Meta Analysis of AI Performance in Diagnosing Breast Cancer Mammography

SEMERE WUBSHET BERHANU ADELE ROBALDO
swbe | robaldo @kth.se

January 15, 2024

Abstract

Breast cancer remains a significant global health challenge despite advances in medical technology in the previous decades. Artificial Intelligence has emerged as an innovative tool to aid healthcare professionals in diagnosing breast cancer. This research paper is a meta-analysis of AI performance in breast cancer diagnosis from mammography screenings. It produces timely best estimates for the values of AI performance metrics. The four key metrics are sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve.

This study produces and adheres to inclusion criteria and curates a collection of previous papers on specific models from which performance metrics are extracted from. The paper employs standard statistical tools to evaluate them. By collecting a diverse set of papers, the papers aims to show the current capabilities of AI models as well as their limitations. The accuracy of AI models was hypothesized to be above 90% based on the recent successes and breakthroughs in computer vision models.

However, this paper finds that the mean values of sensitivity, specificity, accuracy, and area under the curve are 87%, 80%, 89%, and 88% respectively. This disproves our hypothesis and shows AI models still have to improve before matching the capabilities of human radiologists who have sensitivity and specificity of 87% and 89% respectively. This research can help inform researchers and healthcare professionals on the current status of AI in the field of breast cancer diagnosis.

Future research can address the limitations of this study by evaluating a broader range of AI models as well as those trained on diverse datasets, such as the VinDr-Mammo database from Vietnam and the Chinese Mammography Database from China. These improvements can enhance the statistical significance of the study and overcome potential biases introduced by regional variations.

Contents

List of Acronyms and Abbreviations

AI Artificial Intelligence

AUC Area Under the Curve

BC Breast Cancer

ROC Receiver Operating Characteristic

1 Introduction

1.1 Overview of breast cancer

Breast Cancer (BC) is a condition characterized by the uncontrolled growth of abnormal breast cells, leading to the formation of tumors. BC typically originates within the milk producing lobules of the breast. The earliest form, know as ”In situ”, is not immediately life-threatening. However, if left untreated, these tumors have the potential to metastatize and become fatal [1].

BC is the most common cancer among women worldwide. In 2020, 2.3 million of women were diagnosed with BC and 685,000 have died globally [1].

The exact cause of BC is unknown, but several factors can increase the risk of developing it. Alarmingly, nearly half of all BC cases occur in women who exhibit no specific risk factor for BC. Among the main factors that can increase the risk of BC, there are: alcohol consumption, body mass index, height, lack of physical activity, mammographic density, age at menarche or menopause, smoking, and type 2 diabetes mellitus (T2DM) [2].

Additionally, the Collaborative Group on Hormonal Factors in BC projected that in developed countries, the cumulative incidence of BC could potentially decrease by over 50%, dropping from 6.3 to 2.7 cases per 100 women by the age of 70, if women adopted practices such as having more children and breastfeeding for longer periods, as commonly observed in some developing countries [3].

BC can be categorized into various types depending on the specific cells that are impacted. Among the most prevalent are ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC)[4].

1.2 Introduction to AI applications

Artificial Intelligence (AI) has been a field of study since the 1950s and has gained momentum in the past ten years [5]. This field intends to develop computer systems that are capable of doing intelligent tasks that we humans do. Machine learning is a subset of AI that enables systems to enhance their capabilities by processing large amounts of data [6]. Neural networks are a subset of machine learning and their structure is inspired by the human brain [7]

AI has moved from limited practical interest to public conversation and has entered the business environment [5]. For example, AI can be used in education, assisting students with their studies and providing support. These assistants would available round-the-clock and can be accessed instantly even if they can sometimes give unreliable responses [8].

One of the promising use cases of AI is in the field of healthcare. Healthcare is a vital domain of work that is essential for humanity’s modern way of life. Healthcare providers undertake years of professional training to obtain their professional license. There was an estimated 15 million healthcare worker shortage worldwide in 2020 and this number is estimated to decrease down to a shortage of 10 million in 2030 [9]. AI-enabled systems could assist professionals deliver more precise and tailored quality-healthcare to their patients [10].

1.3 Problem Statement

In recent years, artificial intelligence has been increasingly used to aid in the diagnosis of BC using mammography. However, there are many different AI models with varying performances, and there is a lack of general view of the accuracy of AI in diagnosing BC.

