gradients method. Since there is no one-size-fits-all approach to learning machine explanation, it needs a comprehensive evaluation of published papers and tools to bridge the gap in research to practice.

The research that does not consider objective metrics for evaluating XAI may lack significance and experience controversy, especially if negative reviews are not used.[8] To avoid the issues, a study[8] suggests four metrics based on performance differences, D, between the explanation's logic and the agent's actual performance, the number of rules, *R*, outputted by the explanation, the number of features, F, used to generate the explanation, and the stability, S, of the explanation. It is believed that user studies that focus on D, R, F and S metrics in their evaluations are inherently more valid.

The main contributions of this systematic review are:

1. Investigating XAI methods popularly applied for breast cancer diagnosis.
2. Identifying the algorithm's explainability and their performance relation in breast cancer diagnosis.
3. Summarise the evaluation metrics used for breast cancer diagnosis using XAI methods.
4. Summarise existing ethical challenges that XAI overcomes in breast cancer diagnoses.
5. Analysing the research gaps and future direction for XAI for breast cancer detection.

METHODOLOGY

The methodology employed in this systematic review is devoid of any medical (either prospective or retrospective) data of patients. This study applies the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guiding principles for conducting systematic reviews.[26] PRISMA 2020 was adopted because

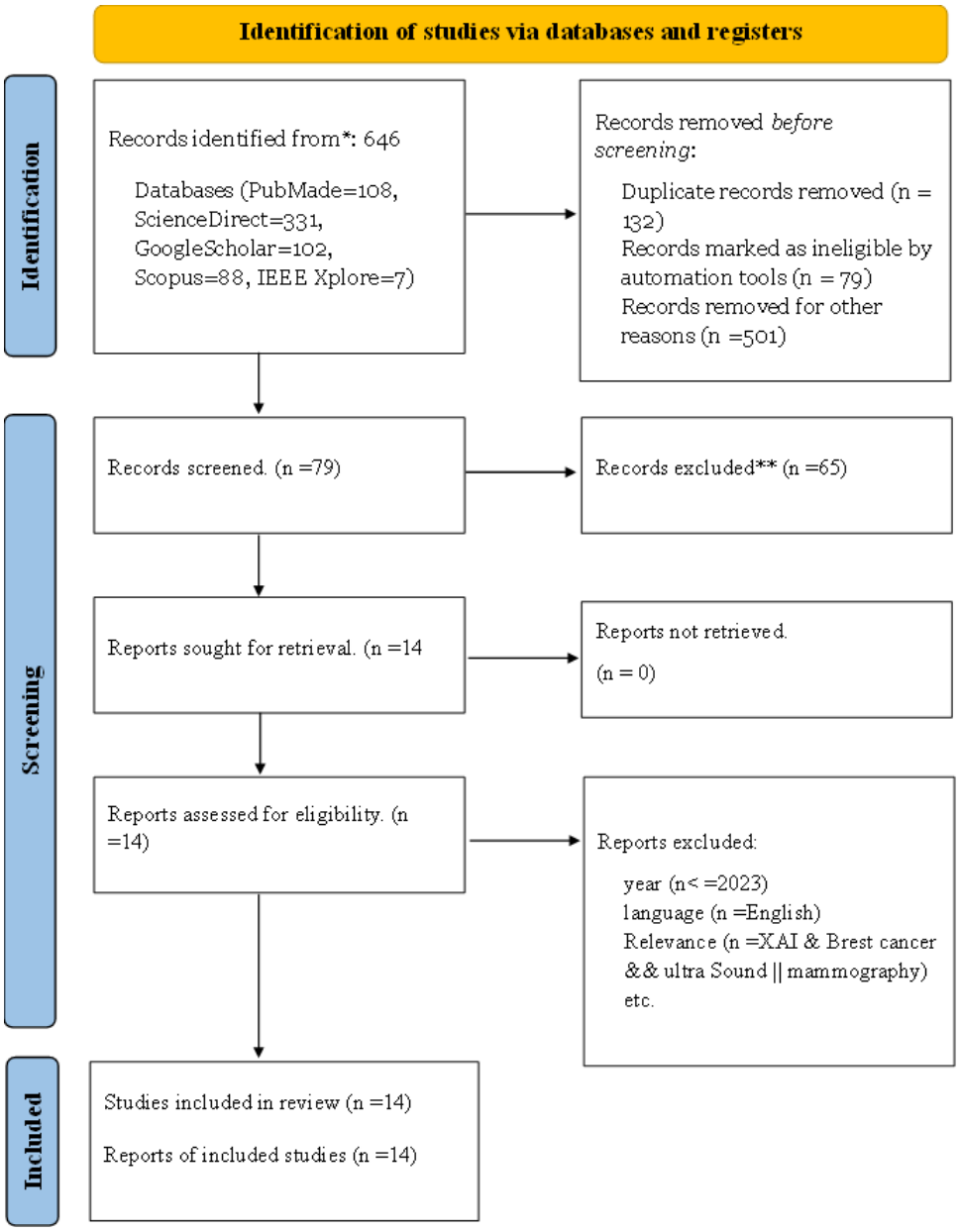

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart of explainable artificial intelligence (XAI) for breast cancer diagnosis.

