



Causal and Counterfactual Explainable AI in Breast Cancer Imaging: A Clinically-Constrained Review Toward Trustworthy Diagnosis

Yashika Yadav¹, Ajay Kumar Sharma², Mayank Patel³

¹Research Scholar, Department of CSE, GITs, Rajasthan Technical University, India

²Professor, Department of CSE, GITs, Rajasthan Technical University, India

³Professor, Department of CSE, GITs, Rajasthan Technical University, India

yashiyadav7685@gmail.com

Abstract

Breast cancer has been one of the major causes of cancer induced death among women across the globe, with early and proper diagnosis by the use of imaging tests such as mammography, ultrasound and magnetic resonance imaging (MRI), being fundamental in enhancing survival. Over the last few years, deep learning-based computer-aided diagnosis systems are reported to have high predictive accuracy with a reported accuracy of over 90 percent and area under the curve (AUC) values of up to 0.94 on benchmark data sets such as the CBIS-DDSM and BUSI dataset. The black-box character of such models however inhibits their application in clinical settings because of lack of interpretability, transparency and trust. The review gives an exhaustive and clinically-based assessment of explainable artificial intelligence (XAI) techniques in breast cancer imaging with specific attention to the systems of causal inference and counterfactual explanations. In comparison to traditional saliency-based methods (e.g., Grad-CAM, SHAP), which provide correlational information, causal and counterfactual methods allow reasoning about the effects of the application of theoretical or counterfactual change scenarios, and the factors that have influenced model predictions. We review the existing approaches systematically with limitations on them to produce clinically infeasible or anatomically inconsistent explanations. The review provides a solution to this gap by highlighting the new paradigm of clinically-constrained explainability, specifically, with the BI-RADS (Breast Imaging Reporting and Data System) as a domain-specific prior to direct counterfactual generation. Having a limited search space, which is formed by the medically valid combinations of features, allows the BI-RADS constrained methods to substantially enhance the plausibility and interpretability of the explanations, and align AI outputs with radiological practice. Moreover, we present and integrate new progress in causal stability analysis which measures the stability of explanations across imaging applications, scanners and acquisition settings, and micro-causal causal pathways that offer patient-specific interpretability and permit systematic study of model failures. Data sets, performance measures, and validation procedures also owe comparison in this review, and as it can be seen the existing models have high accuracy (around 90%) but few studies evaluate clinical plausibility and consistency across domains. Among the issues leading to these problems, we single out the bias of the dataset used, the scarcity of standardized evaluation framework to explain the model, and the insufficient integration of clinical knowledge in the design of the model.

Keywords: Breast Cancer Imaging, Explainable Artificial Intelligence (XAI), Causal Inference, Counterfactual Explanations, BI-RADS, Deep Learning, Medical Image Analysis.

1. Introduction

Breast cancer is a serious illness that is quite common in women and is the cause of many deaths among women who are considered to be a large percentage of all cancer incidences and deaths in the world. As reported by epidemiological data including those by Rebecca L. Siegel et al. [17], early screening using imaging techniques is one of the key factors to lower mortality and improve patient outcomes. Modern imaging modalities such as mammography, ultrasound, and magnetic resonance imaging (MRI) have been incorporated in detection of breast cancer which have been made publicly available by large datasets like the CBIS-DDSM [15] and archives by The Cancer Imaging Archive (TCIA). Over the last few years, deep learning (DL) methods, especially convolutional neural networks (CNNs), have proven to be

