Baltic J. Modern Computing, Vol. 13 (2025), No. 2, pp. 528–550
https://doi.org/10.22364/bjmc.2025.13.2.12

LLMs and XAI for Breast Cancer Transparency: A Review

Sobia DASTGEER, Povilas TREIGYS

Vilnius University, Institute of Data Science and Digital Technologies, Akademijos str. 4, Vilnius, LT-08412, Lithuania

Sobia.dastgeer@mif.stud.vu.lt, Povilas.treigys@mif.vu.lt

ORCID 0009-0005-8016-7197, ORCID 0000-0002-6608-5508

Abstract. The rising global mortality rate of women due to breast cancer highlights the urgent need for advancements in its diagnosis and early detection. Early identification of breast cancer significantly improves patient prognosis and survival outcomes. Artificial intelligence (AI), particularly Deep Learning (DL) and Large Language Models (LLMs), shows transformative potential in enhancing the diagnostic and prognostic capabilities in breast cancer detection. However, their clinical adoption remains challenged due to their "black-box" nature. Intelligent systems in healthcare, understanding the reasoning behind AI decisions is as critical as ensuring their performance, accuracy as well as patient safety and trust. Explainable AI (XAI) addresses these challenge by making AI reasoning transparent, allowing clinicians to interpret, validate, and trust model outputs. This paper reviews the application of XAI methods like SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM) in improving the transparency of DL models for breast cancer detection. This paper explores advanced XAI strategies that balance accuracy with interpretability, including attention-based mechanisms and LLM-driven explanations. In particular we discuss LLMs embedded within XAI systems, act as translational interfaces, decoding complex model outputs into clinician-friendly explanations. By adapting technical explanations to the end user's context and needs, LLMs enhance the accessibility and interpretability of complex model explanations. Collectively, these approaches help to bridge the gap between AI behavior and human understanding, ultimately improving transparency, trust and decision support especially in healthcare domain.

Keywords: Artificial Intelligence, Deep Learning, Breast Cancer, Healthcare, Explainable AI, Large Language Models

1 Introduction

Opaque decision-making systems have increased dramatically in the last fifteen years. Machine learning (ML) and deep learning (DL) models are used in a wide range of

methods in this rapidly developing field. The majority of these models are referred to as "Black-Box" due to their inherent complexity and lack of explanations of the decision-making process (Sabol et al., 2019). These "black-box" systems use advanced machine-learning algorithms to evaluate and forecast individual data, frequently containing private or sensitive data. One of the main obstacles to their adoption in mission-critical application domains, including banking, e-commerce, healthcare, public services, and safety, is their interpretability (Malhi and Främling, 2023). The European Union's General Data Protection Regulation (GDPR), effective in 2018, places strict limitations on automated decision-making systems that significantly affect users while mandating a right to explanation for affected individuals (Goodman and Flaxman, 2017). This regulation highlights the urgency for industries to adopt nondiscriminatory machine learning practices and the critical role of computer scientists in developing interpretable algorithmic frameworks that align with compliance and ethical standards.

The high failure rates of digital innovation adoption in the healthcare industry are noticeable (Guidotti et al., 2018). Artificial intelligence (AI) systems can analyze medical images, such as "Computed Tomography" (CT), "Magnetic Resonance Imaging" (MRI), "Ultrasound", "X-ray", and "Infrared Scans" to identify specific anatomical structures and identify anomalies (Raghavan, Balasubramanian and Veezhinathan, 2024). As a result, the widespread use of AI has caused people to questions: "How comfortable are we blindly trusting these AI-generated detection and results and When anything goes wrong, who is going to be responsible?". Notably, the highly effective predictions of AI models come from Deep Neural Networks (DNNs), built from incredibly complicated non-linear statistical models with countless parameters. However, the complexity of DNNs which consist of numerous non-linear layers and millions of parameters often compromises the transparency and interpretability of these models.

The most notable example of AI application in healthcare is cancer prediction (Bray et al., 2018). According to World Health Organization(WHO) Breast Cancer (BC) is the most prevalent disease worldwide, with over 2.3 million new cases annually. The most significant risk factor for breast cancer is being a woman. Women are affected by breast cancer in about 99% of cases, whereas men are affected in 0.5–1% of cases. Women with breast cancer who live in high-income nations have a 60% higher chance of surviving than those who live in low- and middle-income countries (WHO, 2024). Furthermore, 70% of breast cancer deaths occur in resource-limited environments because of challenges in early diagnosis and treatment. A study by (McKinney et al., 2020) shows that using AI may significantly enhance breast cancer diagnosis statistically, but it doesn't thoroughly examine how these developments fit into routine clinical procedures. The study ignores common problems, such as describing how the system works, making sure it is easy to use, and comprehending how it fits into collaborative practices that allow for a smooth transition into standard clinical work.

Another vital factor regarding interpretability is knowing why a system, service, or method needs to be interpretable. In some situations, explanations may not be required if no critical outcomes depend on the prediction's outcome. For instance, if the objective is to determine whether an image contains a tomato, and this information has no significant consequences, in this situation, an interpretable model may not be required, and a black-box approach might be sufficient. Consequently, explaining and interpreting the

model's outcome and functionality are essential to enhance the applicability of these systems across diverse clinical applications. Explainable AI (XAI) aims to provide researchers with a wide range of tools to understand the opaque nature of black-box AI systems, with a focus on transparency and the interpretability of AI models utilized to make decisions (Croce et al., 2024).

This article focuses on the current state of research, contributions made in this area of XAI and LLMs, and an investigation into what is still to be discovered. Our ultimate goal is to give a comprehensive taxonomy in the field of XAI, which helps those who are new in this field that they can use it as a guide to advance future research while also motivating professionals and experts from other fields to embrace the advantages of AI in their respective fields, free from assumptions about its interpretability. This study, explored the potential of integrating LLMs into XAI pipelines to enhance the interpretability of breast cancer prediction models. Initially, an image-based machine learning model is trained using a breast cancer dataset to predict diagnostic outcomes. Post-hoc explanation techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) or SHapley Additive exPlanations (SHAP), are then employed to generate visual or textual insights into the model's decision making process. However, these explanations often lack clarity and are not easily understandable by non expert users. To address this gap, emerging research has investigated the use of LLMs to generate user friendly, human centered narratives that describe the reason behind the model's predictions. This integration helps to bridge the gap between complex model behavior and end user interpretability by making explanations more accessible, trustworthy, and actionable. Figure 1 illustrates the yearly publication trend from 1999 to 2024 in the healthcare domain using interpretable, explainable, and transparent AI approaches. The core contribution of this study are following:

- An analysis of the role of XAI with a focus on healthcare domain.
- Emerging trends and tools in XAI and LLMs for enhancing interpretability in AI.
- An exploration of how LLMs enhance the explanation by translating them into more understandable format.

