Environment. Technology. Resources. Rezekne, Latvia
Proceedings of the 16th International Scientific and Practical Conference. 2025. Volume II, 77-82

Evaluation of the Diagnostic Accuracy of Two AI Models for Analyzing Mammographic Images

Bora Dogan
Medical University of Varna „Prof. Dr. Paraskev Stoyanov“
Varna, Bulgaria
bdogan008@gmail.com

Yanka Baneva
Medical University of Varna „Prof. Dr. Paraskev Stoyanov“
Varna, Bulgaria
yanysh@abv.bg

Yordanka Eneva
Medical University of Varna „Prof. Dr. Paraskev Stoyanov“
Varna, Bulgaria
stemivada@abv.bg

Elena Ivanova
Medical Imaging Clinic
University Hospital St. Marina
Varna, Bulgaria
ed.ivanova@abv.bg

***Abstract*—Early diagnosis of breast cancer is crucial for improving prognosis and reducing mortality. Mammographic screening is a standard method for detecting malignant lesions, but its effectiveness depends on accurate interpretation and clear communication with patients. In recent years, artificial intelligence (AI) has shown promising results in medical diagnostics, yet its ability to analyze mammographic images and provide reliable patient guidance remains insufficiently studied. This pilot study evaluates the diagnostic accuracy and informational quality of two AI models—ChatGPT and X-Ray Insight. To assess their performance, we use a standardized database of mammographic images, including benign and malignant findings. Additionally, patient inquiry scenarios are used to evaluate the AI-generated explanations, recommendations, and overall clarity in response to common patient concerns. The results are compared against expert radiologist assessments. The AI models are evaluated based on diagnostic accuracy, clarity and correctness of explanations, usefulness of recommendations, and potential risks. Both models performed equally in determining the BI-RADS category, Differential Diagnosis, and Next Clinical Steps. The study underscores the challenges of integrating generative AI models into medical diagnostics, revealing both accurate insights and notable inaccuracies compared to expert interpretations. These discrepancies highlight the need for further training and refinement to improve the reliability and applicability of AI in clinical settings.**

Keywords*—*artificial intelligence, ChatGPT, mammography, X-Ray Insight.

I. INTRODUCTION

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide, accounting for approximately 460,000 deaths annually [1]. Despite advances in early detection and treatment, it continues to pose a significant public health challenge. The development of breast cancer is influenced by a combination of genetic, hormonal, and lifestyle factors. Several key risk factors are identified that increase a woman's likelihood of developing the disease: genetic and family history, personal medical history – women who have previously been diagnosed with breast cancer face a greater risk of recurrence or developing cancer in the other breast, breast tissue density, hormonal and reproductive factors, such as early onset of menstruation and late menopause, result in prolonged exposure to estrogen and progesterone, increasing breast cancer risk. While breast cancer is a significant health concern, early detection through regular screening plays a crucial role in improving survival rates. Mammograms, clinical breast exams, and self-examinations help identify cancer at an early stage when treatment is most effective. Tumor size, proliferative activity, and lymph node metastasis are fundamental in defining treatment plans and patient prognosis [2], [3]. Mammography is the gold standard for breast cancer screening, used extensively to detect early signs of breast cancer, often before clinical symptoms appear [4]. Despite its efficacy, mammography interpretation is a challenging task. False positives and false negatives can occur, leading to unnecessary biopsies or missed diagnoses. Interpretation of mammograms is inherently subjective, influenced by factors like the radiologist's experience, the quality of the mammogram, and the complexity of the tissue being examined. Studies have shown that even experienced

Online ISSN 2256-070X
https://doi.org/10.17770/etr2025vol2.8572

radiologists can miss certain abnormalities, especially in dense breast tissue [5].

Artificial intelligence (AI) has the potential to transform medical diagnostics by enhancing the accuracy and efficiency of diagnostic processes [6], [7]. AI models can be trained on vast datasets of mammograms and learn to identify complex patterns in images, enabling them to detect even subtle abnormalities that might be overlooked by human eyes. AI can assist radiologists by providing a second opinion, reducing the risk of errors, and enhancing productivity [8]. Additionally, AI systems can help prioritize cases, flagging high-risk images for immediate attention, thus improving workflow efficiency. AI models can analyze mammogram images with great detail, examining microcalcifications, masses, and architectural distortions that indicate potential tumors or abnormalities [9]. This is especially important in cases of dense breast tissue, where traditional mammograms might struggle to identify lesions.