This research aims to address this issue by conducting a meta-analysis of AI performance in diagnosing BC. This meta-analysis is a quantitative review of data from multiple studies, with the aim to provide a comprehensive view of the actual accuracy of AI to this day.

2 Theoretical Framework / Literature Review

2.1 Overview of Breast Cancer Screening

The traditional method of screening for BC relies on mammography, which is a low-dose X-ray imaging technique employed for early BC detection [11]. During a mammogram, the breast is gently compressed between two plates, and an X-ray image is captured. The image is then examined by a radiologist for any signs of BC, such as a lump or mass, calcifications, or distortions in the breast tissue [12]. Regarding the involvement of radiologists in the screening process, some programs utilize a single radiologist to interpret mammograms, whereas others employ two radiologists who independently analyze the mammograms and collaborate to reach a consensus in case of any discrepancies [13].

Mammography is the most commonly utilized screening method for detecting BC and has demonstrated its ability to reduce BC mortality among women aged 50 to 69 years [11]. However, it is essential to acknowledge that mammography, while effective, is not infallible, as it can occasionally miss certain cancers or yield false-positive results.

It is important to note that the effectiveness of mammography screening is influenced by the level of breast density. A study published in the European Journal of Public Health found that the effectiveness of mammography screening decreases with decreasing breast density. The study concluded that the effectiveness of mammography screening is limited in women with low breast density [14].

In the following sections, we will explore further the potential of AI in early BC detection.

2.2 AI in Breast Cancer Diagnosis

Computer vision has leveraged AI and developments in deep learning to rapidly evolve and significantly improve image recognition accuracy [15]. As medical imaging devices digitize, such as film mammography X-rays are getting replaced by digital mammography X-rays, machine learning improvements can better ease the workload for medical experts [16]. Currently, there are several AI models that were developed to detect BC from Digital Mammography. But is it possible to judge their performance? This section will describe how we measure the capabilities of a breast cancer screening AI model.

The performance levels of AI models are measured through four diagnostic metrics. These are sensitivity, specificity, accuracy, and Area Under the Curve (AUC). Sensitivity measures how well the system correctly identifies cases with cancer by dividing the number of true positives divided by all the true cases. Specificity measures how well the system correctly identifies cases without cancer by dividing the number of true negatives by all negatives. Accuracy measures how correctly the model identifies cases with and without cancer by dividing the number of true positives and negatives by the number of values [17].

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad \text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

AI models have thresholds scientists can set to control the trade-off between sensitivity and specificity. The Receiver Operating Characteristic (ROC) curve plots the sensitivity versus the false positive rate, or one minus the specificity, as the threshold for the AI model moves through all possible values. The AUC measures the area under this curve and has values between 0 and 1 [18]. A higher score means the system can better tell between positive and negative cases. The AUCs metric is widely used to compare AI models and radiologists in screening breast cancer. Figure 1 is an example of an ROC and shows the relationship between true positive rates, false positive rates, and the AUC.

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{Positive + Negative}$$

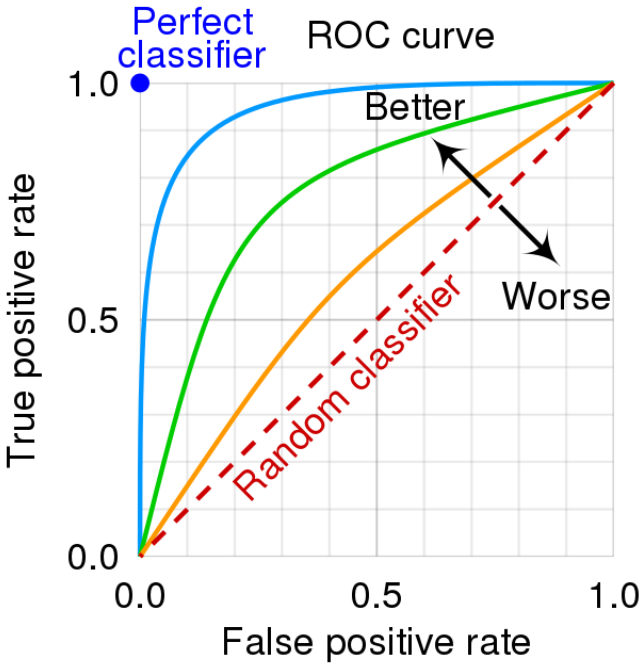

Figure 1: Example ROC Curve for a binary classifier[19]

This research paper shall use the four metrics in assessing the capabilities of existing AI models.