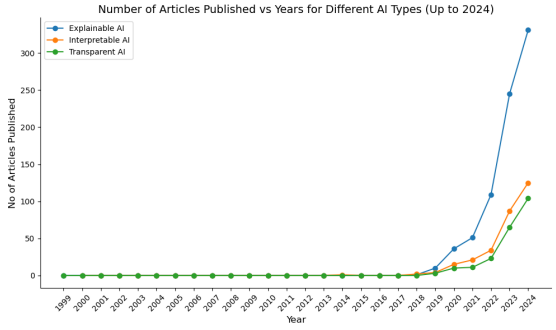

Figure 1: Yearly publication for interpretable, explainable and transparent AI in healthcare(Data derived from SCOPUS)

2 Survey Strategy

To ensure a comprehensive review, we structured our survey strategy into three phases: identifying relevant studies, implementing inclusion and exclusion criteria, and extracting the most suitable articles for detailed analysis.

Step 1: Identifying studies To systematically compile peer-reviewed research on the use of XAI in breast cancer diagnosis, we employed an automated search strategy using specific keywords. Our search targeted reputable academic databases known for publishing high-impact healthcare AI research, including PubMed, IEEE Xplore, Scopus, and Google Scholar.

Step 2: Inclusion and exclusion criteria

- Articles focusing on the use of XAI and LLMs methods and tools for breast cancer diagnosis across multiple imaging or data modalities (e.g., Mammography, MRI, Histopathology).
- Studies employing black-box models (e.g., non-interpretable AI) for breast cancer diagnosis that lack explicit explanation methods or recent approaches integrate LLMs to enhance interpretability by generating human understandable explanations from these opaque systems.
- This survey excluded those articles published prior to 2019, to prioritize recent advancements in XAI and breast cancer research.
- Articles are excluded other than English language, due to potential inconsistencies in translation and accessibility.

Step 3: Extracting suitable articles To ensure the quality and relevance of our review, we applied predefined inclusion and exclusion criteria. Selected studies had to be original research articles published in the aforementioned peer-reviewed journals and must have employed at least one explainable artificial intelligence (XAI) methodology or Large Language Model (LLM) in the context of breast cancer. We conducted a thorough screening of titles and abstracts, excluding studies that did not meet the inclusion criteria. Specifically, we excluded studies that focused on XAI or LLMs without addressing breast cancer, studies on breast cancer without XAI components, preprints pending peer review, duplicate entries, and non-research materials such as books, dissertations, and technical notes. After this screening process, 224 studies were excluded, resulting in a final selection of 54 articles that met all inclusion criteria and were included in our comprehensive analysis.

Table 1 presents the search strings used for article selection, covering publications from January 2020 to December 2024.

3 Fundamental Concepts and Background

Traditionally, radiologists analyze mammograms to identify and diagnose malignancies. This is often done in consultation with other medical professionals for a final decision, but in rural areas and developing countries, access to qualified experts is limited. The complex structure of breast tissue and the peculiarities of breast tumors further

Table 1: Review articles published from 2020 to 2024 were selected based on keywords, with the number of papers retrieved from different databases according to predefined inclusion and exclusion criteria.

Database	Keywords	Paper count
Scopus	(TITLE-ABS-KEY (”Explainable Artificial Intelligence” OR ”Explainable AI” OR ”XAI” OR ”Large Language Model” OR ”LLMs”)) AND TITLE-ABS-KEY (”Breast Cancer”) AND PUBYEAR $>$ 2019 AND PUBYEAR $<$ 2025	192
IEEE Xplore	(”Abstract”:”Explainable Artificial Intelligence” OR ”Abstract”:”Explainable AI” OR ”Abstract”:”XAI” OR ”Abstract”:”Large Language Models” OR ”Abstract”:”LLMs”) AND (”Abstract”:”Breast Cancer”)	2
Google Scholar	(”Explainable Artificial Intelligence” OR ”Explainable AI” OR ”XAI” OR ”Large Language Models” OR ”LLMs”) AND (”Breast Cancer”)	30
PubMed	(”Explainable Artificial Intelligence”[Title/Abstract] OR ”Explainable AI”[Title/Abstract] OR ”XAI”[Title/Abstract] OR ”Large Language Models”[Title/Abstract] OR ”LLMs”[Title/Abstract]) AND (”Breast Cancer”[Title/Abstract]) AND (2020[Date - Publication] : 2024[Date - Publication])	54

complicate the manual analysis process. In contrast to human inspection, AI-based automated image analysis expedites the screening process by saving time and effort by effectively collecting valuable and relevant information from vast amounts of images (Schaffter et al., 2020). To automate the identification of breast cancer, researchers have used a variety of imaging modalities, including CT, MRI, Ultrasound, Thermography, Mammography, and Histopathological imaging (Thakur et al., 2024).

3.1 Datasets & Breast Cancer Screening Approaches

To detect breast cancer, several imaging techniques have been developed. The different methods depend on many factors, like the cancer's size, location inside the body and aggressiveness. Among the most widely recognized methods for diagnosing and determining breast cancer in its early stages are Mammography, Thermography, MRI, Positron Emission Tomography (PET), CT, Ultrasound and Histopathology (Karthiga et al., 2024). Table 2 provides an overview of publicly available breast cancer datasets, including their imaging modalities and corresponding access links. This section examines existing diagnostic techniques, reviews key studies and validating their effectiveness, and outlines evidence based guidelines for clinical application.

3.1.1 Manual Physical Breast Cancer Checkup A healthcare professional or the patient can perform a Breast Physical Examination (BPE), also known as a Clinical Breast Examination (CBE), to detect abnormalities in breast tissue, such as lumps, asymmetry, or skin changes (Mohamed et al., 2021). To assess texture, mobility, and potential masses, the examiner applies varying pressure levels to palpate the breasts and

Table 2: Public Datasets for Breast Cancer Imaging

Ref	Modality	Dataset name	Dataset availability link
(Spanhol et al., 2015)	Histopathology	BreakHis	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
(Mangasarian et al., 1995)	cytology	Breast Cancer Wisconsin(Diagnostic)	https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
(Araújo et al., 2017)	Histopathology	Breast Histology Dataset	https://rdm.inesctec.pt/dataset/nis-2017-003
(Suckling, 1994)	Mammograpghy	Mini-MIAS	http://peipa.essex.ac.uk/info/mias.html
(Rose et al., 2006)	Mammograpghy	DDSM	http://www.eng.usf.edu/cvprg/mammography/database.html
(Moreira et al., 2012)	Mammograpghy	INBreast	https://biokeanos.com/source/INBreast
(Ramos-Pollán et al., 2012)	Mammograpghy	BCDR	https://bcdr.eu/information/about
(Halling-Brown et al., 2020)	Mammograpghy	RSNA	https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostics-waiting-times-and-activity/imaging-and-radiodiagnostics-annual-data/
(Saha et al., 2021)	Radiology	Duke-Breast-Cancer-MRI	https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/
(Rodrigues, 2017)	Radiology	Breast ultrasound image	https://data.mendeley.com/datasets/wmy84gzngw/1
(Institute, 2025)	Radiology	The Cancer Genome Atlas Program (TCGA)	https://portal.gdc.cancer.gov/

surrounding tissues, including the axillary lymph nodes. This non-invasive technique is crucial for the early detection of breast cancer, particularly for individuals who are not yet eligible for routine mammography or those living in resource-limited settings. BPE are cost-effective and easily accessible screening methods that do not require specialized equipment, making them particularly valuable for initial clinical assessments (Sultania et al., 2017). Despite being sensitive compared to advanced imaging modalities such as Mammography or MRI, BPE serves as a critical complementary component of screening programs, helping identify suspicious abnormalities requiring advanced diagnostic assessment.