II. Materials and methods

This pilot study is a comparative evaluation of the diagnostic accuracy of two AI models, ChatGPT-4o and X-Ray Insight, in analyzing mammographic images.

The goal was to assess their ability to detect, classify, and provide differential diagnoses for breast abnormalities compared to ground truth diagnoses and expert radiologist interpretations.

Mammographic images were sourced from Radiopaedia [10] – [30], a trusted repository of medical imaging cases. To identify and select images relevant to this study, specific filters were applied within the "Cases" section: the system was set to "breast," and the study modality was restricted to "mammography." The search results were then sorted by case completion percentage, and only cases with a high diagnostic certainty were chosen. A total of 20 clinical cases comprising mammographic images were selected. These cases represented a broad spectrum of breast pathologies, ensuring a diverse and comprehensive dataset for analysis (Table 1).

Table 1 Summary of clinical case characteristics

Category	Details
Pathology Type	Malignant: 14 (70%), Benign: 6 (30%)
Patient Gender	Female: 18 (90%), Male: 2 (10%)
Age Range	40 to 80 years
Imaging View	CC: 17 (85%), MLO: 3 (15%)

Of the 20 cases, 70% (14/20) represented malignant pathologies, while 30% (6/20) were benign. The cases included 90% (18/20) female patients and 10% (2/20) male patients, with an age range of 40 to 80 years. In terms of imaging views, 85% (17/20) were craniocaudal (CC) views, and 15% (3/20) were mediolateral oblique (MLO) views. All of them were 2D mammograms.

Next, the following study prompt was carefully crafted to be entered into the AI models:

“This is a [2D/3D] mammogram in the [CC/MLO view] of a [age]-year-old [gender] patient, presenting with [clinical presentation, e.g., screening, palpable lump, nipple discharge]. Please describe the findings in the image, including masses (size, shape, margins, density), calcifications (morphology, distribution), architectural distortions, asymmetries, or other abnormalities. Provide a differential diagnosis, listing the most likely diagnosis first, and justify your reasoning. Assess the ACR breast density (Category A, B, C, or D) and assign a BI-RADS category (0-6), explaining your rationale. Estimate the likelihood of malignancy (low, intermediate, high) and recommend next steps for the physician, such as additional imaging, biopsy, or follow-up intervals."

The prompt's design was critical, as it needed to reflect real-world clinical workflows to ensure practical applicability. It incorporated key elements essential for mammogram interpretation, including ACR density and BI-RADS classification. Structured to guide the AI through a systematic analysis, the prompt segmented the task into specific components: findings, differential diagnosis, ACR density, BI-RADS assessment, next steps, and malignancy likelihood. This approach ensured that AI models provided comprehensive and methodical responses. Additionally, the prompt encouraged AI to justify its assessments with rationales and supporting evidence, a crucial factor in evaluating the model’s reasoning capabilities. To maintain consistency and facilitate fair comparisons across AI models and datasets, the prompt was standardized.

The final step involved inputting the prompt, along with a single mammographic image from each case, into a new chat session within the AI models. This method was essential for several reasons. First, it ensured that AI evaluated each image independently, free from influence or bias from prior cases—mirroring real-world clinical practice, where radiologists assess each case on its own merits. Second, starting a new session prevented the AI from retaining context from previous interactions, which could otherwise skew results and compromise evaluation fairness. Finally, this approach maintained consistency throughout the study, enabling accurate comparisons of AI performance across different cases and models. By isolating each case, the study upheld scientific rigor and ensured the reliability of its findings.

The AI-generated responses, which were recorded in a shared document, and the relevant portions were organized into a table to facilitate comparison and evaluation. The radiologist’s remarks for each case were also included in the comparison to ensure a balanced assessment. Discrepancies between AI models, radiologist, and ground truth diagnoses were analyzed to assess AI diagnostic accuracy, sensitivity, specificity, and clinical reliability.

In this study, we use ChatGPT-4o of OpenAI and X-Ray Insight, developed by AiWebTools.Ai (Table 2). It is a custom version of ChatGPT that is specifically designed to assist in interpreting medical images and supporting doctors in diagnosis. ChatGPT-4o is an AI model based on natural language processing (NLP). It is not trained for direct analysis of mammographic images and uses a pre-trained medical knowledge base to provide insights into mammography reports. On the other hand, X-Ray Insight is an image-based deep learning model that specializes in analyzing x-ray images, i.e. mammographic images. It can classify findings based on the BI-RADS categories, identify suspicious lesions, and suggest next steps.