of the clear guidelines it offers to ease robust systematic reviews. Therefore, this review article follows the recommendations of the guidelines. There is no stated date limit to filter the papers. The papers were searched on 19 September 2023. Peer-reviewed manuscripts and conference proceedings from PubMed, IEEE Explore, ScienceDirect, Scopus and Google Scholar databases published were searched. Rayyan for systematic review was used for duplicate removing, inclusion and exclusion term visualisations. The systematic review protocol was registered through PROSPERO with ID CRD42023458665.[27] Preplanned subgroup analyses were detailed.

Search strategy

Five databases (PubMed, IEEE Explore, ScienceDirect, Scopus and Google Scholar) were searched systemically on 19 September 2023. There is no stated date limit to filter the papers. The terms and logical operations are combined and arranged as per tables 1 and 2.

Inclusion and exclusion criteria

After applying the search equation, the criteria for inclusion and exclusion are as follows:

- Literature or systematic review articles were excluded.
- All articles focusing specifically on using XAI and strategies for breast cancer diagnosis using US, mammography or both (practical or theoretical) were included.
- Articles dealing with relevant technologies but, used procedures other than breast cancer diagnosis using US, mammography, or both were excluded, even if these systems were mentioned elsewhere in the article.

- Articles published in languages other than English were excluded.
- Articles by year of publication were not excluded, given the novelty of using XAI for breast cancer diagnosis using US mammography or both.

Study selection

The selection process of the articles was conducted based on the inclusion and exclusion criteria defined (figure 1). A bibliography of 646 papers was extracted from databases (PubMed=118, ScienceDirect=331, Scopus=88, Google Scholar=102 and IEEE Xplore=7). All the extracted papers were imported into the Rayyan online platform for systematic review. In total, 132 articles were found to be duplicates and were deleted. Moreover, 501 papers were excluded (systematic review, scoping review, breast cancer diagnosis without explainable AI and explainable AI without breast cancer diagnosis). In total, 79 papers with XAI for breast cancer terms were retained. Their full documents were downloaded and reviewed. From these, 65 papers with XAI for breast cancer without mammography or US terms were excluded again. Finally, 14 studies with XAI for breast cancer and mammography or US or both terms were included and used for this systematic review.

Risk of bias (quality) assessment method

Quality and risk of bias are assessed using Risk of Bias Visualization assessment tool in a systematic review assessment tool.[28] The tool creates traffic light plots of the domain-level judgments for each result and weighted bar plots of the distribution of risk-of-bias judgments within each bias domain.[28]

RESULTS AND DISCUSSION

Results

A total of 646 papers were extracted using search queries and terms defined in tables 1 and 2 from the selected databases. From a total of 646 papers, 134 were duplicates and removed. As depicted in figure 1, based on inclusion and exclusion stated in section Inclusion and exclusion criteria above, 79 papers (14%) with XAI for breast cancer were included (figure 1). Figure 2 depicts the included and excluded ratios. All screenshots added to these results are taken from Rayyan for a systematic review online platform.

US and mammography are the most recommended methods for breast cancer diagnosis. From 79 included papers based on XAI for breast cancer, 14 papers with XAI for breast cancer and mammography or US or both terms were either included or excluded based on criteria set in section Inclusion and exclusion criteria above. So, table 3 presents that 64.29% (9 papers from included 14) of papers were on US images, whereas 35.71% (5 papers from included 14) of papers were on mammography images.

Figure 2 shows that 97% were excluded and 3% were included based on inclusion criteria. Table 3 shows that 100% of the included papers visualised are XAI for breast cancer from mammography, US or both. It shows that 50% of them used heatmaps for visualisation.

The main objective of XAI is to encounter ethical challenges and to increase doctors' and patients' thrust on XAI. Different XAI are used for breast cancer. However, only one paper compared doctors' trust in the system.