3 Research questions, hypotheses

In this meta-analysis, we aim to evaluate the properties of AI in diagnosis breast cancer from mammography. The main question that drives this research is: given many machine learning models for detection in the medical field, what is the accuracy of the technology today for detecting breast cancer? Our hypothesis is that the accuracy of AI in diagnosing BC is above 90%. To justify this hypothesis, we refer to recent studies that have shown that AI has been improving lately, especially in computer vision. A case of study is the ImageNet challenge, a computer vision competition where participants are tasked with developing algorithms that can classify and detect objects in images [20]. This challenge has been instrumental in advancing the field of computer vision and has led to significant improvements in image classification accuracy. In 2012, the winning team achieved an error rate of 15.3%, which was a significant improvement over the previous year's error rate of 26.2%. Since then, the error rate has continued to decrease, and in 2015, the winning team achieved an error rate of just 3.6% [21]. While the ImageNet challenge is not directly related to BC diagnosis, it is a good example of how AI has been improving in computer vision. Researchers have used deep learning techniques to develop models that can accurately classify images, and these techniques are now being applied to medical imaging, including mammography.

4 Method

The method chosen for the research consist of a meta-analysis. A meta-analysis is a statistical technique for combining data from multiple scientific studies. Meta-analyses are considered the most trustworthy source of evidence by the evidence-based medicine literature [22]. This approach was chosen because is a powerful and rigorous statistical method for synthesizing data from multiple studies that allow to obtain a more accurate estimate of the effect size. In the context of our research question, it is the most suitable to produce a comprehensive overview of performance metrics in AI-based breast cancer diagnosis. The aim is to use approaches from statistics to derive a pooled estimate closest to the unknown common truth based on how this error is perceived. In fact, by combining the results of multiple studies, we increase the sample size and statistical power, which can help reduce the risk of type 2 error [23]. Other types of analysis, such as a narrative review or a case study, would not have been suitable because they lack in objectivity and rigor, and they do not provide a quantitative estimate of the effect size.

While meta-analysis is a valuable tool for evaluating AI-based BC diagnosis, it is important to recognize and address potential methodological challenges. One such challenge is heterogeneity, which arises from the use of different AI models, datasets, and methodologies across studies. To address this issue, we will apply rigorous inclusion criteria to select comparable studies, perform careful data extraction, and employ appropriate statistical methods to assess and adjust for heterogeneity. To enhance the transparency and reproducibility of the analysis, we will provide a clear and detailed description of our study selection process, data extraction methods, statistical analyses, and any potential limitations or biases. We want to write it for researchers and professional healthcare.

4.1 Inclusion Criteria for Research Papers

For the meta-analysis, we performed a selection of data guided by specific inclusion/exclusion criteria. The purpose of defining it was to ensure the integrity, relevance, and reliability of the studies included in the analysis. Studies were elegible for inclusion if they reported the following features:

1. Ground Truth Time frame: to ensure data accuracy, only studies with ground truth established within 12 months before the meta-analysis were included, recognizing the dynamic nature of breast cancer diagnoses.

2. Quantitative Performance Metrics: studies considered in the data analysis reported quantitative data on critical performance metrics such as sensitivity, specificity, accuracy, and AUC. These parameters are fundamental to assess the diagnosis precision of AI-based mammographies.

3. Publication Year: eligible papers, published after 2016, were incorporated to integrate the latest technological advancements in AI-driven breast cancer diagnosis, while excluding outdated methodologies.

4. Sample Size Requirement: studies with a minimum of 200 subjects were considered for inclusion, justified by the specificity, focus and similarity in effect sizes among selected studies, ensuring relevance and reliability [24].

4.2 Statistical methods employed

The main purpose of this paper was to find the values for each AI attribute that best represents this attribute. The metrics that measure the capabilities of an AI model were properly defined, therefore, we needed to employ certain statistical tools to extract representative values from the collected data. It was decided that the best way to do that would be to provide the mean for each attribute and the interval within which there is certain confidence level of its existence. The standard deviation was also included to give the reader information on how the data relates to the best estimates of the four attributes.