3.1.2 Mammograpghy Mammography is a key imaging technique used for the early detection of breast cancer, which uses low-dose X-rays to visualize internal breast tissue. It effectively identifies microcalcifications, lumps or structural distortions that may indicate malignancy frequently before symptoms show up (Welch et al., 2016). The breast is compressed between two plates to get high resolution images typically in

craniocaudal and mediolateral oblique views. Although the test can identify abnormal regions, it cannot determine that they are cancer. Although challenges such as false-positive results, false negatives (notably in dense breast tissue) and patient discomfort exist, innovations like digital mammography and 3D tomosynthesis have enhanced diagnostic precision and minimized recall rates. Its sensitivity is still limited, especially in high risk patients which increases the likelihood of false positives and raises questions about the careless use of population based screening. Additionally, age and breast density affect mammography accuracy with younger people or those with dense tissue showing lower sensitivity (Geisel et al., 2018).

3.1.3 Magnetic Resonance Imaging (MRI) Breast MRI is an advanced, non-invasive diagnostic tool that utilizes powerful magnetic fields and radio waves to generate highly detailed, cross-sectional images of breast tissue. Compared to Mammography or Ultrasound, breast MRI excels in soft tissue contrast and it is beneficial for high risk patients such as those with BC gene mutations or dense breasts, and for evaluating complex cases where other imaging results are inconclusive (Kuhl, 2024). It is particularly used to assess tumor extent, monitor chemotherapy response and screen for cancer recurrence. The procedure often involves a gadolinium based contrast agent which enhances visualization of abnormal blood flow patterns associated with malignancies. While breast MRI boasts high sensitivity in detecting cancers it has lower specificity sometimes leading to false positives and unnecessary biopsies (Gao and Heller, 2020). Additionally, it is more time-consuming, costly and requires careful consideration for patients with certain implants, renal impairment or claustrophobia.

3.2 Challenges in breast cancer recognition using AI

We review several articles on breast cancer detection using AI and address the issues identified in these studies:

- **Limited Public Datasets**: The lack of publicly available datasets limits the progress of breast cancer diagnostic research and poses an obstacle to model development.
- **Unbalanced and insufficient data**: Unbalanced datasets and small sample sizes can negatively impact model performance. This makes it challenging to get reliable results.
- **Data loss in data preprocessing**: Techniques such as data scaling solve the problem of small data sizes. This often leads to data loss. This may affect the quality of the input data.
- **Bias in AI Algorithms**: Sometimes AI algorithms can produce biased results. This challenges the development of models that can be generalized to diverse communities.

4 Impact of Explainability on AI Systems

Machine learning and Deep learning models are often criticized as 'black boxes' due to their inherently opaque and complex structures (Dastgeer and Treigys, 2024). The

opaque nature of their decision-making processes makes it challenging for researchers to justify their outputs in human-understandable terms. This lack of transparency has gained significant interest in XAI. A concept significance arises from its social need. Given the increasing focus on explainability in AI algorithms, we identify a few crucial areas where XAI might result in bringing about transformative change.

4.1 Explainable AI (XAI)

The concept of XAI has deep roots in computational history. Early research on this topic can be traced back to literature published over four decades ago. Early examples include rule based expert systems that explained their outcomes based on applied rules (Swartout, 1985). The term XAI refers to the characteristics that explain how the AI model makes its predictions (Shi et al., 2022). According to (Sadeghi et al., 2024), XAI focuses on creating an interface that makes AI decision-making accessible and helps people to understand. Interpretability focuses on creating human understandable rules that explain how a system makes its decisions. In the healthcare industry, where decisions can have critical consequences, it is essential to understand how AI algorithms generate their recommendations. Healthcare workers may find it difficult to assess and trust AI systems outputs if they lack explainability potentially leading to hesitation in adopting these technologies.

4.2 Need of Explainable AI (XAI)

AI models, often called ”black boxes” frequently produce unjustifiable, unexplainable and unaccountable outcomes. In recent years, the XAI field has received more attention. These days, it is crucial for AI systems not only provide precise diagnoses but also offer supplementary information that clarifies or supports the complex classifier decisions. A study (Moxey et al., 2010) highlights that physicians typically do not prefer black boxes in medical systems because they would rather know how the system generates this decision. According to (Lamy et al., 2019) the primary goal of XAI is to develop intelligent systems that can clearly and understandably communicate their choices, predictions, and behaviors to users. This approach aims to develop models that produce correct results and explain the reasoning behind them. This makes it easier for users to trust and communicate with AI systems, particularly in crucial fields like healthcare, finance, and law. XAI focuses on enhancing the transparency, accountability, and fairness of AI systems which help users to understand the model’s behaviour better and make defensible judgements based on its recommendations (Hassija et al., 2024). It draws attention to essential methods like LIME, SHAP, Grad-CAM, and other vital factors that advance explainability and interpretability. Figure 2 illustrates the overview of the problem of black-box AI in medical diagnosis and the proposed solution using XAI and LLMs to provide interpretable and user friendly explanations for clinical decision making.

4.3 Explainable Artificial Intelligence in Medical Diagnostics

The use of XAI to explain medical diagnostic conclusions has recently come into the spotlight of the scientific community. Therefore, it can be understood that the healthcare

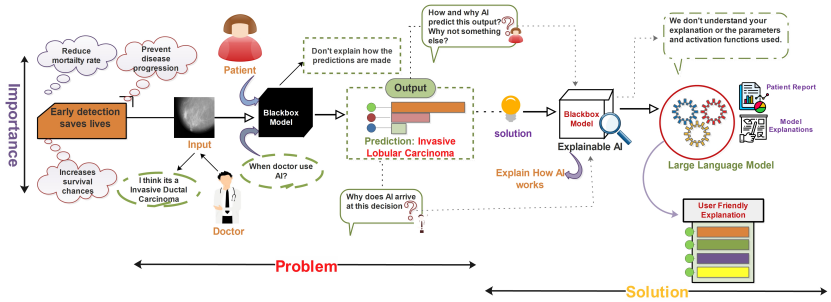

Figure 2: Explainable AI Framework for Medical Diagnosis

sector is unique in its need where user acceptability of AI algorithms depends on both explainability and accuracy (Zhang et al., 2022). Medical professionals must ensure the models are appropriately trained and the parameters they rely on align with their expertise (Aziz et al., 2024). For example, suppose an ML model's post-hoc analysis findings indicate that fatigue is an indication of breast cancer. In that case a medical expert may instantly suggest that the ML model is unreliable. In applications such as sentiment analysis, spam detection, or recommendation systems the lack of user participation may not pose significant issues. This is because experts can independently analyze the outputs of XAI methods to debug models and identify gaps in the training data. However, in the medical domain the situation is different. Even if XAI methods provide plausible explanations, only clinicians can properly analyze the outputs and understand the causes of failure cases in ML models especially in critical areas like breast cancer diagnosis.