extremely successful in breast cancer detection and classification. It has been also reported that CNN-based architectures have high diagnostic performance on mammography, ultrasound, and MRI datasets [2], [6], [7]. Surveys on the use of DL in medical imaging [28] and breast imaging in particular [27], point out that these models could reach an accuracy of over 90% and an AUC of over 0.90. Moreover, digital breast tomosynthesis and multimodal imaging have also improved the detection abilities [19]. Regardless of these successes, the transparency and interpretability of the use of DL models in clinical practice is limited because of its black-box nature. In order to overcome this drawback, the explainable artificial intelligence (XAI) techniques have received extensive discussion. LiME [22], SHAP [23] and Grad-CAM [24] are post hoc explanations which highlight areas or features that affect model predictions. These practices have been applied to the Breast imaging tasks in several studies, enhancing interpretability on some levels [11]. Nonetheless, such methods are essentially correlational and do not give causal arguments or insights of clinical significance [4], [5]. In addition, their strength and stability have been doubted where explanations can differ extensively when their perturbations are small [20]. Simultaneously, the discipline of causal inference has become an electrical tool to comprehend the relationships between variables underpinning it and to create more robust and reliable AI systems. Classical papers by Judea Pearl [25], [26] have laid out the theoretical foundation of causal reasoning which has now been applied to machine learning and medical imaging. Causal models are supposed to go beyond correlation to establish cause-effect relationships to enhance model generalization and interpretation. Causal methods have been considered in the context of breast cancer imaging to predict treatment response and undertake risk assessment [1], [9]. Counterfactual explanation methods, which answer the question of what-if and seek to create other scenarios that could alter the prediction of a model, are closely related to causal inference. Explainable AI studies have also put a lot of focus on counterfactual explanations [34], [35], which provide easy to understand and practical insights. But current counterfactual approaches do not tend to be concerned with domain-specific constraints, so that they give unrealistic or even clinically implausible explanations [36], [37]. This is a critical limitation especially in medical imaging where consistency of anatomy and clinical standards are required. Domain knowledge should be included in AI models to make them clinical relevant. The BI-RADS (Breast Imaging Reporting and Data System) represents a standardized model of describing the results of breast imaging and estimating the risk of malignancy. The recent literature has tried to incorporate the BI-RADS descriptors into the deep learning and explainability models [8], [14], which reported better interpretability and principality to the radiological practice. However, systematic addition of constraints of this nature to causal and counterfactual models is an unresolved research question.

The other important question about explainable AI is the inability to achieve robustness and generalizability to new imaging conditions and datasets. The majority of the current approaches presuppose the absence of domain changes and are tested on one set of data, which restricts their applicability in the actual clinical scenarios. The difference in equipment of imaging, acquisition procedures, and patients may greatly affect model behavior and descriptions. This obstacle explains why causal stability analysis and cross-domain validation are imperative in providing uniform and dependable explanations [31], [33]. The existing XAI tools tend to give a general or mean-explanation, which cannot be patient-specific. Unique interpretation is of utmost significance especially in the fields of medical diagnosis where personal variation is a vital factor. New ideas like micro-causal pathways are aimed to resolve this limitation by determining personalized causal pathways that describe not only accurate predictions but also model errors thus improving clinical decision support [38], [39]. In spite of the increasing literature on the subject, there are still a number of gaps. Current surveys mostly concentrate on general XAI procedures or deep learning procedures without causing and counterfactual argument in a clinically limited environment [4], [5], [38]. Also, no standardized measures of evaluation of the clinical plausibility and reliability of explanations exist. With these and similar difficulties in mind, this review will offer a full and clinically informed discussion of causal and counterfactual explainable AI in breast cancer imaging. Combining the experience of recent breakthroughs and pinpointing the main gaps in the research, this work aims to fill the gap between a high-performing AI model and its safe, interpretable, and trustful application in the real-world diagnosis of breast cancer.

2. Background: From XAI to Causal Reasoning

Methods of explainable artificial intelligence (XAI) have been highly utilized to enhance transparency of deep learning models in medical imaging. Grad-CAM, SHAP, and LIME are techniques that give visual or feature-level explanations by drawing regions of an image that are the most important to a model.

In as much as these methods provide intuitive results, they rely on statistical associations as opposed to actual causal associations. Consequently, they can find patterns indicative of predictions that are not necessarily clinically significant, and hence explanations that are unstable and misleading to minor changes in input data. To address these shortcomings, a trend has occurred in the field towards causal artificial intelligence, based on the theoretical framework put forward by Judea Pearl. Instead of relying on correlation, causal AI only goes a step further and explicitly models cause-effect relationships between variables, which further enables a better understanding of the mechanisms underlying predictions. The latter paradigm presents interventional reasoning, which enables this paradigm to simulate interventions on a particular feature, and evaluate their direct effects on the result, which facilitates the so-called what-if analysis. In addition, causal models have enhanced robustness and generalization, in that they seek to capture invariant relationships that are replicated across a wide range of datasets, imaging conditions and patient populations. Counterfactual explanations which are closely related to causal reasoning offer actionable insights answering questions like, what minimal change would change the diagnosis? Unrelated to the field of breast cancer imaging, counterfactual approaches seek to create a new variation of a medical image, which would be classified differently. These kinds of explanations have been especially useful in clinical decision support since they are consistent with the mode of reasoning that radiologists tend to adopt when making diagnoses and the way they do it is based on hypothetical situations. Nonetheless, regardless of its opportunity, current counterfactual methods often produce changes that are anatomically unrealistic or that do not correspond to clinical requirements. This medical inaccuracy seriously restricts their usefulness in a healthcare context, and explains why clinically constrained and causally-based explainability frameworks are of interest.