These four metrics are all based on the underlying number of cases classified as positive or negative whether correctly or incorrectly. Even though the sensitivity of a model and its specificity are not related, there are known and direct mathematical relationships among the 4 metrics. They all depend on the true positive rate, the true negative rate, or both. Therefore, analyzing the relationships between the metrics is not useful and was not done in this research paper.

The mean was chosen to represent each attribute because the mean is generally considered the best measure of central tendency [25]. There are other statistical measures that could be used to express this central tendency such as the median and the mode. The median is preferred when there are extreme values in the data-set, exist values with undetermined values, is an open ended distribution, or is measurement in an ordinal scale [25]. The data-set that used in this research paper does not contain extreme values nor any missing values. The values were measured from 0 to 1 and measured in a continuous scale. Therefore, taking the median was not an appropriate measure. The mode was the preferred measure when the data was measured in a nominal scale [25]. However, that was not the case for this paper.

The mean gave us a point estimate for each AI attribute that were previously defined. However, this was just an estimate of the true mean value which could vary from our estimate. Therefore, we needed to provide a confidence interval around our estimated mean value. Confidence levels are selected by the researcher according to the particularities of the study and commonly set at 95% confidence interval [26]. Therefore, the confidence level for this study was set to 95% and the confidence intervals were reported.

Apart from the main statistical analysis, graphical visualizations were presented. These aid the reader in understanding how the point estimates and confidence intervals relate to the overall data-set. The histogram of values as well as a box-plot of the data-set were included in this research paper. A box-plot visualizes medians and quartiles in one graph and shows outliers in a data set [27].

The complete description of the hardware and software used in the statistical analysis are provided in the appendix in Table 4 and Table 3.

5 Results

Data from various scientific publications on BC screening AI models were collected before processing. The data collected includes the title, publishing journal, publishing date, the four diagnostic metrics, training data set name and size. After collection, they were processed in accordance with the well defined inclusion criteria. In the end, this paper identified 16 models to use in the study which can be found in the appendix in Table 2.

The collected data were analyzed according to the method described above and the results of the analysis of each AI attributes is presented in Table 1 below.

	Sensitivity	**Specificity**	**Accuracy**	**AUC**
Mean	0.87	0.80	0.89	0.88
Confidence Interval	0.8295 - 0.9187	0.7512 - 0.8528	0.8540 - 0.9256	0.8364 - 0.9188
Standard Deviation	0.073813	0.075588	0.046586	0.071341

Table 1: Means, Confidence Intervals, and Standard Deviations for AI attributes

Figure 2 is a histogram that shows the frequency of occurrence of a value for each of the AI attribute.

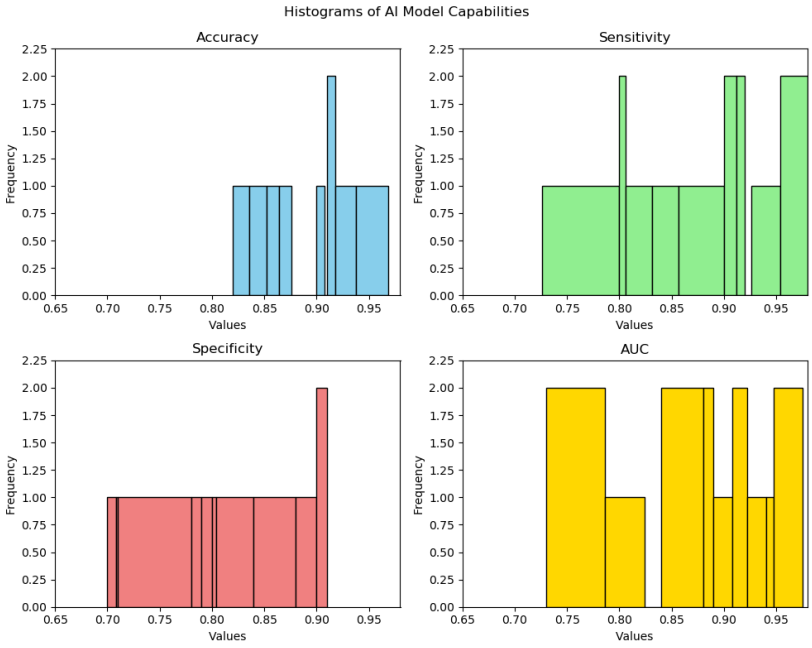

Figure 2: Histogram of AI attributes

Figure 3 is a box plot that shows the medians and quartiles of each of the four AI attribute side by side.