4.4 Four Principles of XAI

The increasing use of AI systems in high stakes fields such as healthcare, finance, and legal decision-making requires these principles since opaque "black-box" models compromise accountability, safety, and trust (Angelov et al., 2021). According to (Phillips et al., 2020) an AI system must meet these four essential guidelines to be classified as an XAI:

- **Explanation**: AI systems must provide a transparent justification for their outcomes providing relevant contextual evidence or specific reasons.
- **Meaningful**: To ensure clarity, explanations should match the user's expertise level and be conveyed in an understandable, clear, and appropriate format.
- **Accuracy**: The system's explanation must clearly and accurately represent its internal processes and decision pathways, ensuring that they are neither oversimplification or misrepresentation.
- **Knowledge limits**: This principle asserts that AI systems must recognize limitations outside of their intended design where their responses may not be reliable.

The explanation promotes user confidence in AI results by allowing users to examine judgments such as medical diagnoses. For example, a medical professionals need explanations that differ from those of patients or regulators to achieve a meaningful explanation. Accurate explanations can help avoid misunderstandings that bias a model's reasoning essential for evaluating regulatory compliance and identifying algorithmic bias. Lastly, knowledge limits reduce the risk of damage by preventing overconfident or outside of scope predictions such as an AI that detects uncommon diseases it was never trained on.

5 Insights into Explainable AI (XAI)

Explainability in AI models has been attained using a variety of methods and strategies. Table 3 summarizes key studies that have applied XAI techniques in breast cancer diagnosis. The table lists the datasets used, the specific explanation techniques employed (such as Grad-CAM, SHAP, or LIME), machine learning models applied and the limitations identified in each study. This comparison provides insight into the current landscape of XAI applications in medical imaging helping to identify which methods are frequently used as well as their associated challenges.

5.1 Explainability Methods

The method for explaining AI behavior depends on the type of machine learning algorithm. Some algorithms produce inherently transparent models (e.g., Decision Trees, Bayesian Classifiers, Random Forest), while others like deep learning algorithms create complex black-box models that require specialized techniques to interpret their decisions for users to understand (Hall and Gill, 2019). The explainability method is categorized into two categories: How explanations are generated and When explanation are provided? Another key criterion for classifying XAI techniques is the scope of explanations which can be categorized into local explanations focusing on individual predictions and global explanations providing a broader understanding of overall model behavior.

5.1.1 Model-specific vs Model-agnostic Based on how explanations are generated, explainability methods in machine learning are categorized into two types: Model-specific and Model-agnostic. Model-specific approaches are designed to analyze particular types of models by examining their internal structure and parameters to generate insights (Ai and Narayanan. R, 2021). For instance, in Random Forest Models, feature importance is calculated using techniques directly tied to the model's structure. One such technique is the Gini importance metric which evaluates how much each feature reduces prediction uncertainty or impurity. Alternatively, permutation importance assesses a feature's impact by randomly shuffling its values and measuring the resulting decline in model performance. These techniques help identify which features contribute the most to the model's predictions.

Table 3: Comparison of Explainability methods used in different breast cancer studies

Ref no	Dataset	Method	Explanation Technique	Limitation
(Dhiman et al., 2024)	OCT images	TOPSIS & CSA	SHAP	The model's performance may vary if a different dataset is used
(Maheswari et al., 2024)	Fine Needle Aspirate (FNA) images	KNN, SVM, RF Naive Bayes	LIME & SHAP	When applied to another, unexplored situations with different features, the model's performance might not be as accurate or effective.
(Dihmani et al., 2024)	Infrared Image (DMR-IR)	Hybrid Particle Swarm Optimization & Hybrid Spider Monkey Optimization	SHAP	The study focused solely at one imaging modality; adding MRI or mammography might make feature extraction and interpretability more difficult.
(Briola et al., 2024)	Wisconsin Breast Cancer(tabular dataset)	XGBoost	SHAP	The effectiveness of federated learning depends on the consistency and quality of the data from all sources differences affect performance.
(Raghavan, B and v, 2024)	Infrared breast images	DenseNet201, VGG19 & EfficientNetB7	Attention guided grad cam	Using explanation maps resulted in a 42.5% decrease in performance, indicating a compromise between accuracy and interpretability.
(Kaushik et al., 2023)	Infrared breast imagery	DenseNet201	Grad-CAM	Clear and well-structured data is necessary for denoising autoencoders and classifiers to produce precise predictions.
(Rajpal et al., 2023)	DNA methylation data	MethylMarker Framework (Deep Neural Network-based)	SHAP	A single-omic approach to DNA methylation could miss information from combining data from several omics.
(Khater et al., 2023)	WBCD (tabular dataset)	KNN	LIME	Biases in the dataset might affect the model's decision.
(Paudel et al., 2023)	Categorical data	Support Vector Machine,Random Forest,MultiLayer Perceptron	LIME & SHAP	Although it is impressive to get high F1 scores (above 0.98), the study overlooks the possibility of overfitting and generality across other patient groups or datasets.
(Farrag et al., 2023)	INBreast (Mammogram dataset)	DeepLabv3	Grad-CAM	It is in doubt how well this study performs well because it does not compare with other cutting-edge segmentation models (such as U-Net, U-Net++, and Swin UNETR).

However, attention mechanisms in transformer models illustrate model-specific interpretability by revealing how the model processes and prioritizes input elements during prediction. For instance, in Natural Language Processing (NLP) tasks, self-attention layers generate heatmaps that visualize which words or phrases the model focuses on when making predictions. Similarly, in computer vision, attention maps identify salient regions in an image that contribute most to the model's decision. By directly exposing the model's internal focus attention based interpretations help to bridge the gap between the complexity of transformer layers and human understanding of their decision making process. In contrast, model agnostic techniques work across various model types and treat the model as a "black box," avoiding reliance on internal parameters. Methods such as LIME and SHAP provide interpretations by analyzing input output relationships rather than requiring deep expertise in the model's architecture (Wikle et al., 2023). This flexibility makes them particularly useful for explaining complex models like DNNs in a way that is more accessible to non experts.

5.1.2 Post-Hoc vs Transparent Explainability Based on when explanations are provided explainablity methods in machine learning are categorized into two methods:Post-hoc and Transparent. Post-hoc explainability methods systematically examine the internal logic and behavior of trained machine learning models after they generate predictions. These methods may also use surrogate modeling to deconstruct the mechanistic rationale behind the model's input output relationships (Rai, 2020). Transparent models also referred to as ante-hoc methods, intrinsically interpretable or glass-box models prioritize explainability by embedding interpretability directly into a model's architecture or training process. These approaches create inherently understandable systems ensuring transparency from the outset (Retzlaff et al., 2024). Rule Based Systems, such as Bayesian Rule Lists (Letham et al., 2015) are intrinsically interpretable models that classify data using a series of logical "if-then" conditions. These models operate by repeatedly identifying simple, human-readable rules that partition the data into subsets based on feature thresholds or categorical conditions. The interpretability arises from their clear, sequential logic which closely align with human decision making processes. Unlike black-box models, rule based systems provide outcomes that can be directly traced through the applied rules, allowing users to validate each step of the reasoning. Their transparency makes them particularly useful in domains like healthcare where stakeholders need clear justifications for predictions (Rudin, 2019).