In most of the papers, 50% (7 from 14 papers) used heatmaps for visualisation of areas of interest[29–35] and.[36] Additionally, Zhang *et al*[37] used BI-RADS-Net, Zhang *et al*[38] and Shen *et al*[35] used a saliency map, Ortega-Martorell *et al*[39] used uniform manifold approximation and projection (UMAP), Mital and Nguyen[40] used a tornado diagram, Rezazadeh *et al*[41] used histogram and Rezazade Mehrizii *et al*[34] used class activation map (CAM)-based heatmaps.

Shen *et al*'s study[35] used the largest number of datasets when compared with included studies. The study proves that the artificial intelligence system reduces false-positive findings in the interpretation of breast US examinations.[35] Breast cancer is most common in women, based on evidence on the ground in all of the studies most of the data are from women. This implies the ground truth. However, most of the datasets are taken from women and do not keep the existence of breast cancers in the ratio from man to women.

Table 2 Search equations

No	Database	Query	Number of papers
1	ScienceDirect	((('Explainable' ||'Interpretable') && ('AI' ||'Artificial Intelligence' ||'machine learning' ||'deep Learning') && 'Breast Cancer'))	331
2	PubMed	((((Breast Cancer) AND (((Explainable) OR (Explainable))) AND ((AI) OR (Artificial Intelligence) OR (machine learning) OR (deep Learning))	118
3	IEEE Explore	Explainable Machine Learning for Breast Cancer Diagnosis from Mammography and Ultrasound Images	7
4	Scopus	((('Explainable' ||'Interpretable') && ('AI' ||'Artificial Intelligence' ||'machine learning' ||'deep Learning') &&'Breast Cancer'))	88
5	Google Scholar	Explainable Machine Learning for Breast Cancer Diagnosis from Mammography and Ultrasound Images	102

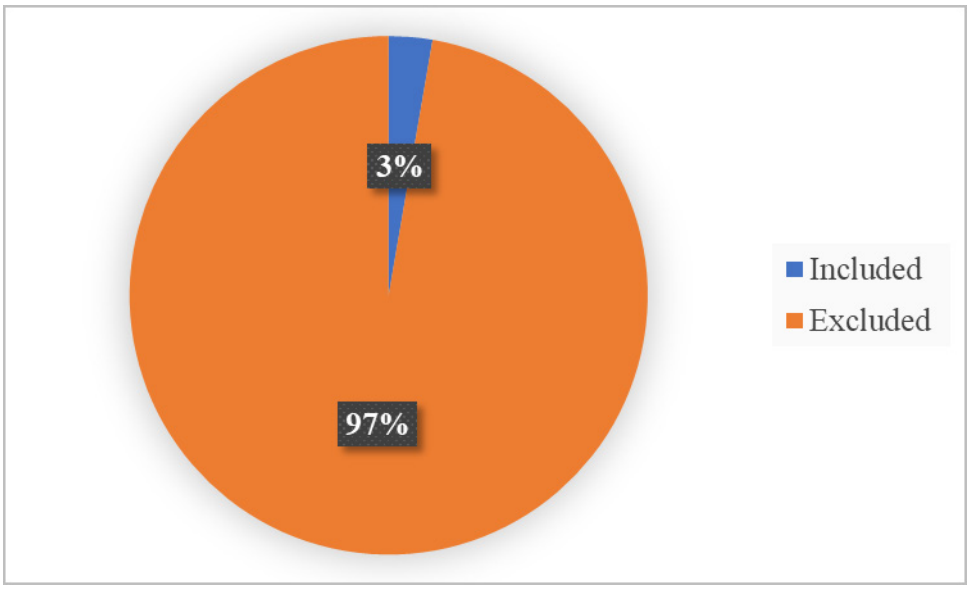

Figure 2 Included and excluded ratio graph for explainable artificial intelligence, breast cancer and mammography or ultrasound.

A total of 5 648 066 datasets are used by all included papers. From all the included papers, US-based datasets were used by 99% of studies. Mammography-based datasets used by only 1% of the total studies. For example, the maximum datasets used by Shen *et al*[35] used 5 442 907 US images, and the study by Mital and Nguyen[40] used 100 000 mammography images. This shows that there are many works left to work on improving the number of datasets on mammography images when compared with US images. We recommend that data should be collected from suspected patients with breast cancer but all the included studies said nothing about it.

Explainable/interpretable algorithms used are deep learning explanation algorithms: Of 14 papers, Explainer alone or with Grad-CAM,[29] interpretable deep learning,[30] Grad-CAM,[31] Fisher information network (FIN),[39] AI and Polygenic Risk Scores (PRS) algorithms,[40] DenseNet,[35] Explainability-partial,[34] Explainability-full,[34] VGG-16,[37] fine-tuned MobileNet-V2 convolutional neural network,[33] OMIG explainability[32] and BI-RADS-Net-V2[38] are used in 11 papers (78.57 %), SHAP[41 42] is used in 2 papers (14.3%) and LIME[36] is used in 1 paper (7.14%).