TABLE 2 COMPARISON OF THE TWO AI MODELS

Feature	ChatGPT-4o	X-Ray Insight
Input Type	Text, images, and files	Text, images, and files
Processing Method	NLP (Natural Language Processing)	Designed to interpret x-ray images
Can Identify Tumors?	Yes	Yes
Provides BI-RADS Classification?	Yes	Yes
Suggests Clinical Management?	Yes (biopsy, MRI, follow-up)	Yes (based on detected abnormalities)
Main Limitation	Not specifically designed for interpreting x-ray images	Lacks full clinical context, potential overdiagnosis
Best Use Case	Assisting with report analysis and decision support	Detecting abnormalities in screening mammograms

III. RESULTS AND DISCUSSION

To assess AI model performance, we compared their outputs against ground truth diagnoses and radiologist interpretations using the following metrics: diagnostic accuracy, sensitivity, specificity , BI-RADS classification agreement with ground truth, false positives, false negatives, agreement with radiologist's assessments (%). A board-certified radiologist independently reviewed AI outputs and rated the accuracy, relevance, and clinical usability of the AI-generated findings. AI results were compared to pathology-confirmed diagnoses. Discrepancies were analyzed to identify common misclassification trends. 20 cases were reported, and Fig.1 includes:

1. **Findings Description:** AI-generated mammographic interpretation from ChatGPT-4o, AI-generated mammographic interpretation from X-Ray Insight, Ground truth diagnosis based on pathology-confirmed findings.
2. **Radiologist's Assessment of AI Interpretations:** The radiologist reviewed and scored the accuracy of the findings provided by both AI models, errors were noted, including missed abnormalities, incorrect lesion descriptions, and misclassified BI-RADS categories.
3. **BI-RADS Classification Comparison:** AI-assigned BI-RADS categories, Ground truth BI-RADS classification for each case, the radiologist's assessment of whether AI overestimated, underestimated, or correctly assigned BI-RADS levels.
4. **Malignancy Likelihood Estimation:** AI models provided malignancy probability estimates based on findings. The radiologist assessed whether AI was overestimated, underestimated, or correctly estimated malignancy risk compared to pathology.
5. **Recommended Next Steps:** AI models suggested clinical management plans, Ground truth clinical recommendations were included for comparison, and the radiologist evaluated whether AI recommendations were appropriate or unnecessary.

Both ChatGPT-4o and X-Ray Insight often provided similar interpretations (table 3). However, errors occurred in identifying key findings such as: Missing architectural distortion (a crucial early malignancy sign), Incorrectly classifying breast density (ACR categories), Differences in malignancy likelihood estimations (e.g., ChatGPT often reported ≥50%, while X-Ray Insight went for ≥95%). Doctor's interpretations of AI outputs varied, some cases showed consensus with AI findings, while others had significant discrepancies, where AI misclassified benign vs. malignant lesions. For the BI-RADS classification consistency, BI-RADS 4 and 5 cases (high suspicion for malignancy) were more consistently identified while BI-RADS 2 and 3 cases (benign or probably benign findings) had more misinterpretations

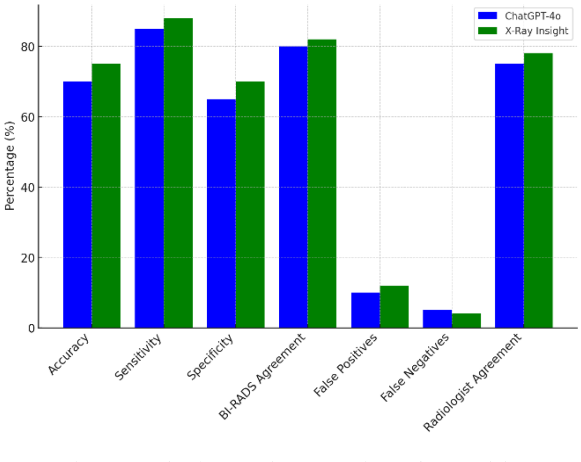

Fig. 1. Evaluation metrics comparison of AI models.

AI models misidentified normal post-surgical scars as potential malignancies. Some benign conditions (e.g., gynecomastia, hamartomas) were flagged as potentially malignant.

Among the 20 cases analyzed, the most common errors were: missed architectural distortion in 3 cases, overestimated malignancy in 4 cases, incorrect BI-RADS classification in 5 cases, missed lymph node involvement in 2 cases.

ChatGPT-4o and X-Ray Insight had an accuracy of ~5-6% when compared strictly to the Ground truth.