5.1.3 Global vs Local Explanation Local interpretability methods aim to explain a model behavior for specific instances or regions within the input space, rather than providing a comprehensive understanding of model's decision making process (Hakkoum et al., 2024). These techniques generate instance specific explanations by approximating the model's behavior locally (e.g., near the data point of interest). While such explanations reveal how the model responds to particular inputs they do not necessarily provide insights into the model's broader decision patterns patterns or generalize to its overall functionality. Global explanation methods offer a holistic understanding of a machine learning model's behavior across the entire dataset rather than explaining an individual predictions (local explanations). These techniques reveal overarching pat-

terns, feature importance and decision logic making them essential for auditing models, ensuring compliance, and building trust (Radensky et al., 2022). Figure 3 illustrates the classification of AI models into transparent (glass-box) models which are intrinsically interpretable (e.g., decision trees, linear models), and black-box models (e.g., deep neural networks, ensemble methods), which rely on post-hoc explanation techniques. The figure further classifies explainability methods by scope (global vs. local) and technique (model-agnostic vs. model-specific), such as feature importance scores, surrogate models or saliency maps. Transparent models enable direct inspection of their logic, while black-box systems require external methods to approximate or extract their decision-making patterns.

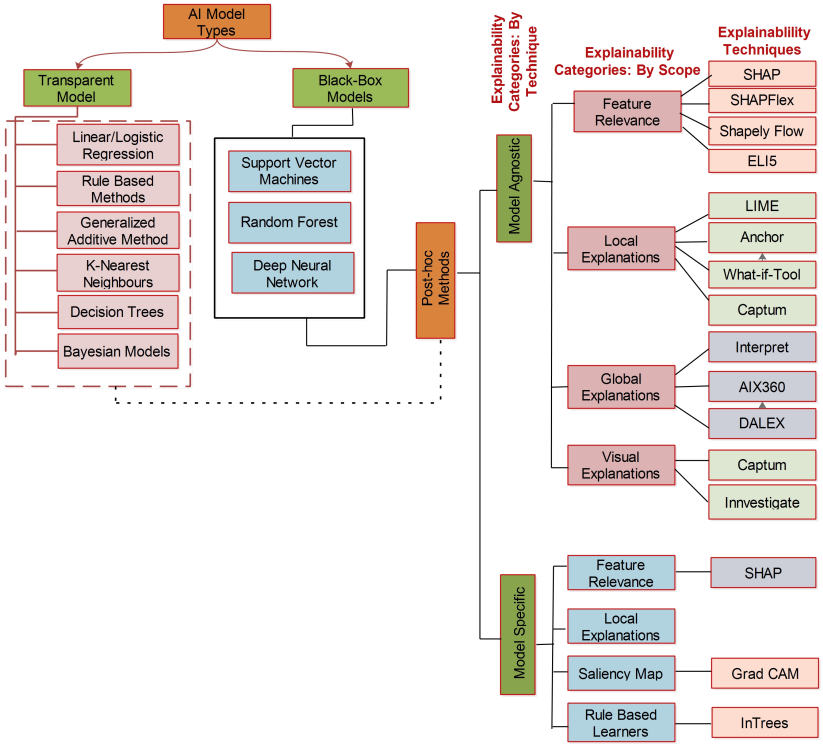

Figure 3: Taxonomy of AI Models and Explainability Approaches

5.2 Hurdles in Enabling XAI

There are several key challenges in achieving explainability for ML models. As noted by (Adadi and Berrada, 2018) all Black-box systems don't have to justify every decision they make because doing this could have a number of negative effects, including

worse system performance and higher development costs. In the absence of comprehensive development framework for XAI, local interpretation methods has become a common practice to explain the cases being investigated. Complex machine learning model analysis requires a foundation in advanced statistics and mathematics concepts. The regions that are crucial to the model's predictions are identified by XAI techniques but the underlying characteristics that give these regions their significance are not explained. Healthcare systems have not yet been able to meet the functional and design requirements for the effective use of machine learning models in the medical field. End users actively participate in the creation of ML models in order to align technology with practical requirements and ensure that the end result meets user expectations. Although XAI methods provide meaningful insights, only healthcare care providers can appropriately assess the results and understand the reasons behind failures of the ML model, particularly in crucial domains such as breast cancer detection (Chaddad et al., 2023). Therefore, ML experts often have to rely on clinicians for debugging and improving their models.

6 The Urgency of Large Language Models in Medical Diagnostics

The opacity of black box models poses a critical challenge in high-stakes healthcare applications, such as breast cancer diagnostics, where clinicians and patients need clear, actionable insights. Traditional Explainable AI (XAI) method like SHAP and Gradients typically rely on visualizations (e.g., heatmaps, saliency maps) or numerical outputs to interpret model outcomes. While these approaches are valuable for ML experts, they may fail to communicate with end users such as radiologists, oncologists, or patients who lack specialized technical training. This gap undermines trust and limits the clinical adoption of AI tools despite their diagnostic potential. LLMs offer a transformative solution by converting complex model behaviors into contextual, natural language explanations (Williams et al., 2024). For instance, in breast cancer diagnostics, an LLM could elucidate why a deep learning model classified a mammogram as "high risk" by summarizing key features (e.g., microcalcifications, tumor morphology in easily understandable terms. This aligns with broader trends in XAI research where frameworks like LLaVA-Med (Li et al., 2023), a multimodal LLM tailored for biomedical applications. It is trained to process medical images (e.g., Radiology scans, Histopathology slides) alongside textual data (e.g., clinical notes, lab reports) and generate natural language responses. Adapting similar approaches to healthcare could empower clinicians to validate AI driven insights and more effectively communicate the reasoning behind diagnoses to patients.

6.1 Enhancing AI Model Transparency via Large Language Models

In recent years, LLMs have shown great promise in enhancing XAI by translating complex machine learning outputs into coherent and accessible human language. The human centered approach further guides the refinement of these explanations by considering user's comprehension levels, contextual needs, and interaction preferences (Zhou et al., 2024). By involving end users in the development process through methods such

as interviews and scenario based evaluations. The explanations generated by the LLMs are not only technically accurate but also socially relevant and easy to understand. This approach ultimately enhances the transparency of AI models in healthcare and supports more informed and confident decision making by medical professionals and patients alike.

6.2 Enhancing Breast Cancer Diagnosis Interpretability with Large Language Models

In recent developments, the integration of Deep Learning based image analysis with LLMs has emerged as a promising approach to enhance interpretability in AI-driven breast cancer diagnosis. While CNNs and vision transformers (ViTs) achieve state of the art performance in classifying Mammographic images (e.g., malignant vs. benign), post hoc explanation methods such as Grad-CAM and LIME typically generate saliency maps and feature importance scores. However, these technical outputs are often not easily interpretable by medical professionals or patients. To address this issue, emerging research has investigated the use of LLMs to generate human centered natural language explanations that better convey the model's reasoning in an accessible manner. By translating complex outputs into coherent narratives, LLMs help bridge the gap between model behavior and user understanding, improving transparency, trust and decision support particularly for medical professionals such as radiologists and oncologists. LLMs (e.g., GPT-4, LLaMA-3) can be incorporated as translation layers that convert structured XAI outputs including prediction confidence, salient image regions (e.g., spiculated masses, microcalcifications), and metadata (e.g., lesion size and location)into natural language explanations (Egli, 2023).

7 Tools for Fairness and Explainability in Interpretable AI

Machine learning practitioners frequently require tools to examine and evaluate their models (Rahman et al., 2023). Essential steps to enhance performance include assessing a model's effectiveness and investigating how input modifications impact its outcomes.