Risk of bias

The study population was known in all articles. We have obtained complete outcome variables in all articles. In all articles involved, selective reporting and publication bias were not obtained (figure 3). 'Traffic light' plots of the domain-level judgments for each result are sh0wn in figure 3.

DISCUSSION

Explainer is the situation that is explainable by itself rather than explaining black box.[29] They proved that physicians perform better when assisted by Explainer than when diagnosing alone. The study compares the use of Explainer with the post hoc technique. Based on this, they prove that Explainer can locate more reasonable and feature-related regions than the classic post hoc technique. Robustness is a characteristic expected from XAI. The study by Song *et al*[29] also tested the robustness of the proposed framework. Explainability[29] is not only related to AI performance but also to responsibility and risk in medical diagnosis. For phantom object detection,[30] accuracy and mean intersection over union were used to test the model over a total of 6369 out of 6400 objects. Finally, Oh *et al*'s study[30] concludes interpretable deep learning model using large-scale data from multiple centres shows high performance.

In the study by Qian *et al*,[31] BI-RADS scores for breast cancer were compared with experienced radiologists, areas under the receiver operating curve (ROC) and CI for multimodal images. Explanation using principal component analysis, visualisation using UMAP, FIN visualisations of the training cases and projecting the test cases onto the trained embedding.[39] the study propose a novel visualisation using FIN containing accurate information about data points' similarities that can provide intelligence about neighbouring data points.

The finding by Mital and Nguyen[40] explained AI's ability to identify high-risk women more accurately than PRS, and family history reduces the possibility of delayed breast cancer diagnosis and fewer false-positive diagnoses from not screening low-risk women.

In Sun *et al*'s study,[42] model-agnostic methods versus model-specific methods, post hoc (black box+SHAP) technique and three algorithms, namely, logistic regression, extreme gradient boosting and random forest performance, were evaluated by sensitivity, specificity and AUC.[42] This evaluation was used to evaluate the black box model only. Moreover, SHAP was used for visualising feature importance using a heatmap but it was not tested.

Table 3 Overview of reviewed articles on explainable artificial intelligence data

Reference	Number of images	Type of data	Population	Image type	Features used	XAI used	Visualisations
29	19341	Image	19341	Ultrasound (US)	Physician-annotated TI-RADS features	Explainer alone or with Grad-CAM	Heatmaps
30	2208 (training 1808, testing 400)	Images	1755	Mammography	Phantom object detection	Interpretable deep learning	Heatmaps
31	10815 (1633 lesions)	Images	775	US	Convolutional features	Grad-CAM	Heatmaps
39	2000	Images	1246 women	Mammography	CNN features	FIN	UMAP
40	100000	Images	Images	US	ICER	AI and PRS algorithms	Tornado diagram
42	11 294: 45–49 years (5709) and 50–54 years (5585)	Images	11294	Mammography	Random forest	SHAP	SHAP values
36	153	Images	153 patients (59 with metastasis and 94 without metastasis)	US	CNN features	LIME	CAM-based heatmaps
35	5442907	Images	143203	US	Coarse and fine ROIs	DenseNet	Saliency maps heatmaps
34	2760	Image	2760	Mammography	Morphological and numerical inputs	Explainability-partial Explainability-full	Heatmap and numerical attributes
37	1192 (BUSIS 562 and BUSI 630)	Image	1192	US	BI-RADS descriptors	VGG-16	BI-RADS-Net
33	624	Image	624	US	BI-RADS descriptors	Fine-tuned MobileNet-V2 convolutional neural network	Heatmaps
41	780	Image	600 female patients (age 25–75 years)	US	GLCM texture features	SHAP	Histogram
32	52800 simulated, 4800 real and 48 augmentations	Image	4800	Mammography	Phantom features	OMIG explainability	Heatmaps
38	1192 (727 benign (negative) and 465 malignant (positive))	Image	1192	US	Morphological features	BI-RADS-Net-V2	Saliency map

*BUSIS, Breast Ultrasound Image Segmentation
AI, artificial intelligence; BI-RADS, breast imaging reporting and data system; BUSI, Breast Ultrasound Image; BUSIS, Breast Ultrasound Image Segmentation; CAM, class activation map; CNN, Convolutional Neural Network; FIN, Fisher information network; GLCM, Gray-level cooccurrence matrix; ICER, incremental cost effectiveness ratio; OMIG, oriented, modified integrated gradient; ROIs, regions of interest; TI-RADS, Thyroid Imaging Reporting & Data System; UMAP, uniform manifold approximation and projection; XAI, explainable artificial intelligence.