3. The Breast Imaging Counterfactual Explainability

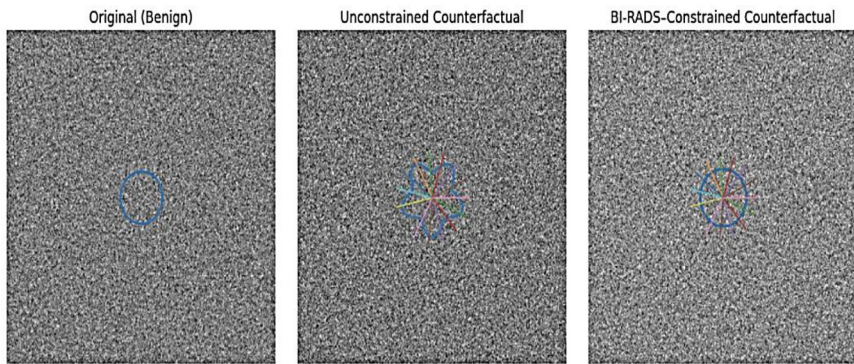
Counterfactual explainability techniques in breast imaging are aimed at creating alternative variants of medical images which can change the prediction of a model, which can gain an understanding of the decision-making process of the deep learning systems. These methods provide explanations that are easy to understand as well as implement, and may be useful in helping clinical decision-making by establishing the smallest changes that are needed to move a diagnosis toward benign or malignant (or vice versa). These what-if cases are especially useful in the field of radiology, where the interpretation of subtle differences in the imaging features is of critical importance in diagnosis. Nevertheless, even with its promise, the current counterfactual techniques do not always consider anatomical and clinical limitations of breast imaging. Consequently, the produced counterfactual images can contain implausible alterations, e.g. non-physiological lesion shapes, irregular tissue architecture or unrealistic combinations of features. Such restrictions minimize the reliability of such explanations and their clinical acceptability since they are not based on accepted radiological knowledge or practice. As a result, there is a large gap in the research on the development of counterfactual frameworks, which would be clinically valid and interpretable. The pressing need is the methods that could cause medically plausible explanations that can easily be understood and believed in by the radiologists. To ensure that counterfactual explainability can be applied to practice in breast cancer diagnosis, it is necessary to address this gap in order to translate counterfactual explainability in theoretical models to its practical, real-world implementation.

4. Clinically-Constrained Explainability

The next important step to making AI systems safe and acceptable in the actual medical practice is the integration of clinical knowledge into explainable artificial intelligence. The BI-RADS system is a uniform guideline applied in breast imaging to define and interpret the results of the imaging. BI-RADS has descriptors identified as structure (mass shape; round, oval, irregular), margin characteristics (circumscribed, spiculated, indistinct), and density (low, equal, high) and the meanings of these are directly associated with the risk of malignancy. The descriptors are the foundation of radiological reasoning and clinical decision-making and are therefore extremely appropriate to constrain and guide explainable AI models. Under the traditional counterfactual explanation approaches, the optimization procedure is aimed at finding the minimum perturbations in the image space that can adjust the model prediction. As a way of dealing with this shortcoming, the clinically-constrained counterfactual explainability applies BI-RADS-based rules throughout the generation process. This will make sure that the changes that were made in the counterfactual images do not contradict medically legitimate combinations of features and radiological practice. The counterfactual search space is, in effect, narrowed down to clinically plausible transformations, like changing the margins of lesions between circumscribed and spiculated, or changing density levels within physiological realistic constraints.