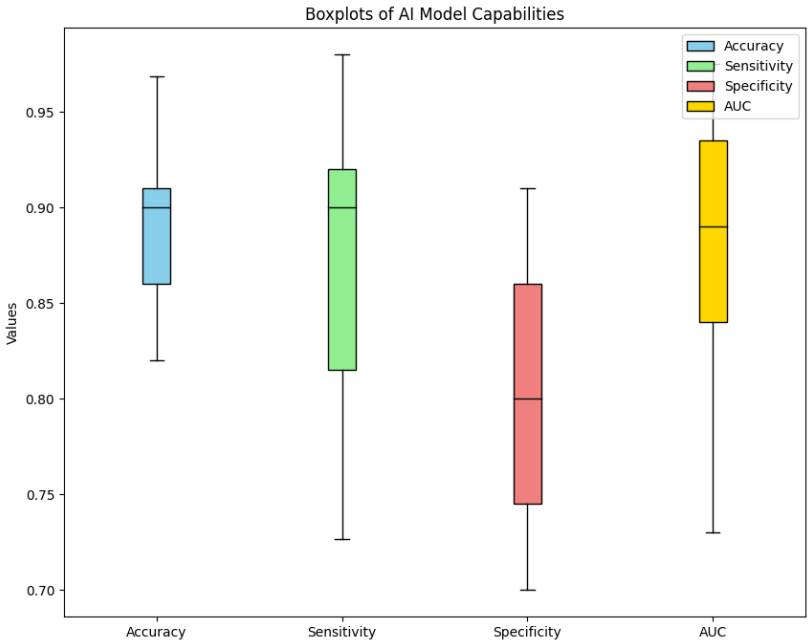

Figure 3: Boxplot of AI attributes

6 Discussion

The capabilities of AI models used in BC mammography were explored in this research paper. It studied only recent models in order to say that this is their performance today. The sensitivity, specificity, accuracy, and AUC of AI models today is found to be 87%, 80%, 89%, and 88% respectively. Radiologists generally have a sensitivity of 87%, similar to AI, but a specificity of 89%, which is better than current AI performance [28].

This research paper had hypothesized that recent models would have accuracies higher than 90%. However, these finding invalidate this hypothesis. Although computer vision has improved significantly over the years, even having an error rate less than 3.6% in the ImageNet Challenge in 2015 [21], it still needs to improve to achieve accuracries higher than 90%. This finding can be used by researchers and healthcare professionals to inform themselves on the current state of AI models in the field of BC mammography screenings.

This study is mainly limited by the number of AI models that were evaluated. The study only examined the performance of 16 models and would have produced more meaningful results if it evaluated more. This study also suffers from the bias introduced by the segment of human population which was used to create the databases on which the AI models were tested. For example, the DDSM database which was used to test 5 of the models, was a collaborative effort between Massachusetts General Hospital and others. This means the patients whose screenings were collected are more likely to live in the United States. The mini-MIAS data set which was used to test one of the models, was curated from BC screenings in the United Kingdom National Breast Screening Programme. Therefore, there is a danger of the models showing their performance on screenings on women who live in Western countries instead of women in general.

The authors recommend future research focus on addressing the limitations of this research paper. Future research should evaluate a wide range of AI models so that the results are statistically significant. The bias introduced by testing AI models in Western countries can be overcome by testing the models on data sets from a diverse set of countries. For example, VinDr-Mammo database from Vietnam [29] and Chinese Mammography Database from China [30] are large BC data sets are publicly available for research.