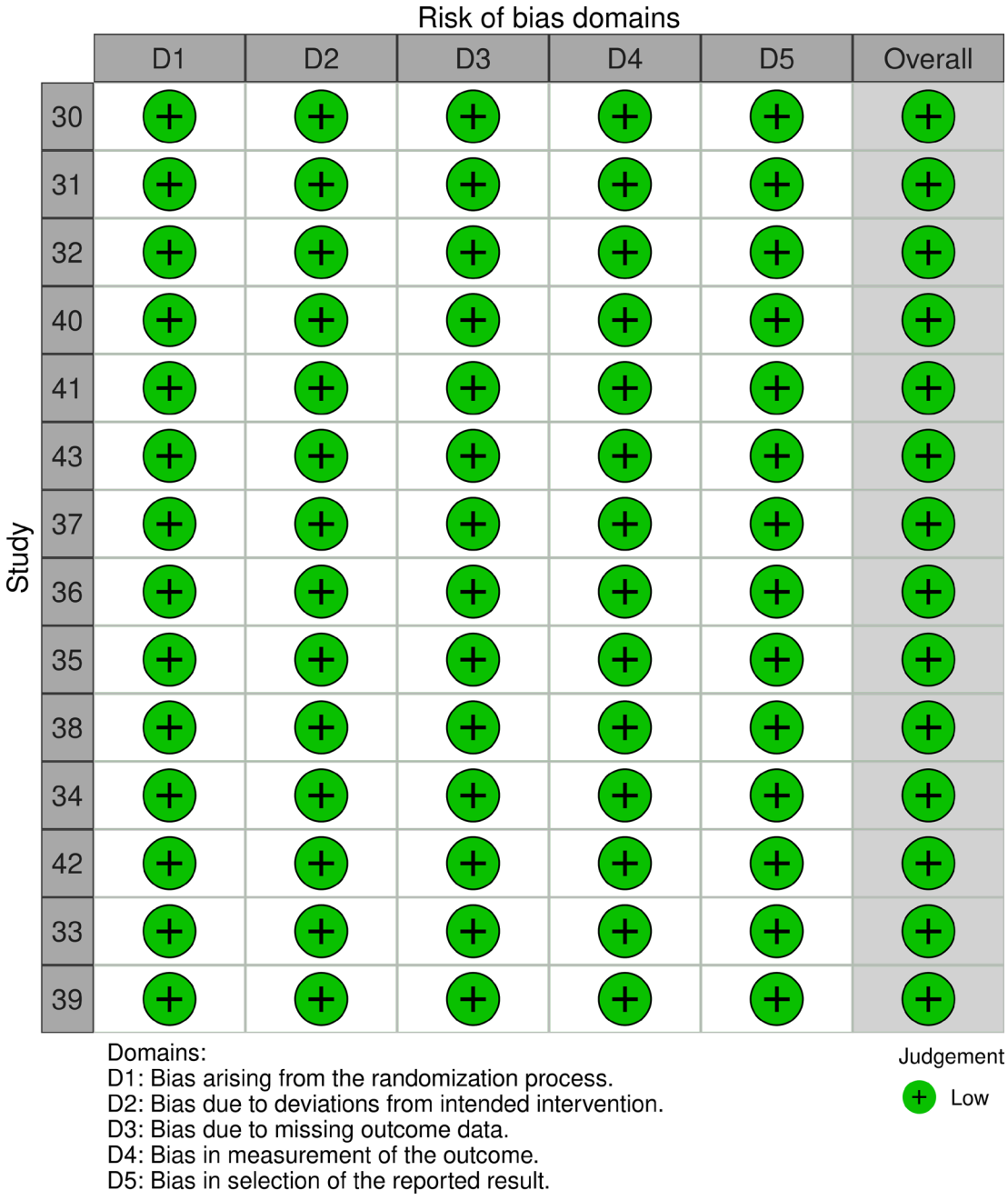

Figure 3 Traffic light plot for risk of bias.

In Lee *et al*'s study,[36] accuracy, sensitivity, specificity and AUC were used. Simple linear iterative clustering superpixel segmentation method and the LIME explanation algorithm were employed to explain how the model makes decisions.

The area under the ROC of machine learning and an average of 10 board-certified breast radiologists were compared.[35] In this case, radiologists decreased their false-positive rates with the help of XAI. They also evaluated an independent external test dataset to prove the potential of XAI in improving the accuracy, consistency and efficiency of breast US diagnosis worldwide. The study [35] discuss accuracy of the VGG backbone to ResNet50 and EfficientNet B0 backbone was evaluated and BI-RADS descriptors were used to evaluate.[37]

In Zhang *et al*'s study,[38] accuracy, sensitivity, specificity, F1 score, R2, Mean Squared Error (MSE), Root Mean Square Error (RMSE),d shape orientation and margin were used to test the likelihood of malignancy. Explainer I was used to explain the classification results semantically. Explainer II constructs a quantitative explanation based on the classifier and Explainer I.