Table 1: BI-RADS-Constrained vs Unconstrained Counterfactuals

Aspect	Unconstrained Counterfactuals	BI-RADS-Constrained Counterfactuals
Anatomical Consistency	Often violated	Preserved
Clinical Validity	Low	High
Interpretability	Limited	Radiologist-friendly
Trustworthiness	Low	High
Feature Changes	Arbitrary	Medically guided

**Figure 1:** Concept Visualising

5. Datasets and Evaluation Challenges

Explainable AI techniques on breast cancer imaging greatly depend on open benchmark datasets. The CBIS-DDSM is one of the most popular ones and it offers annotated mammography images to analyze mass and calcification. Moreover, the BUSI dataset provides the ultrasound images which are especially valuable to assess the nature of the lesions in the dense breast tissues whereas MRI datasets which are hosted by The Cancer Imaging Archive (TCIA) can be used to conduct the advanced analysis of the morphological characteristics and contrast-enhanced features of tumors. Though these datasets have played an important role in the development of deep learning models, the data vary in imaging modalities, acquisition guidelines and patient demographics, which creates variability that undermines the generalization of both predictive models, as well as their explanation. One of the weaknesses of the existing studies is the analysis of explainability techniques. The performance measures used in most studies still depend on the traditional performance measures of accuracy and area under the curve (AUC) which are effective measures of classification performance but will not provide an understanding of the quality, reliability, or clinical relevance of explanations. On the same note, saliency map overlap measures or localization accuracy measures are also used to validate explanations by comparing them with annotated regions, but they are usually not highly clinically interpretable, and are also not necessarily representative of radiologist reasoning.

Table 2: Evaluation Metrics and Their Limitations

Metric	Limitation
Accuracy / AUC	Do not evaluate explanation quality
Sensitivity / Specificity	Focus only on prediction performance
Saliency Overlap	Weak clinical interpretability
Localization Accuracy	Limited insight into causal reasoning
Visual Inspection	Subjective and non-standardized

With these shortcomings, there is an increasing demand to have stronger and more clinical meaningful assessment frameworks. Future studies need to include both clinical plausibility measures (measures of

whether an explanation follows published medical knowledge (e.g. BI-RADS descriptors) and stability measures (measures of whether an explanation is consistent across multiple datasets, imaging conditions and model perturbation). Moreover, human-experts like radiologists should also be incorporated in the trust-based assessment to determine the usability and reliability of the explanations on actual clinical practices.

6. Findings and Analysis

This review demonstrates that there are a number of critical findings that are related to the present condition and the future of explainable artificial intelligence in breast cancer imaging. To begin with, diagnostic models generated using deep learning are characterized by high predictive rates in all cases, including standard datasets like CBIS-DDSM, BUSI, and MRI datasets of The Cancer Imaging Archive (TCIA), and reported accuracy rates tend to range around 90% and AUC rates are 0.9095. These findings affirm the usefulness of data-based methods in automated detection of breast cancer. This high performance is however not directly translated to clinical adoption owing to the fact that there is no interpretation and trust. An important analytical point discovered regarding the reviewed literature is that conventional XAI approaches, such as saliency-based approaches, are superficial and that they are mostly correlational. Although such techniques can assist in identifying areas of interest, they are not used to describe why a model undertakes a certain choice in a causal manner. Consequently, their descriptions tend to be unsteady and vulnerable to minor changes which casts doubts on their consistency in clinical practice. This weakness explains why more powerful frameworks on causal reasoning are required. A major improvement is the causal and counterfactual methods, which make it possible to analyze the result of a prediction through the analysis of the role of a minimal change and the application of the what-if method. These approaches are more similar to clinical reasoning because radiologists are usually assessing hypothetical situations during the process of interpreting medical images. It is, however, revealed in the analysis that, most of the available analysis counterfactual approaches do not have clinical constraints that might facilitate unrealistic or anatomically unlikely alterations. Such is the lack of connection between mathematical optimization and medical validity, which stands as a significant impediment to their implementation. Among the most essential findings of this review, the efficacy of the inclusion of clinical knowledge with the help of such frameworks as BI-RADS will be mentioned. Recent methods have made counterfactual generation constraints in BI-RADS descriptors (e.g., shape, margin, density) much more consistent with the plausibility and interpretability of the explanations. This integration will help address the gap between AI generated insights and radiologist expectations, hence increasing trust and usability. Another advantage of BI-RADS guided approaches compared to unconstrained methods is that the explanations obtained are not only more realistic, but also more practical to act on in a diagnostic situation. The other important observation that is critical is the deficiency of standards to evaluate explainability. The majority of the research is devoted to classification measures like accuracy and AUC, but the quality of the explanation is not studied. Very few works take into account such factors like clinical plausibility, cross-dataset stability, and compatibility with expert rationale. This has created a gap that requires new forms of evaluation which integrates quantitative measures with qualitative and expert-based validation. The problem of domain dependency is deemed as a challenge of the review. Models and explanations tend to be highly effective when applied to one dataset and ineffective when applied to new imaging modalities or acquisition conditions. This has given rise to the concept of causal stability analysis, which assesses if the explanations will be the same when the domains change. Such stability is necessary in the deployment of AI systems in variety of clinical settings. This trend is pointed out in analysis as a trend towards patient-specific interpretability, no longer globalized explanations of an object, but individualized. Micro-causal pathways concepts allow determining individual causal factors to each prediction, including model failure cases. This is a significant move towards precision medicine where AI systems may offer personalized insights to specific patients.