7.1 Tools for ensuring Fairness and reducing Bias

IBM AI Fairness 360 (AIF360): The open source toolkit IBM AI Fairness 360 (AIF360) (Varshney, 2018) offers a collection of measurements, algorithms and bias mitigation strategies to identify discrimination and resolve bias in machine learning models. It has tools for using preprocessing and postprocessing strategies to improve fairness, as well as the ability to measure bias in datasets and models (Blow et al., 2024).
The What-If Tool (WIT): WIT from Google is an open source TensorBoard web application that allows users to evaluate the performance and fairness of machine learning models (Wexler et al., 2019). The tool requires only a sample dataset and trained models.

Fairlearn : This toolkit is designed to help practitioners to evaluate and improve fairness of AI systems (Bird et al., 2020). Its accompanying Python library, supports fairness in AI by allowing practitioners to evaluate model outputs across different populations and includes specific algorithms designed to mitigate bias and fairness issues.

7.2 Agnostic Explainability Tools

SHapley Additive exPlanations (SHAP): SHAP is an Explainable Artificial Intelligence (XAI) method based on game-theoretic principles (Lundberg and Lee, 2017). It interprets machine learning models by considering individual features as team members working together to achieve a common goal, where the model's outcome represents the collective payoff. By calculating each feature's unique contribution to the result, SHAP provides both local and global explanations, offering insights into feature importance across the dataset as well as overall model behavior.
Local Interpretable Model-agnostic Explanations (Lime): LIME (Ribeiro et al., 2016) helps to enhance the interpretability of a machine learning models and make its individual outcome more understandable. It provides local explanations by approximating the model's behavior for a specific single instance, helps it to for understand how a particular outcome was made.
Anchors: It is an open-source toolkit that generates high-precision, rule-based explanations for individual model predictions (Ribeiro et al., 2018). It identifies minimal conditions ("anchors") under which the prediction remains unchanged, thereby enhancing transparency and trust in AI systems through locally faithful and interpretable rules.

7.3 Explainability Methods for Deep Neural Networks

Captum: Captum is an open-source PyTorch library for model interpretability, offering a unified framework to implement and evaluate feature attribution methods. It supports gradient-based and perturbation-based algorithms to explain predictions across diverse models including complex architectures like graph neural networks in both classification and non-classification tasks (Stanchi et al., 2023).
Gradient-weighted Class Activation Mapping (Grad-CAM): Grad-CAM is a visual attribution method for CNNs that generates coarse heatmaps highlighting image regions critical to a model's class prediction. It computes gradients from a target class back to the final convolutional layer, weighting activation maps to reveal influential spatial features (Selvaraju et al., 2017).
Integrated Gradients: Integrated Gradients is an attribution method for DNNs that distributes a model's prediction to input features (Sundararajan et al., 2017). It satisfies two core principles: Sensitivity (non-zero attribution for output-changing features) and Implementation Invariance (identical attributions for functionally equivalent models). The approach requires no model modifications and leverages standard gradient computations.

7.4 Large Language Model Explainability Tools

ExBERT: ExBERT is an interactive visualization tool for exploring attention mechanisms in transformer models (e.g., BERT). It visualizes word-to-word attention across

layers and heads, enabling granular analysis of how LLMs process linguistic relationships (Gajbhiye et al., 2021).

BertViz: BertViz is an open-source visualization tool enables interactive exploration of transformer attention mechanisms, visualizing token to token patterns and head dynamics in self or cross attention (Vig, 2019). It advances explainability by revealing how input tokens influence predictions and exposing biases or spurious correlations. Researchers leverage it to debug behaviors, improve model design, and enhance trust in AI outputs.

ExplainaBoard: ExplainaBoard is an open-source toolkit for interpretable evaluation of NLP models (Yuan et al., 2021). ExplainaBoard converts standard NLP evaluation into an interpretable, diagnostic, and comparative analysis, empowering researchers to decode model behavior beyond metrics.

Figure 4 tracks the rising GitHub star counts of machine learning explainability repositories over time. The x-axis represents the period from 2015 to 2025, while the y-axis quantifies repository popularity through accumulated stars. The repositories include well known tools include SHAP (slundberg/shap), LIME (marcotcr/lime), Captum (pytorch/captum), and other libraries dedicated to AI interpretability. SHAP exhibits the most pronounced growth trajectory, surpassing LIME and Captum in adoption. The trend suggests a rising interest in post-hoc explainability techniques, with certain repositories gaining significant traction over time.

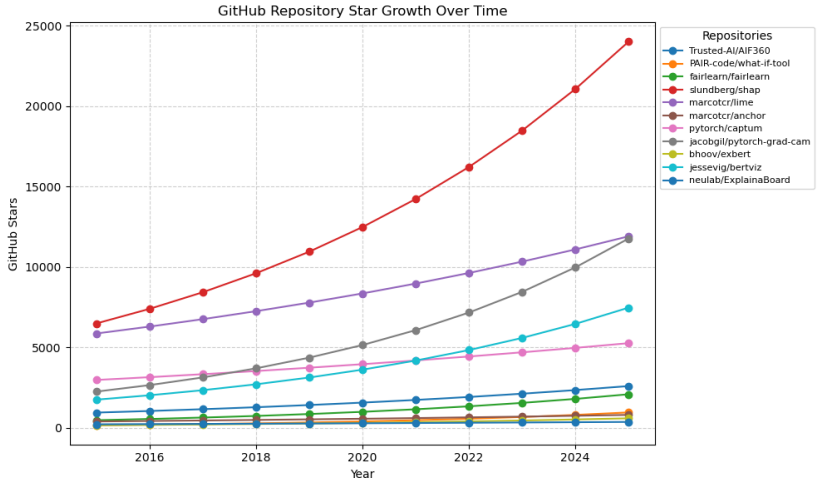

Figure 4: GitHub Repository Star Growth Over Time for Explainability Techniques

8 Conclusion

In this paper, we review the existing literature and provide a comprehensive analysis of XAI, with a focus on its applications in healthcare and cancer diagnostics, while also highlighting the emerging role of LLMs enhancing AI interpretability and user-centered explanations. This study explored the inherent interpretability challenges of DL models, clarifying why they are often described as 'black boxes. We discussed the limitations and challenges associated with current XAI methods, particularly in providing clear and meaningful explanations to end users. By leveraging the tools discussed in this study, practitioners can build interpretable models that promote the responsible and widespread adoption of AI in sensitive and high-impact domains such as healthcare. We highlight the critical need for trust between humans and AI, particularly in medical contexts, where even small errors in model predictions can have severe consequences. Futhermore, we explored the transformative potential of integrating LLMs into XAI systems, particularly in the context of AI-driven breast cancer diagnosis. Recent developments in the filed of interpretable machine learning, particularly in local interpretation methods, provide insights into the decision-making process of complex models by explaining individual predictions. It is crucial to explore approaches that make these explanations more accessible and comprehensible to a wider range of stakeholders, ensuring the effective translation of AI insights into actionable and understandable information. In future work, we will address the need to integrate visual tools with textual explanations, enabling end users to better understand the critical regions of an image by visualizing them using techniques such as saliency maps or heatmaps, ultimately enhancing transparency and trust in AI systems.