The study by Amanova *et al*[32] proposes and applies a new explainability method: OMIG method. The study proved that the proposed approach yields substantially more expressive and informative results for our specific

use case. To avoid issues like limited meaning and confirmation bias due to low-fidelity explanations unnecessarily, Gurmessa and Jimma[8] suggest four metrics based on performance (D, R, F and S), but none of the selected studies used these metrics.

Bad stuff (bad decision, bad medical diagnosis and bad prediction) is the most common drawback of AI algorithms today. However, XAI could resolve this drawback. Robustness is also a characteristic expected from XAI. The study by Song *et al*[29] tested the robustness of the proposed framework. This study puts explainability as not only related to AI performance but also to responsibility and risk in medical diagnosis. XAI proves that the performance of algorithms is complementary but not enough alone. The complementing of both performance and explainability satisfaction increases the system's acceptance of legal and personal recognition.

XAI and ethical challenges

XAI overcomes ethical challenges[37 38 42 43] by providing confidence, trustworthiness, transparency, accountability and interpretability in the decision-making process. It provides an opportunity to know the reason behind the prediction for patients, clinicians and doctors.[37]

The study by Song *et al*[29] recommends focusing on augmenting AI systems to extract relevant information from past US examinations as future research. Another limitation of this work is the design of the reader study.[29] A limitation of the method proposed by Ortega-Martorell *et al*[39] is that the calculation of the FI distances when creating the embedding might be slow depending on the number of data points and the sizes of the images. However, existing implementations can be used in a high-performance computing cluster which can reduce the time considerably.[39] Future studies could re-examine the cost-effectiveness of using AI to guide breast cancer screening not just among women aged 40–49 years but also in women across the entire candidate age range, including those over age 50 years.[40] To further enhance the applicability and accuracy parameters of the model, a larger dataset across multiple centres is necessary to enhance the data quality.[42] While Sun *et al*'s study[42] focuses on age groups with the highest incidence of breast cancer, future analysis encompassing older age groups would yield significant conclusions, especially about the postmenopausal population.[42] The retrospective nature of the study[42] makes it prone to selection bias[42] and also a small size dataset used.[36]

The study by Shen *et al*[35] did not provide an evaluation of patient cohorts stratified by risk factors such as family history of breast cancer and breast and ovarian cancer are the breast cancer (BRCA) gene test results and it was only provided with US images, patients' ages and notes from the operating technician.

It is important to investigate how the experience of working with these algorithms impacts the way radiologists make decisions.[34] The image's 'low-resolution' restriction remained a limitation. In future work, it is recommended to conduct a study for qualitative assessment of the level of explainability of this approach with BUS clinicians via structured interviews and questionnaires.[37] The study by Zhang *et al*[37] stated that using a more diverse dataset, trying different convolutional neural network architectures, building a multimodal model and implementing denoising algorithms can be done to improve this research.[33] It also states that combining convolutional networks with decision trees is an interesting future work.[41] To do so OMIG is used. OMIG reveals a complex pattern behind the prediction; this pattern could also be the subject of future work.[32]

Future research can also focus on augmenting AI systems to extract relevant information from past US examinations. Another limitation of this work is the design of the reader study.[29] A limitation of the method proposed by Ortega-Martorell *et al*[39] is that the calculation of the FI distance when creating the embedding might be slow depending on the number of data points and the sizes of the images. However, existing implementations can be used in a high-performance computing cluster which can reduce the time considerably.[39] Re-examine the cost-effectiveness of using AI to guide breast cancer screening not just among women aged 40–49 years but also in women across the entire candidate age range, including those over age 50 years.[40] To further enhance the applicability and accuracy parameters of their model, a larger dataset across multiple centres is necessary to enhance the data quality.[36 42] The study by Addala[33] recommended a more diverse dataset, trying different convolutional neural network architectures, building a multimodal model and implementing denoising algorithms as a future work, combining convolutional neural networks with decision trees.[41] OMIG reveals a complex pattern behind the prediction; this pattern was the subject of future work by the study.[32]

Shen *et al*'s study[35] recommends focusing on augmenting AI systems to extract relevant information from past US examinations as future research. In addition, Shen *et al*'s study[35] did not provide an evaluation of patient cohorts stratified by risk factors such as family history of BRCA gene test results. To provide a fair comparison with the AI system, readers in the study were only provided with US images, patients' ages and notes from the operating technician.[35]

Finally, it is important to investigate how the experience of working with these algorithms impacts the way radiologists make decisions.[34] The study by Zhang *et al*[37] recommended conducting a study for qualitative assessment of the level of explainability with Breast ultrasound (BUS) clinicians via structured interviews and questionnaires.