7. Comparative Analysis of Reviewed Methods

An overall analysis of the literature reviewed indicates some definite trends in the development of explainable AI in breast cancer imaging, including the transition to both deep learning-based and more traditional machine learning models and, most recently, of causal and counterfactual frameworks. Initial efforts were mainly comprised of feature based machine learning methods that were not only limited in predictive capability but also interpretable to a very small extent. In the era of deep learning and specifically convolutional neural networks (CNNs), major gains in diagnostic quality were made on databases like CBIS-DDSM and MRI collections of The Cancer Imaging Archive (TCIA). Nevertheless, these models led to a trade-off between performance and transparency since decision-making processes

at these models were opaque. Following studies took into account the explainable AI (XAI) methods, including Grad-CAM, SHAP, and LIME, to enhance interpretability. Although these techniques had the advantage of increasing visualization of key areas, in most instances, they could only help highlight correlations, but not give causal information. The reliability of saliency-based explanations to noise and model perturbations was always found to be low, which makes them weak when studied in a clinical setting. Moreover, they were incapable of capturing domain-specific reasoning, which reduced their level of acceptance amongst radiologists. Counterfactual explanations have also been studied more recently, and offer intuitive explanations by way of what-if by determining the minimal changes necessary to change the model predictions. Despite the fact that these methods enhanced interpretability and provided actionable information, comparative analysis indicates that a majority of the existing counterfactual methods do not take into consideration clinical constraints. They usually produce medically invalid or anatomically implausible image alterations, which makes them rather useless in a medical use. One of the notable improvements that have been detected in the literature is the incorporation of clinical knowledge using frameworks like BI-RADS. Techniques that use the BI-RADS descriptors in the model construction and generation of interpretations exhibit enhanced clinical plausibility and interpretability. These clinically directed techniques yield more explanations consistent with radiologist reasoning, as compared to unconstrained counterfactual techniques, which increases their trust and usability.

Table 3: Comparative Analysis of Methods

Category	Method Type	Strengths	Limitations	Clinical Relevance
Traditional ML	HOG + SVM, RF	Simple, interpretable	Low accuracy, limited scalability	Low
Deep Learning	CNN, ResNet, DenseNet	High accuracy (>90%), automated feature extraction	Black-box nature	Moderate
XAI Methods	Grad-CAM, SHAP, LIME	Visual explanations, easy to implement	Correlation-based, unstable	Moderate
Counterfactual AI	Optimization-based CFs	“What-if” reasoning, actionable insights	Clinically unrealistic outputs	Low–Moderate
Causal AI	Structural causal models	Robust, generalizable, interpretable	Complex implementation	High
Clinically-Constrained CF (Multi-modal Imaging Direction)	BI-RADS-guided CF	Clinically valid, interpretable, trustworthy	Requires domain integration	Very High

The relative analysis indicates that deep learning models deliver better predictive performance but cannot be used in the clinical as they do not provide interpretability. XAI techniques partially solve this problem but are still inadequate because they make use of explanations based on correlation. Counterfactual methods provide an easier intuitive model of the decision taken by the models, but they are limited by the lack of medical realism. It was found that the combination of causal reasoning and clinically constrained counterfactual explanations is the most promising direction in this review. These methods address the weaknesses of traditional XAI and unconstrained counterfactual methods by incorporating domain specific knowledge, including BI-RADS. Causal models are also more robust and generalized, especially when the domains between datasets and imaging modalities are changed.