Doctors agreed with AI models in ~64% of cases but had disagreements in specific borderline cases.

TABLE 3 THE COMPARISON BETWEEN CHATGPT-4O AND X-RAY INSIGHT BASED ON THEIR PERFORMANCE ACROSS TWENTY CASES

Metric	ChatGPT-4o	X-Ray Insight
Overall Accuracy (%)	65%	70%
BI-RADS 4-5 Accuracy (%)	85%	87%
BI-RADS 1-3 Accuracy (%)	50%	55%
False Positives (Overestimated Risk)	4 cases	5 cases
False Negatives (Missed Malignancies)	2 cases	1 case
Doctor Agreement (%)	64%	66%
Correct Next Step Recommendation (%)	85%	88%

ChatGPT-4o and X-Ray Insight had an accuracy of ~5-6% when compared strictly to the Ground truth. Doctors agreed with AI models in ~64% of cases but had disagreements in specific borderline cases.

For the Findings Description, AI models detected abnormalities correctly in ~60% of cases. They accurately described tumor shape, density, and calcifications, while missing key features in ~40% of cases.

AI models have a high ability to identify the correct pathology. Almost ~75% of the cases were correctly listed by the models. Difficulty appeared when distinguishing rare conditions and post-surgical changes from malignancy. For the AI models' performance in categorizing breast tissue, the agreement with the Ground truth is ~80%. For the malignancy likelihood estimation, AI risk estimation was more aligned with human experts in high-suspicion cases but overcalled risk in borderline lesions.

Both AI models generally provided the correct Next steps for the treatment recommendations but sometimes failed in triaging lower-risk cases.

Additionally, the AI model's responses for each clinical case were analyzed across six components: Findings Description, Differential Diagnosis, ACR Density, BI-RADS, Malignancy Likelihood, and Next Steps. The radiologist evaluated each component across all cases, assigning scores on a scale from 2 to 6, with 2 representing the lowest grade and 6 the highest.

Fig. 2 illustrates the percentage distribution of individual scores assigned to ChatGPT-4o and X-Ray Insight across all components of the clinical cases. The score distribution indicates that X-Ray Insight received a higher proportion of top scores (6) at 57.50%, compared to 50.83% for ChatGPT-4o, suggesting a stronger overall performance. Lower scores (3 and 2) were relatively infrequent for both models, though ChatGPT-4o had a higher percentage of the lowest score (2) at 13.33% compared to 7.50% for X-Ray Insight, suggesting occasional inconsistencies.

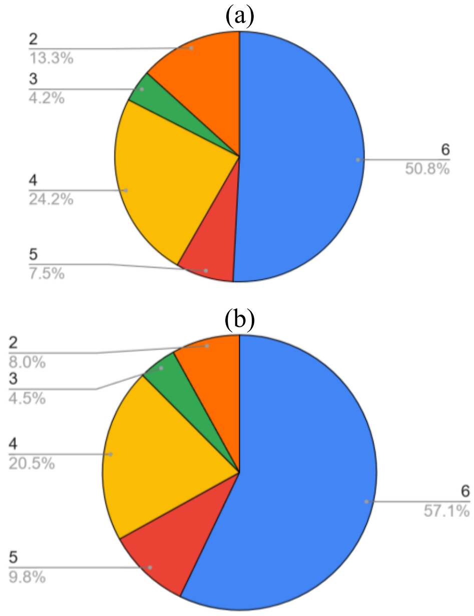

Fig. 2. Percentage distribution of scores assigned to ChatGPT-4o (a) and X-Ray Insight (b)

Fig. 3 presents the average scores assigned to AI models by the radiologist for each clinical case. X-Ray Insight generally scored higher in malignancy-related cases such as Invasive Ductal Carcinoma (IDC), Multifocal Breast Cancer, and Plasma Cell Mastitis, aligning with its strength in Findings Description and Malignancy Likelihood. ChatGPT-4o performed better in Ductal Carcinoma in Situ (DCIS), Invasive Solid Papillary Breast Carcinoma, and Axillary Tail Breast Cancer, suggesting potential advantages in interpretative aspects of ACR Density. Some cases, including Fibroadenoma, Fibrocystic Disease, and Invasive Breast Carcinoma, showed equal performance, indicating areas where AI models provide comparable insights.

IV. CONCLUSIONS

AI models perform well in high-suspicion cancer cases (BI-RADS 4-5) but struggle with low-risk or benign cases. AI is useful as a second opinion tool but cannot yet replace radiologists. Training AI on more diverse mammography datasets could improve its accuracy, especially for borderline cases.