References

Adadi, A., Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai), *IEEE Access* **6**, 52138–52160. http://doi.org/10.1109/ACCESS.2018.2870052

Ai, Q., Narayanan. R, L. 2021. Model-agnostic vs. model-intrinsic interpretability for explainable product search, *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 5–15.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., Atkinson, P. M. 2021. Explainable artificial intelligence: an analytical review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **11**(5), e1424.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A. 2017. Classification of breast cancer histology images using convolutional neural networks, *PloS one* **12**(6), e0177544.

Aziz, N. A., Manzoor, A., Mazhar Qureshi, M. D., Qureshi, M. A., Rashwan, W. 2024. Unveiling explainable ai in healthcare: Current trends, challenges, and future directions, *medRxiv* pp. 2024–08.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai, *Microsoft, Tech. Rep. MSR-TR-2020-32* .

Blow, C. H., Qian, L., Gibson, C., Obiomon, P., Dong, X. 2024. Comprehensive validation on reweighting samples for bias mitigation via aif360, *Applied Sciences* **14**(9), 3826.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., Jemal, A. 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: a cancer journal for clinicians* **68**(6), 394–424.

Briola, E., Nikolaidis, C. C., Perifanis, V., Pavlidis, N., Efraimidis, P. 2024. A federated explainable ai model for breast cancer classification, p. 194 – 201.
http://doi.org/10.1145/3655693.3660255

Chaddad, A., Peng, J., Xu, J., Bouridane, A. 2023. Survey of explainable ai techniques in healthcare, *Sensors* **23**(2), 634.

Croce, D., Smirnov, A., Tiburzi, L., Travaglini, S., Costa, R., Calabrese, A., Basili, R., Levialdi Ghiron, N., Melino, G. 2024. Ai-driven transcriptomic encoders: From explainable models to accurate, sample-independent cancer diagnostics, *Expert Systems with Applications* **258**.

Dastgeer, S., Treigys, P. 2024. Transforming black-box models into explainable ai for breast cancer recognition, *DAMSS: 15th conference on data analysis methods for software systems, Druskininkai, Lithuania, November 28-30, 2024.*, Vilniaus universiteto leidykla, pp. 19–20.

Dhiman, B., Kamboj, S., Srivastava, V. 2024. Explainable ai based efficient ensemble model for breast cancer classification using optical coherence tomography, *Biomedical Signal Processing and Control* **91**.
http://doi.org/10.1016/j.bspc.2024.106007

Dihmani, H., Bousselham, A., Bouattane, O. 2024. A new computer-aided diagnosis system for breast cancer detection from thermograms using metaheuristic algorithms and explainable ai, *Algorithms* **17**(10).
http://doi.org/10.3390/a17100462

Egli, A. 2023. Chatgpt, gpt-4, and other large language models: the next revolution for clinical microbiology?, *Clinical Infectious Diseases* **77**(9), 1322–1328.

Farrag, A., Gad, G., Fadlullah, Z. M., Fouda, M. M. 2023. Mammogram tumor segmentation with preserved local resolution: An explainable ai system, p. 314 – 319.
http://doi.org/10.1109/GLOBECOM54140.2023.10436915

Gajbhiye, A., Moubayed, N. A., Bradley, S. 2021. Exbert: An external knowledge enhanced bert for natural language inference, *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, Springer, pp. 460–472.

Gao, Y., Heller, S. L. 2020. Abbreviated and ultrafast breast mri in clinical practice, *Radiographics* **40**(6), 1507–1527.

Geisel, J., Raghu, M., Hooley, R. 2018. The role of ultrasound in breast cancer screening: the case for and against ultrasound, *Seminars in Ultrasound, CT and MRI*, Vol. 39, Elsevier, pp. 25–34.

Goodman, B., Flaxman, S. 2017. European union regulations on algorithmic decision-making and a "right to explanation", *AI magazine* **38**(3), 50–57.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. 2018. A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* **51**(5), 1–42.

Hakkoum, H., Idri, A., Abnane, I. 2024. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data, *Engineering Applications of Artificial Intelligence* **131**, 107829.

Hall, P., Gill, N. 2019. *An introduction to machine learning interpretability*, O'Reilly Media, Incorporated.

Halling-Brown, M. D., Warren, L. M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M. G., Wilkinson, L. S., Given-Wilson, R. M., McAvinchey, R., Young, K. C. 2020. Optimam mammography image database: a large-scale resource of mammography images and clinical data, *Radiology: Artificial Intelligence* **3**(1), e200103.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A. 2024. Interpreting black-box models: a review on explainable artificial intelligence, *Cognitive Computation* **16**(1), 45–74.
https://doi.org/10.1007/s12559-023-10179-8

Institute, N. C. 2025. The cancer genome atlas (tcga). Accessed: 2025-01-26.
https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Karthiga, R., Narasimhan, K., Amirtharajan, R. et al. 2024. Review of ai & xai-based breast cancer diagnosis methods using various imaging modalities, *Multimedia Tools and Applications* pp. 1–52.

Kaushik, R., Sivaselvan, B., Kamakoti, V. 2023. Integrating explainable ai with infrared imaging and deep learning for breast cancer detection, *OCIT 2023 - 21st International Conference on Information Technology, Proceedings*, p. 82 – 87.
http://doi.org/10.1109/OCIT59427.2023.10431160

Khater, T., Hussain, A., Mahmoud, S., Yasen, S. 2023. Explainable ai for breast cancer detection: A lime-driven approach, *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, p. 540 – 545.
http://doi.org/10.1109/DeSE60595.2023.10469341

Kuhl, C. K. 2024. Abbreviated breast mri: state of the art, *Radiology* **310**(3), e221822.

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., Séroussi, B. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, *Artificial Intelligence in Medicine* **94**, 42–53.
https://doi.org/10.1016/j.artmed.2019.01.001

Letham, B., Rudin, C., McCormick, T. H., Madigan, D. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, *Advances in Neural Information Processing Systems* **36**, 28541–28564.

Lundberg, S., Lee, S. 2017. Advances in neural information processing systems. 2017, *A unified approach to interpreting model predictions* pp. 4765–4774.

Maheswari, B. U., Aaditi, A., Avvaru, A., Tandon, A., De Prado, R. P. 2024. Interpretable machine learning model for breast cancer prediction using lime and shap.
http://doi.org/10.1109/I2CT61223.2024.10543965

Malhi, A., Främling, K. 2023. An evaluation of contextual importance and utility for outcome explanation of black-box predictions for medical datasets, *Communications in Computer and Information Science* **1901 CCIS**, 544 – 557.
http://doi.org/10.1007/978-3-031-44064-9_29

Mangasarian, O. L., Street, W. N., Wolberg, W. H. 1995. Breast cancer diagnosis and prognosis via linear programming, *Operations research* **43**(4), 570–577.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A. et al. 2020. International evaluation of an ai system for breast cancer screening, *Nature* **577**(7788), 89–94.

Mohamed, S. K., Sakr, N. A., Hikal, N. A. 2021. A review of breast cancer classification and detection techniques, *International Journal of Advanced Science Computing and Engineering* **3**(3), 128–139.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., Cardoso, J. S. 2012. Inbreast: toward a full-field digital mammographic database, *Academic radiology* **19**(2), 236–248.