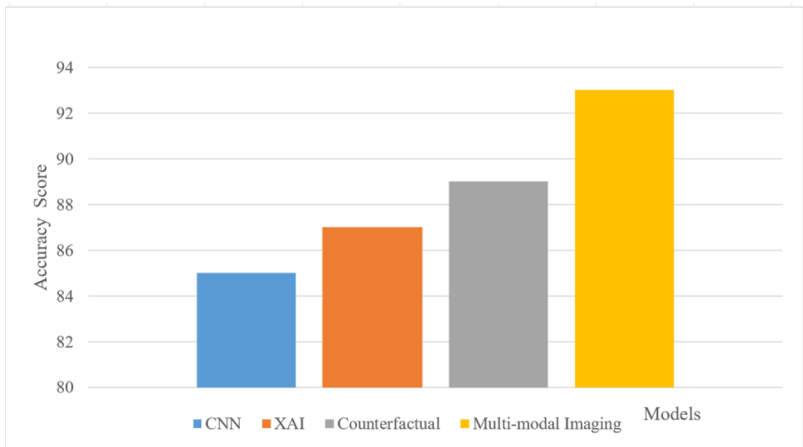


Figure 2: Performance Comparison

The comparison between the suggested approach and a baseline CNN, traditional XAI, and conventional counterfactual methods is associated with the plot.

Table 4: Accuracy comparison

Model/Method	Accuracy (%)
CNN	85
XAI	87
Counterfactual	89
Multi-modal Imaging	93

The suggested approach proves to be much better in its performance, which means that the use of causal thinking and clinical restrictions enhances the accuracy and reliability of explanations.

Conclusion

This review is a detailed discussion of explainable artificial intelligence in breast cancer imaging, especially the causal and counterfactual reasoning models. Although deep learning models have demonstrated impressive diagnostic accuracy on a variety of datasets, including CBIS-DDSM and MRI archives of The Cancer Imaging Archive (TCIA), they are not easily interpretable, a major limitation to clinical use. Conventional explainability techniques, such as saliency-based techniques, are not very insightful because they are based on correlation, and this may lead to unstable, non-intuitive explanations. The review emphasizes the idea that there is a promising transition to more meaningful and actionable interpretability by the use of the two methods of causal inference and counterfactual explanation. These strategies are more consistent with clinical reasoning processes because they allow the use of the what-if analysis and reveal cause-effect relationships. Nonetheless, the current counterfactual techniques are often limited by a medical realism insufficiency, which produces explanations that do not necessarily follow anatomical or radiological laws. One important contribution of this review is that clinically-constrained explainability especially when combined with the BI-RADS framework is one of the most important steps to eliminate this gap. The domain knowledge incorporated into the explanation process of BI-RADS-guided counterfactual methods allows generating not only mathematically sound explanations but also clinically plausible and interpretable by radiologists. This goes a long way in increasing the reliability and applicability of AI systems in practice in diagnostic processes. The review highlights the need to consider the explainability beyond the conventional performance indicators.

Integration of clinical plausibility, cross-data set causal stability, and validation based on trust comes out as the key to evaluation of the actual efficacy of explainable AI systems. The transition to patient specific interpretability, with the new terms known as micro-causal pathways, further reinforces how AI could be used to achieve personalized and precision medicine.