ChatGPT-4o is best for interpreting mammography reports and helping doctors make decisions based on text-based data. X-Ray Insight is ideal for image-based detection, helping radiologists spot potential cancers directly from mammograms. Combining the two AI models could create a hybrid system, where text-based insights and image-based detections work together for more accurate diagnoses.

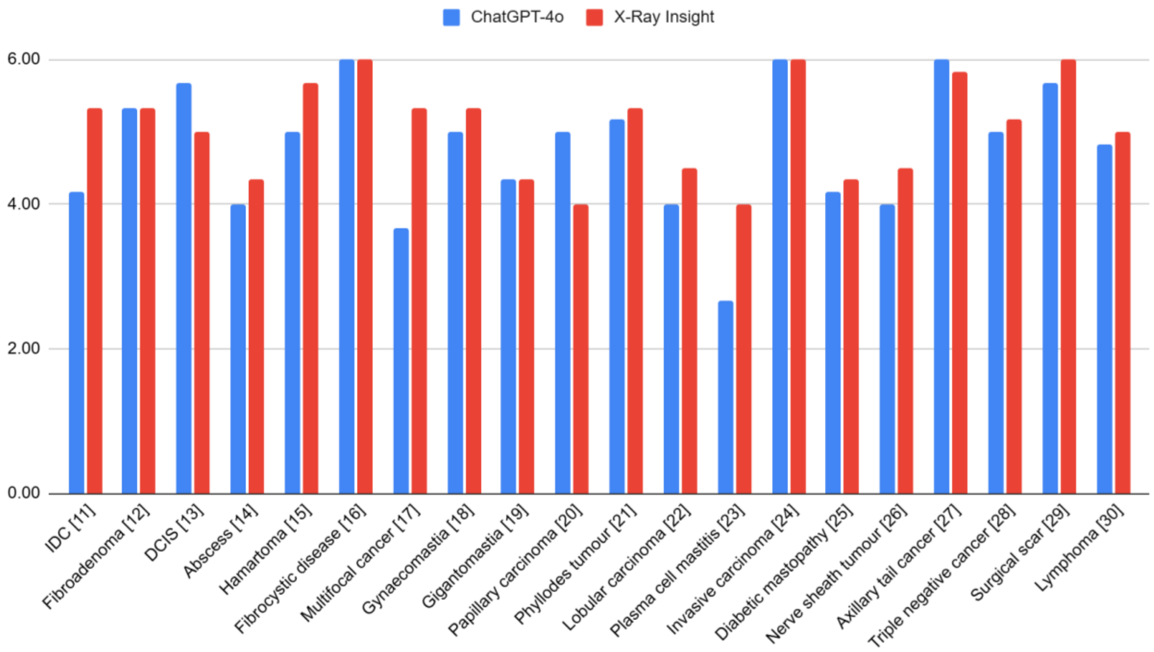

Fig. 3. Radiologist-assigned average scores, ranging from 2 to 6, for ChatGPT-4o and X-Ray Insight for each clinical case.

The integration of AI into mammography is a game-changer for breast cancer detection, improving accuracy, efficiency, and early detection rates. By automating routine analysis, providing a second opinion, reducing human error, and supporting radiologists in decision-making, AI is enhancing the ability to detect breast cancer earlier, which is key to improving patient outcomes. As AI technology continues to evolve, its integration with multi-modal imaging, its ability to learn from diverse data, and its potential to reduce health disparities will only continue to make mammography a more effective tool in the fight against breast cancer.

X-Ray Insight provides results without sufficient justification, limiting its use for learning while with ChatGPT-4o medical students can input case descriptions and get step-by-step diagnostic reasoning, and may receive an explanation of why a case falls into a certain BI-RADS category.

V. References

[1] G. Dimitrov, I. Gavrilov, and T. Sedloev, Eds., *Breast Cancer*. Paradigm, 2014.

[2] V. Parvanova, *Breast Cancer - Early Detection and Modern Treatment*, Sofia: Tip - Top Press, 2006.

[3] T. Deliyski and D. Damyanov, *Recommendations for Complex Treatment of Breast Cancer*, Pleven: University Hospital-Pleven EAD, 2005.

[4] N. Day and R. Warren, "Mammographic screening and mammographic patterns," *Breast Cancer Res.*, vol. 2, no. 4, pp. 247-251, 2000, doi: 10.1186/bcr64.