XAI toolkits

The most popularly used toolkits that we can access from this review are DALEX and AIX360. DALEX[21 22] is a library used by R Studio. It only supports a few functionalities (ie, local post-hoc and global post-hoc), whereas AIX360[12] is a library used by Python. This toolkit supports

all functionalities (ie, data explanations, directly interpretable, local post-hoc, global post-hoc and persona-specific explanations) including the evaluation matrix.

CONCLUSION

In addition to increasing accuracy, reducing human error and technological advancement, XAI for breast cancer diagnosis overcomes ethical challenges by providing the right to know, robustness, transparency, accountability and interpretability in the decision-making process of machine learning models. However, it is not approved that it increases users' and doctors' trust in the system. Effective and systematic evaluation of its usefulness in this scenario is also lacking. Additionally, further work is needed to enhance the interpretability of deep learning algorithms through overcoming explainable to accuracy trade-offs, as well as to investigate the potential insights they can provide for clinicians' decision-making.

Twitter Daraje kaba Gurmessa @darajejaba

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD

Daraje kaba Gurmessa http://orcid.org/0000-0002-1526-7547

REFERENCES

1 Han H-J, Chu Y-C, Wang J, *et al*. Characteristics of breast cancers detected by screening mammography in Taiwan: a single institute's experience. *BMC Womens Health* 2023;23:330.
2 Arnold M, Morgan E, Rumgay H, *et al*. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast* 2022;66:15–23.
3 Lawrence RA. *2 - Anatomy of the Breast', in Breastfeeding*. Ninth Edition. Philadelphia: Elsevier, 2022: 38–57.
4 Berg WA, Bandos AI, Mendelson EB, *et al*. Ultrasound as the Primary Screening Test for Breast Cancer: Analysis From ACRIN 6666. *J Natl Cancer Inst* 2016;108:4.
5 Meenalochini G, Ramkumar S. Survey of machine learning algorithms for breast cancer detection using mammogram images. *Materials Today: Proceedings* 2021;37:2738–43.
6 Zhang Y-D, Satapathy SC, Guttery DS, *et al*. Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network. *Information Processing & Management* 2021;58:102439.
7 Zhang Y-D, Pan C, Chen X, *et al*. Abnormal breast identification by nine-layer convolutional neural network with parametric rectified linear unit and rank-based stochastic pooling. *Journal of Computational Science* 2018;27:57–68.
8 Gurmessa DK, Jimma W. A comprehensive evaluation of explainable Artificial Intelligence techniques in stroke diagnosis: A systematic review. *Cogent Engineering* 2023;10:2273088.
9 Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *J Biomed Inform* 2021;113:103655.
10 Pfeuffer N*et al*. Explanatory Interactive Machine Learning: Establishing an Action Design Research Process for Machine Learning Projects. *Business and Information Systems Engineering* 2023.
11 Carvalho DV, Pereira EM, Cardoso JS. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 2019;8:832.
12 Gutti G, Arya K, Singh SK. Latent Tuberculosis Infection (LTBI) and Its Potential Targets: An Investigation into Dormant Phase Pathogens. *Mini Rev Med Chem* 2019;19:1627–42.
13 Graziani M, Dutkiewicz L, Calvaresi D, *et al*. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artif Intell Rev* 2023;56:3473–504.
14 Klaise J, Van Looveren A, Vacanti G. Alibi explain: Algorithms for explaining machine learning models Alexandru Coca. 2021. Available: http://jmlr.org/papers/v22/21-0017.html
15 Wu L, Huang R, Tetko IV, *et al*. Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets. *Chem Res Toxicol* 2021;34:541–9.
16 Ribeiro PH, Orzechowski P, Wagenaar J, *et al*. Benchmarking Automl Algorithms on a collection of synthetic classification problems. 2022. Available: http://arxiv.org/abs/2212.02704
17 Ledell E, Poirier S. H2O Automl: Scalable automatic machine learning. 2020. Available: https://scinet.usda.gov/user/geospatial/#tools-and-software
18 Nori H, Jenkins S, Koch P, *et al*. Interpretml: A unified framework for machine learning Interpretability. 2019. Available: http://arxiv.org/abs/1909.09223
19 Maxwell AE, Sharma M, Donaldson KA. Explainable Boosting Machines for Slope Failure Spatial Predictive Modeling. *Remote Sensing* 2021;13:4991.
20 Rasheed K, Qayyum A, Ghaly M, *et al*. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Comput Biol Med* 2022;149:106043.
21 Baniecki H*et al*. Dalex: responsible machine learning with interactive Explainability and fairness in python monitoring of AI regulations view project Explainable machine learning view project Dalex: responsible machine learning with interactive Explainability and fairness in python. 2021. Available: http://jmlr.org/papers/v22/20-1473.html
22 Baniecki H, Kretowicz W, Piatyszek P, *et al*. Dalex: responsible machine learning with interactive Explainability and fairness in python. 2021. Available: http://jmlr.org/papers/v22/20-1473.html
23 Egger R. Applied data science in tourism. In: Applications RE, ed. *Interpretability of Machine Learning Models', in Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies*. Cham: Springer International Publishing, 2022: 275–303.
24 Kawakura S, Hirafuji M, Ninomiya S, *et al*. Adaptations of Explainable Artificial Intelligence (XAI) to Agricultural Data Models with ELI5, PDPbox, and Skater using Diverse Agricultural Worker Data. *EJAI* 2022;1:27–34.
25 Meng C, Trinh L, Xu N, *et al*. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep* 2022;12:7166.
26 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
27 Gurmessa WJD. Explainable machine learning for breast cancer diagnosis from Mammography and ultrasound images: A systematic review; 2023.
28 McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments. *Res Synth Methods* 2021;12:55–61.
29 Song D, Yao J, Jiang Y, *et al*. A new xAI framework with feature explainability for tumors decision-making in Ultrasound data: comparing with Grad-CAM. *Comput Methods Programs Biomed* 2023;235:107527.
30 Oh J-H, Kim H-G, Lee KM, *et al*. Reliable quality assurance of X-ray mammography scanner by evaluation the standard mammography phantom image using an interpretable deep learning model. *Eur J Radiol* 2022;154:110369.
31 Qian X, Pei J, Zheng H, *et al*. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat Biomed Eng* 2021;5:522–32.
32 Amanova N, Martin J, Elster C. Explainability for deep learning in mammography image quality assessment. *Mach Learn: Sci Technol* 2022;3:025015.
33 Addala V. BREAST AI: low cost, Explainable artificial intelligence based App for efficient diagnosis of breast cancer in developing areas. 2023 IEEE 3rd International Conference on Electronic