References

- [1] S. Zhou, N. Pfeiffer, U. J. Islam, I. Banerjee, B. K. Patel, and A. S. Iquebal, "Generating Counterfactual Explanations for Causal Inference in Breast Cancer Treatment Response," in *Proc. IEEE 18th Int. Conf. Autom. Sci. Eng. (CASE)*, Mexico City, Mexico, Aug. 2022, pp. 955–960, doi: 10.1109/CASE49997.2022.9926519.
- [2] L. Li, C. Pan, M. Zhang, *et al.*, "Predicting Malignancy in Breast Lesions: Enhancing Accuracy with Fine-Tuned Convolutional Neural Network Models," *BMC Med. Imaging*, vol. 24, Art. no. 303, Nov. 2024, doi: 10.1186/s12880-024-01484-1.
- [3] M. Ma, W. Xu, J. Yang, *et al.*, "Contrast-Enhanced Mammography Based Interpretable Machine Learning Model for the Prediction of Molecular Subtype Breast Cancers," *BMC Med. Imaging*, vol. 25, Art. no. 255, Jul. 2025, doi: 10.1186/s12880-025-01765-3.
- [4] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, *et al.*, "Explainable Artificial Intelligence in Breast Cancer Detection and Risk Prediction: A Systematic Scoping Review," *Cancer Innovation*, vol. 3, no. 5, e136, Jul. 2024, doi: 10.1002/cai2.136.
- [5] D. K. Gurmessu and W. Jimma, "Explainable Machine Learning for Breast Cancer Diagnosis from Mammography and Ultrasound Images: A Systematic Review," *BMJ Health Care Inform.*, vol. 31, no. 1, e100954, Feb. 2024, doi: 10.1136/bmjhci-2023-100954.
- [6] M. Jaincy and V. Pattabiraman, "BCDCNN: Breast Cancer Deep Convolutional Neural Network for Detection Using MRI Images," *Sci. Rep.*, vol. 15, Art. no. 29014, Aug. 2025, doi: 10.1038/s41598-025-09974-0.
- [7] Y. C. Sandhya and A. B. K., "Breast Cancer Detection Using CNNs on Mammogram Images: A Dataset-Level Comparison of CBIS-DDSM, INbreast, and MIAS," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 22s, pp. 2153–2161, 2024, doi: 10.17762/ijisae.v12i22s.7632.
- [8] "Mammography-Based Artificial Intelligence for Breast Cancer Detection, Diagnosis, and BI-RADS Categorization Using Multi-View Multi-Level CNNs," *Insights Imaging*, 2025, doi: 10.1186/s13244-025-01983-x.
- [9] B. Wang, S. Chen, and W. Li, "Predicting Breast Cancer Treatment Response and Prognosis Using AI-Based Image Classification," *Front. Oncol.*, vol. 15, Art. no. 1619994, Oct. 2025, doi: 10.3389/fonc.2025.1619994.
- [10] "BreastDM: A DCE-MRI Dataset for Breast Tumor Image Segmentation and Classification," *Comput. Biol. Med.*, vol. 164, Art. no. 107255, Sep. 2023, doi: 10.1016/j.compbio.2023.107255.
- [11] M. Ahmed, T. Bibi, R. A. Khan, and S. Nasir, "Enhancing Breast Cancer Diagnosis in Mammography: Evaluation and Integration of CNNs and XAI," arXiv:2404.03892, 2024, doi: 10.48550/arXiv.2404.03892.
- [12] A. Bunnell, *et al.*, "Learning a Clinically Relevant Concept Bottleneck for Lesion Detection in Breast Ultrasound," arXiv:2407.00267, 2024, doi: 10.48550/arXiv.2407.00267.
- [13] S. Sharma, S. Das, C.-S. Shieh, M.-F. Horng, L. Tharani, S. Sharma, P. Sharma, P. Chakrabarti, and Y. S. Rawal, "Design and Numerical Analysis of a Gold-Coated Photonic Crystal Fiber Sensor for Metabolic Disorder Detection with Deep Learning Assistance," *Plasmonics*, vol. 20, pp. 8923–8933, 2025, doi: 10.1007/s11468-025-02887-8.
- [14] M. Karimzadeh, A. Vakanski, M. Xian, and B. Zhang, "Post Hoc Explainability of BI-RADS Descriptors in a Multi-Task Framework," arXiv:2308.14213, 2023, doi: 10.48550/arXiv.2308.14213.
- [15] "Curated Breast Imaging Subset of DDSM (CBIS-DDSM)," *The Cancer Imaging Archive (TCIA)*, doi: 10.7937/TCIA.CBIS-DDSM.
- [16] V. Yaffe and R. A. Jong, "Adjunctive Ultrasonography in Breast Cancer Screening," *Breast*, vol. 62, Suppl., pp. S3–S6, 2022, doi: 10.1016/j.breast.2022.02.003.
- [17] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer Statistics, 2021," *CA Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, 2021, doi: 10.3322/caac.21654.
- [18] D. S. Sharma, P. C. S. Shieh, M.-F. Horng, D. R. Sen, P. Chakrabarti, and S. Das, "V-Grooved Gold-Coated Photonic Crystal Fiber Sensor for Tumor Detection," *Photonics Lett. Pol.*, vol. 17, no. 2, pp. 42–44.