References

[1] W. H. Organization, "Breast cancer," *World Health Organization*, july 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=In%202020%2C%20there%20were%202.3,the%20world's%20most%20prevalent%20cancer

[2] S. C. J. C.-C. S. K. W. Z. S. E. B. D. E. P. D. P. P. . M. K. S. Maria Escala-Garcia, Anna Morra, "Breast cancer risk factors and their effects on survival: a mendelian randomisation study," November 2020. [Online]. Available: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01797-2

[3] "Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease," National Library of Medicine, July 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/12133652/

[4] "Breast cancer—epidemiology, risk factors, and genetics," National Library of Medicine, Sep 2000. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1118507/

[5] M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California Management Review*, pp. 1–10, 2019. doi: 10.1177/0008125619864925 Article reuse guidelines: sagepub.com/journals-permissions. [Online]. Available: https://doi.org/10.1177/0008125619864925

[6] Google Cloud. What is machine learning (ml)? Accessed on: 2023-11-06. [Online]. Available: https://cloud.google.com/learn/what-is-machine-learning

[7] IBM. What are neural networks? Accessed on: 2023-11-06. [Online]. Available: https://www.ibm.com/topics/neural-networks

[8] K. Seo, J. Tang, I. Roll, S. Fels, and D. Yoon, "The impact of artificial intelligence on learner–instructor interaction in online learning," *International Journal of Educational Technology in Higher Education*, vol. 18, p. 54, 2021. doi: 10.1186/s41239-021-00292-9. [Online]. Available: https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-021-00292-9

[9] M. Boniol, T. Kunjumen, T. S. Nair, A. Siyam, J. Campbell, and K. Diallo, "The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage?" *BMJ Glob Health*, vol. 7, no. 6, p. e009316, 2022. doi: 10.1136/bmjgh-2022-009316. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35760437/#:~:text=Results%3A%20In%202020%2C%20the%20global,to%2065.1%20million%20health%20workers

[10] M. . C. EIT Health, "Transforming healthcare with AI: The impact on the workforce and organizations," *McKinsey & Company*, March 10 2020. [Online]. Available: https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai

[11] N. C. Institute, "Pdq breast cancer screening," June 2023. [Online]. Available: https://www.cancer.gov/types/breast/hp/breast-screening-pdq

[12] E. Commission, "European commission initiative on breast cancer (ecibc): European guidelines on breast cancer screening and diagnosis," October 2019. [Online]. Available: https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/EtDs/ECIBC_GLs_EtD_mammography_readers.pdf

[13] F. Gaillard, "Mammography," March 2023. [Online]. Available: https://radiopaedia.org/articles/mammography

[14] H. S. Søren R Christiansen, Philippe Autier, "Change in effectiveness of mammography screening with decreasing breast cancer mortality:a population-based study," June 2022. [Online]. Available: https://academic.oup.com/eurpub/article/32/4/630/6609838

[15] Y. Matsuzaka and R. Yashiro, "Ai-based computer vision techniques and expert systems," *AI*, vol. 4, p. 13, 2021, this article belongs to the Special Issue Feature Papers for AI. [Online]. Available: https://www.mdpi.com/2673-2688/4/1/13

[16] M. E. Tschuchnig and M. Gadermayr, "Anomaly detection in medical imaging - a mini review," in *Data Science – Analytics and Applications*. Springer Fachmedien Wiesbaden, 2022, pp. 33–38. [Online]. Available: https://doi.org/10.1007%2F978-3-658-36295-9_5

[17] M. Salim, E. Wåhlin, K. Dembrower, E. Azavedo, T. Foukakis, Y. Liu, K. Smith, M. Eklund, and F. Strand, "External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms," *JAMA Oncology*, vol. 6, no. 10, pp. 1581–1588, Oct 2020. doi: 10.1001/jamaoncol.2020.3321. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7453345/

[18] T. Gneiting and E.-M. Walz, "Receiver operating characteristic (roc) movies, universal roc (uroc) curves, and coefficient of predictive ability (cpa)," *Machine Learning*, vol. 111, pp. 2769–2797, 2022. doi: 10.1007/s10994-021-06114-3. [Online]. Available: https://doi.org/10.1007/s10994-021-06114-3

[19] C. Lee, "Receiver operating characteristic (roc) curve with false positive rate and true positive rate," Wikimedia Commons, 2018, drawn by CMG Lee based on http://commons.wikimedia.org/wiki/File:roc-draft-xkcd-style.svg. [Online]. Available: https://commons.wikimedia.org/wiki/File:Roc_curve.svg

[20] J. D. Olga Russakovsky, "The box plot: A simple visual method to interpret data," *Springer Link*, 2015. doi: https://doi.org/10.1007/s11263-015-0816-y. [Online]. Available: https://link.springer.com/article/10.1007/s11263-015-0816-y?sa_campaign=email/event/articleAuthor/onlineFirst

[21] G. E. H. Alex Krizhevsky, Ilya Sutskever, "Imagenet classification with deep convolutional neural networks," 2012.