Communications, Internet of Things and Big Data (ICEIB); Taichung, Taiwan.2023:164–7
34 Rezazade Mehrizi MH, Mol F, Peter M, *et al*. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Sci Rep* 2023;13:1.
35 Shen Y, Shamout FE, Oliver JR, *et al*. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun* 2021;12:1.
36 Lee Y-W, Huang C-S, Shih C-C, *et al*. Axillary lymph node metastasis status prediction of early-stage breast cancer using convolutional neural networks. *Comput Biol Med* 2021;130:104206.
37 Zhang B, Vakanski A, Xian M. BI-RADS-NET: AN EXPLAINABLE MULTITASK LEARNING APPROACH FOR CANCER DIAGNOSIS IN BREAST ULTRASOUND IMAGES. IEEE Int Workshop Mach Learn Signal Process 2021;2021:1–6.
38 Zhang B, Vakanski A, Xian M. BI-RADS-NET-V2: A Composite Multi-Task Neural Network for Computer-Aided Diagnosis of Breast Cancer in Ultrasound Images With Semantic and Quantitative Explanations. *IEEE Access* 2023;11:79480–94.
39 Ortega-Martorell S, Riley P, Olier I, *et al*. Breast cancer patient characterisation and visualisation using deep learning and fisher information networks. *Sci Rep* 2022;12:14004.
40 Mital S, Nguyen HV. Cost-effectiveness of using artificial intelligence versus polygenic risk score to guide breast cancer screening. *BMC Cancer* 2022;22:501.
41 Rezazadeh A, Jafarian Y, Kord A. Explainable Ensemble Machine Learning for Breast Cancer Diagnosis Based on Ultrasound Image Texture Features. *Forecasting* 2022;4:262–74.
42 Sun J, Sun C-K, Tang Y-X, *et al*. Application of SHAP for Explainable Machine Learning on Age-Based Subgrouping Mammography Questionnaire Data for Positive Mammography Prediction and Risk Factor Identification. *Healthcare (Basel)* 2023;11:2000.
43 Dong F, She R, Cui C, *et al*. One step further into the blackbox: a pilot study of how to build more confidence around an AI-based decision system of breast nodule assessment in 2D ultrasound. *Eur Radiol* 2021;31:4991–5000.