[5] E. U. Ekpo, M. Alakhras, and P. Brennan, "Errors in Mammography Cannot be Solved Through Technology Alone," *Asian Pac. J. Cancer Prev.*, vol. 19, no. 2, pp. 291-301, 2018, doi: 10.22034/APJCP.2018.19.2.291.

[6] Y. Eneva and B. Dogan, "A Qualitative Assessment of Medical Diagnosis Capabilities of Three Artificial Intelligence Models: ChatGPT-4o, CodyMD, and Dr. Gupta," *J. Biosci. Med.*, vol. 12, pp. 243-254, 2024, doi: 10.4236/jbm.2024.1211021.

[7] Y. Eneva and B. Dogan, "Evaluation of Medical Diagnosis Capabilities of Three Artificial Intelligence Models – ChatGPT-3.5, Google Gemini, Microsoft Copilot: Sustainable Development Goals (SDGs)," *J. Lifestyle and SDGs Review*, vol. 5, no. 2, e03545, 2025, doi: 10.47172/2965-730X.SDGsReview.v5.n02.pe03545.

[8] D. Killock, "AI outperforms radiologists in mammographic screening," *Nat. Rev. Clin. Oncol.*, vol. 17, no. 3, p. 134, 2020, doi: 10.1038/s41571-020-0329-7.

[9] S. Al Muhaisen, O. Safi, A. Ulayan, et al., "Artificial Intelligence-Powered Mammography: Navigating the Landscape of Deep Learning for Breast Cancer Detection," *Cureus*, Mar. 2024, doi: 10.7759/cureus.56945.

[10] Radiopaedia.org, "The peer-reviewed collaborative Radiology Resource," [Online]. Available: https://radiopaedia.org/. [Accessed: Mar. 1, 2025].

[11] H. Al Ja'afreh, "Invasive ductal carcinoma," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-174470. [Accessed: Feb. 3, 2025].

[12] A. Ashraf, "Giant fibroadenoma," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-90946. [Accessed: Feb. 3, 2025].

Online ISSN 2256-070X
https://doi.org/10.17770/etr2025vol2.8572

[13] W. Lee, "Ductal carcinoma in situ," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-160609. [Accessed: Feb. 3, 2025].

[14] A. Ruiz Gaviria, "Breast abscess - male," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-163546. [Accessed: Feb. 3, 2025].

[15] V. Pai, "Hamartoma of the breast," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-27023. [Accessed: Feb. 7, 2025].

[16] M. Altintakan, "Multiple fibroadenomas and fibrocystic disease," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-75028. [Accessed: Feb. 7, 2025].

[17] M. Niknejad, "Multifocal breast cancer - 2D mammography and tomosynthesis," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-156732. [Accessed: Feb. 7, 2025].

[18] A. Abougazia, "Gynaecomastia," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-23657. [Accessed: Feb. 7, 2025].

[19] A. Ashraf, "Gigantomastia," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-150940. [Accessed: Feb. 7, 2025].

[20] W. Lee, "Invasive solid papillary carcinoma of the breast," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-161313. [Accessed: Feb. 7, 2025].

[21] A. Ashraf, "Malignant phyllodes tumour," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-89374. [Accessed: Feb. 7, 2025].

[22] S. Sorrentino, "Invasive lobular carcinoma," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-34821. [Accessed: Feb. 8, 2025].

[23] A. Ranchod, "Plasma cell mastitis," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-175533. [Accessed: Feb. 8, 2025].

[24] I. Khatatbeh, "Invasive breast carcinoma," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-190669. [Accessed: Feb. 8, 2025].

[25] N. Abidin, "Diabetic mastopathy," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-58607. [Accessed: Feb. 8, 2025].

[26] G. Kruger, "Nerve sheath tumour - breast," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-18778. [Accessed: Feb. 8, 2025].

[27] H. Knipe, "Axillary tail breast cancer," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-43262. [Accessed: Feb. 9, 2025].

[28] A. Abdelrahman, "Triple negative breast cancer," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-78448. [Accessed: Feb. 9, 2025].

[29] G. Kruger, "Surgical scar of the breast," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-22389. [Accessed: Feb. 9, 2025].

[30] G. Baratelli, "Non-Hodgkin lymphoma arising from an intramammary lymph node," *Case study, Radiopaedia.org*, [Online]. Available: https://doi.org/10.53347/rID-42463. [Accessed: Feb. 9, 2025].