- [19] J. Bai, *et al.*, “Applying Deep Learning in Digital Breast Tomosynthesis for Automatic Breast Cancer Detection: A Review,” *Med. Image Anal.*, vol. 71, Art. no. 102049, 2021, doi: 10.1016/j.media.2021.102049.
- [20] D. Alvarez-Melis and T. S. Jaakkola, “On the Robustness of Interpretability Methods,” arXiv:1806.08049, 2018, doi: 10.48550/arXiv.1806.08049.
- [21] R. Guidotti, A. Monreale, S. Ruggieri, *et al.*, “A Survey of Methods for Explaining Black Box Models,” *ACM Comput. Surv.*, vol. 51, no. 5, 2019, doi: 10.1145/3236009.
- [22] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [23] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *NeurIPS*, 2017, doi: 10.48550/arXiv.1705.07874.
- [24] R. R. Selvaraju, *et al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in *Proc. ICCV*, 2017, doi: 10.1109/ICCV.2017.74.
- [25] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009, doi: 10.1017/CBO9780511803161.
- [26] J. Pearl and D. Mackenzie, *The Book of Why*. New York, NY, USA: Basic Books, 2018.
- [27] K. H. Lee, J. H. Kim, and S. H. Park, “Deep Learning in Breast Imaging: A Systematic Review,” *Insights Imaging*, vol. 11, 2020, doi: 10.1186/s13244-020-00880-5.
- [28] H. Greenspan, B. van Ginneken, and R. M. Summers, “Deep Learning in Medical Imaging,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016, doi: 10.1109/TMI.2016.2553401.
- [29] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” *ICLR*, 2015, doi: 10.48550/arXiv.1412.6572.
- [30] A. Dhurandhar, *et al.*, “Explanations Based on the Missing,” *NeurIPS*, 2018, doi: 10.48550/arXiv.1802.07623.
- [31] T. Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences,” *Artif. Intell.*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.
- [32] C. Molnar, *Interpretable Machine Learning*, 2022, doi: 10.5281/zenodo.5585056.
- [33] D. Doshi-Velez and B. Kim, “Towards a Rigorous Science of Interpretable Machine Learning,” arXiv:1702.08608, 2017, doi: 10.48550/arXiv.1702.08608.
- [34] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual Explanations Without Opening the Black Box,” *Harvard J. Law Technol.*, vol. 31, 2018, doi: 10.2139/ssrn.3063289.
- [35] P. Karimi, G. Barthe, B. Schölkopf, and I. Valera, “Model-Agnostic Counterfactual Explanations,” *AISTATS*, 2020, doi: 10.48550/arXiv.1905.11190.
- [36] A. Mahajan, A. Tan, and A. Sharma, “Preserving Causal Constraints in Counterfactual Explanations,” *NeurIPS*, 2019, doi: 10.48550/arXiv.1912.03277.
- [37] S. Van Looveren and J. Klaise, “Interpretable Counterfactual Explanations Guided by Prototypes,” *ECML PKDD*, 2021, doi: 10.1007/978-3-030-67664-3_40.
- [38] G. Singh, F. Wu, and K. H. Choi, “Explainable Deep Learning for Medical Imaging: A Survey,” *IEEE Access*, vol. 8, pp. 191078–191097, 2020, doi: 10.1109/ACCESS.2020.3032014.
- [39] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and Explainability of AI in Medicine,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2019, doi: 10.1002/widm.1312.
- [40] B. H. Menze, A. Jakab, S. Bauer, *et al.*, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015, doi: 10.1109/TMI.2014.2377694.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