Moxey, A., Robertson, J., Newby, D., Hains, I., Williamson, M., Pearson, S.-A. 2010. Computerized clinical decision support for prescribing: provision does not guarantee uptake, *Journal of the American Medical Informatics Association* **17**(1), 25–33.

Paudel, P., Saud, R., Karna, S. K., Bhandari, M. 2023. Determining the major contributing features to predict breast cancer imposing ml algorithms with lime and shap. http://doi.org/10.1109/ICECET58911.2023.10389217

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., Przybocki, M. A. 2020. Four principles of explainable artificial, *Technical report*, NIST Interagency/Internal Report (NISTIR). Gaithersburg, MD: National ....

Radensky, M., Downey, D., Lo, K., Popovic, Z., Weld, D. S. 2022. Exploring the role of local and global explanations in recommender systems, *Chi conference on human factors in computing systems extended abstracts*, pp. 1–7.

Raghavan, K., B, S., v, K. 2024. Attention guided grad-cam : an improved explainable artificial intelligence model for infrared breast cancer detection, *Multimedia Tools and Applications* **83**(19), 57551 – 57578. http://doi.org/10.1007/s11042-023-17776-7

Raghavan, K., Balasubramanian, S., Veezhinathan, K. 2024. Explainable artificial intelligence for medical imaging: Review and experiments with infrared breast images, *Computational Intelligence* **40**(3). http://doi.org/10.1111/coin.12660

Rahman, M. S., Khomh, F., Hamidi, A., Cheng, J., Antoniol, G., Washizaki, H. 2023. Machine learning application development: practitioners' insights, *Software Quality Journal* **31**(4), 1065–1119.

Rai, A. 2020. Explainable ai: From black box to glass box, *Journal of the Academy of Marketing Science* **48**, 137–141.

Rajpal, S., Rajpal, A., Saggar, A., Vaid, A. K., Kumar, V., Agarwal, M., Kumar, N. 2023. Xaimethylmarker: Explainable ai approach for biomarker discovery for breast cancer subtype classification using methylation data, *Expert Systems with Applications* **225**. http://doi.org/10.1016/j.eswa.2023.120130

Ramos-Pollán, R., Guevara-López, M. Á., Oliveira, E. 2012. A software framework for building biomedical machine learning classifiers through grid computing resources, *Journal of medical systems* **36**, 2245–2257.

Retzlaff, C. O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M. E., Holzinger, A. 2024. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities, *Journal of Artificial Intelligence Research* **79**, 359–415.

Ribeiro, M. T., Singh, S., Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Ribeiro, M. T., Singh, S., Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Rodrigues, P. S. 2017. Breast ultrasound image, *Mendeley Data* **1**(10).

Rose, C., Turi, D., Williams, A., Wolstencroft, K., Taylor, C. 2006. Web services for the ddsm and digital mammography research, *Digital Mammography: 8th International Workshop, IWDM 2006, Manchester, UK, June 18-21, 2006. Proceedings 8*, Springer, pp. 376–383.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature machine intelligence* **1**(5), 206–215.

Sabol, P., Sincak, P., Ogawa, K., Hartono, P. 2019. Explainable classifier supporting decision-making for breast cancer diagnosis from histopathological images, Vol. 2019-July. http://doi.org/10.1109/IJCNN.2019.8852070

Sadeghi, Z., Alizadehsani, R., CIFCI, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S. et al. 2024. A review of explainable artificial intelligence in healthcare, *Computers and Electrical Engineering* **118**, 109370.

Saha, A., Harowicz, M. R., Grimm, L. J., Weng, J., Cain, E., Kim, C., Ghate, S., Walsh, R., Mazurowski, M. A. 2021. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set], *The Cancer Imaging Archive* .

Schaffter, T., Buist, D. S., Lee, C. I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S. et al. 2020. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, *JAMA network open* **3**(3), e200265–e200265.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shi, H., Yang, D., Tang, K., Hu, C., Li, L., Zhang, L., Gong, T., Cui, Y. 2022. Explainable machine learning model for predicting the occurrence of postoperative malnutrition in children with congenital heart disease, *Clinical Nutrition* **41**(1), 202–210. `https://doi.org/10.1016/j.clnu.2021.11.006`

Spanhol, F. A., Oliveira, L. S., Petitjean, C., Heutte, L. 2015. A dataset for breast cancer histopathological image classification, *Ieee transactions on biomedical engineering* **63**(7), 1455–1462.

Stanchi, O., Ronchetti, F., Quiroga, F. 2023. The implementation of the rise algorithm for the captum framework, *Conference on Cloud Computing, Big Data & Emerging Topics*, Springer, pp. 91–104.

Suckling, J. 1994. The mammographic images analysis society digital mammogram database, *Exerpta Medica. International Congress Series, 1994*, Vol. 1069, pp. 375–378.

Sultania, M., Kataria, K., Srivastava, A., Misra, M. C., Parshad, R., Dhar, A., Hari, S., Thulkar, S. 2017. Validation of different techniques in physical examination of breast, *Indian Journal of Surgery* **79**, 219–225.

Sundararajan, M., Taly, A., Yan, Q. 2017. Axiomatic attribution for deep networks, *International conference on machine learning*, PMLR, pp. 3319–3328.

Swartout, W. R. 1985. Explaining and justifying expert consulting programs, *Computer-assisted medical decision making*, Springer, pp. 254–271.

Thakur, N., Kumar, P., Kumar, A. 2024. A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities, *Multimedia Tools and Applications* **83**(12), 35849–35942.

Varshney, K. 2018. Introducing ai fairness 360, *IBM Research blog* .

Vig, J. 2019. Bertviz: A tool for visualizing multihead self-attention in the bert model, *ICLR workshop: Debugging machine learning models*, Vol. 3.

Welch, H. G., Prorok, P. C., O'Malley, A. J., Kramer, B. S. 2016. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness, *New England Journal of Medicine* **375**(15), 1438–1447.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J. 2019. The what-if tool: Interactive probing of machine learning models, *IEEE transactions on visualization and computer graphics* **26**(1), 56–65.

WHO 2024. No-one should face breast cancer alone, *https://www.who.int/news-room/fact-sheets/detail/breast-cancer* .

Wikle, C. K., Datta, A., Hari, B. V., Boone, E. L., Sahoo, I., Kavila, I., Castruccio, S., Simmons, S. J., Burr, W. S., Chang, W. 2023. An illustration of model agnostic explainability methods applied to environmental data, *Environmetrics* **34**(1), e2772.

Williams, C. Y., Zack, T., Miao, B. Y., Sushil, M., Wang, M., Kornblith, A. E., Butte, A. J. 2024. Use of a large language model to assess clinical acuity of adults in the emergency department, *JAMA Network Open* **7**(5), e248895–e248895.

Yuan, W., Neubig, G., Liu, P. 2021. Bartscore: Evaluating generated text as text generation, *Advances in neural information processing systems* **34**, 27263–27277.

Zhang, Y., Weng, Y., Lund, J. 2022. Applications of explainable artificial intelligence in diagnosis and surgery, *Diagnostics* **12**(2), 237.

Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D. et al. 2024. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities, *IEEE Communications Surveys & Tutorials* .

Received June 17, 2025 , accepted June 27, 2025