[22] S. H. . B. J. Hansen, C., "How to conduct a meta-analysis in eight steps: a practical guide," 2021. doi: https://doi.org/10.1007/s11301-021-00247-4

[23] J. S. M. R. Huecker., "Type i and type ii errors and statistical power," March 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK557530/

[24] C. S. D. T. S. J. A. C. S. T.-P. Karoline Freeman, Julia Geppert, "Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy," Sep 2021.

[25] S. Manikandan, "Measures of central tendency: Median and mode," *Journal of Pharmacology and Pharmacotherapeutics*, vol. 2, no. 3, pp. 214–215, July 2011. doi: 10.4103/0976-500X.83300 Accessed online: www.jpharmacol.com.

[26] A. Hazra, "Using the confidence interval confidently," *Journal of Thoracic Disease*, vol. 9, no. 10, 2017. [Online]. Available: https://jtd.amegroups.org/article/view/16406

[27] D. Williamson, R. Parker, and J. Kendrick, "The box plot: A simple visual method to interpret data," *Annals of internal medicine*, vol. 110, pp. 916–21, 07 1989. doi: 10.1059/0003-4819-110-11-916

[28] C. D. Lehman, R. F. Arao, B. L. Sprague, J. M. Lee, D. S. M. Buist, K. Kerlikowske, L. M. Henderson, T. Onega, A. N. A. Tosteson, G. H. Rauscher, and D. L. Miglioretti, "National performance benchmarks for modern screening digital mammography: Update from the breast cancer surveillance consortium," *Radiology*, vol. 283, no. 1, pp. 49–58, April 2017. doi: 10.1148/radiol.2016161174. [Online]. Available: https://doi.org/10.1148/radiol.2016161174

[29] H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao, and V. Vu, "Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography," *Scientific data*, vol. 10, no. 1, p. 277, 2023. doi: 10.1038/s41597-023-02100-7 This work was funded by the Vingroup JSC. The funder had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript. [Online]. Available: https://www.nature.com/articles/s41597-023-02100-7

[30] C. Cui, L. Li, H. Cai, Z. Fan, L. Zhang, T. Dan, J. Li, and J. Wang, "The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast," *The Cancer Imaging Archive*, 2021. doi: 10.7937/tcia.eqde-4b16 Data Citation. [Online]. Available: https://doi.org/10.7937/tcia.eqde-4b16

A Appendix

Journal	Year	Accuracy	Sensitivity	Specificity	AUC	Dataset	DOI Link
Computers	2016	0.91	0.92	0.84	0.91	600.00	link
PeerJ	2019	0.87	0.86	0.88	0.94	2,620.00	link
PubMed	2018	0.86	0.93	0.78	0.88	600.00	link
PubMed	2018	0.91	0.92	0.91	-	2,482.00	link
PubMed	2017	0.93	0.96	0.90	-	216.00	link
PubMed	2017	0.84	0.91	0.80	0.92	761.00	link
PubMed	2017	0.84 - 0.95	0.98	0.70	0.69 - 0.76	410.00	link
PubMed	2017	0.82	0.792 - 0.81	0.694 - 0.723	0.88	1,874.00	link
PubMed	2019	-	0.8	-	0.84	2652	link
Scientific Reports	2018	-	0.9	-	0.95	2949	link
PubMed	2018	0.9685	-	-	0.975	6116	link
PubMed	2018	-	0.7265	0.708	0.8	2292	link
PubMed	2017	-	0.842	0.804	0.84	1264	link
PubMed	2017	-	-	-	0.941	18453	link
PubMed	2017	-	-	-	0.78	2242	link
PubMed	2017	-	0.815	0.79	0.9	1090	link

Table 2: Collected publications of BC screening AI models and their key metrics

Software	Version
Python	NVIDIA T1000 / PCIe / SSE2
matplotlib	3.7.3
pandas	2.0.3
scipy	1.10.1

Table 3: Software running statistical analysis

Specification	Details
Processor	12th Gen Intel® Core™ i7-12700 × 20
Graphics	NVIDIA T1000 / PCIe / SSE2
OS Name	Ubuntu 20.04.6 LTS
OS Type	64-bit
GNOME Version	3.36.8
Windowing System	X11

Table 4: Hardware running statistical analysis