Original Article

Eur J Breast Health 2026; 22(1): 35-43
DOI: 10.4274/ejbh.galenos.2025.2025-4-1

Comparative Evaluation of Machine Learning and Specialist Physicians in Breast Care Triaging: A Real-World Observational Study

Aswini Misro[1], Naim Kadoğlou[2], Hüseyin Doğan[3]

[1]Department of Innovation, YouDiagnose Limited, London, United Kingdom
[2]Department of Breast and General Surgery, London North-West NHS Trust, London, United Kingdom
[3]Department of Media, Science and Technology Bournemouth University, Poole, United Kingdom

ABSTRACT

Objective: To evaluate the diagnostic accuracy and efficiency of a proprietary breast-specific machine learning (ML) model—built upon the open-source Open Triage platform—in comparison to specialist physicians, using standardized real-world clinical data for breast referral triaging.

Materials and Methods: A retrospective observational study was conducted using 174 standardized breast cases obtained from proprietary industry datasets, spanning 46 disease types, 23 of which were cancers. The cohort ranged from 19 to 75 years (mean: 39.4±12.0). Physicians and an ML model each generated three diagnostic predictions per case. Both modalities were compared after benchmarking their predictions against a gold-standard diagnosis established by imaging and biopsy. Performance was evaluated using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and receiver operating characteristic (ROC) analysis. Time efficiency was also assessed to compare diagnostic turnaround times between physician- and ML-generated predictions.

Results: The ML model demonstrated superior diagnostic accuracy (100%) compared to physicians (83.9%), with higher sensitivity (0.947 *vs.* 0.826) and PPV (0.500 *vs.* 0.442). Both groups achieved comparable specificity and NPV values. ROC analysis showed an AUC of 0.91 for the ML model's first prediction versus 0.83 for the doctor's first prediction, indicating superior predictive power of the ML model.

Conclusion: The ML model demonstrated diagnostic accuracy comparable to or better than that of physicians while significantly reducing the time required. These findings suggest that AI-powered triage tools could enhance the efficiency and standardization of breast triage.

Keywords: Artificial intelligence; clinical decision support systems; breast surgery; machine learning; predictive models; diagnostic accuracy

Cite this article as: Misro A, Kadoğlou N, Doğan H. Comparative evaluation of machine learning and specialist physicians in breast care triaging: a real-world observational study. Eur J Breast Health. 2026; 22(1): 35-43

Key Points

- Artificial intelligence
- Clinical decision support systems
- Predictive models

Introduction

Triage is the critical process of prioritising patients based on the urgency of their condition to ensure timely care and effective resource allocation. In breast care, this involves identifying individuals at elevated risk—such as those with suspected or confirmed cancer, genetic predispositions, or concerning symptoms—so that diagnostic evaluation and intervention can be appropriately expedited, ultimately improving patient outcomes (1-3).

In breast care, triage is typically implemented through several modes of entry, shaped by patient circumstances and the nature of the clinical encounter. Within population-based screening workflows, radiologists traditionally conduct image reviews and make recall decisions guided

Corresponding Author:
Aswini Misro MD; draswini@gmail.com

Received: 10.04.2025
Accepted: 11.09.2025
Online Publication Date: 25.12.2025

by established protocols. Artificial intelligence (AI)-enabled triage mechanisms are increasingly being explored to prioritise abnormal findings, accelerate diagnostic follow-up, and reduce radiologist workload without compromising cancer detection rates (4-8). Opportunistic triage, by contrast, occurs during unrelated healthcare encounters, such as routine clinical visits, during which clinicians take the opportunity to initiate breast assessment outside formal referral systems (9). Referral triage in breast care is predominantly symptom-driven; primary care providers submit structured referrals for patients presenting with concerning features, and these referrals are then assessed by specialist teams (3, 10).

Based on these varied clinical entry points, breast triage models have diversified to encompass manual sorting, telephonic and virtual platforms, and increasingly digital or algorithmic approaches alongside traditional pathways (11, 12). During the COVID-19 pandemic, temporary innovations such as virtual triage gained prominence; however, resource-intensive manual triage processes remain the norm today (13, 14). Research highlights that the effectiveness of triage is heavily influenced by multiple variables-most notably clinician expertise, workload fluctuations, infrastructure constraints, and the availability of supporting systems and personnel (12, 15-17). Notably, risk-averse triage may drive unnecessary investigations and overtreatment (18, 19), whereas purely rule-based digital systems may increase workload without demonstrable improvement in clinical outcomes (2, 15, 16, 20, 21). Thus far, many digital triage tools have largely digitized existing protocols without addressing broader systemic pressures such as legal anxieties and patient demand (22, 23).

Such practical constraints are reflected in referral trends: breast cancer referrals have seen a marked increase—from 432 to 1,027 per 100,000 individuals between 2009 and 2023—while conversion rates have declined from 2% to 1% (24). Evidence further suggests that consistent application of evidence-based triage protocols could raise conversion rates to 14%, highlighting the urgent need for innovation and more effective triage processes (25).

In response to these challenges, AI initiatives have gained traction across specialties-from emergency medicine and dermatology to emerging applications in radiology. Although AI adoption in breast triage remains limited, lessons from emergency medicine (26) and dermatology underscore its potential (27, 28). One major modality is multi-layered, data-driven triage, where machine learning (ML) and deep learning models analyse diverse structured inputs—such as electronic health records, imaging data, and patient histories—to deliver nuanced risk stratification and consistent decision-making (26).

A second modality is image-based triage, which leverages clinical photographs or scans, such as mammograms, to enable direct visual assessment by ML models. This pathway represents a distinct form of triage, allowing algorithms to process and prioritise visual information independently of structured datasets (29). For instance, in tele-dermatology, image-based triage enabled remote resolution of over 80% of consultations, halving the need for in-person visits (30). The third modality involves reinforcement learning (RL), where algorithms learn optimal triage policies by observing expert clinical decisions. Some evidence suggests RL models can achieve safe, consistent decision-making that mirrors expert-level reasoning, reinforcing their potential to adapt to complex, dynamic triage environments (31). Finally, natural language processing offers a flexible modality by integrating unstructured narrative data—such as free-text nursing notes—with structured clinical inputs. This fusion enables models to surpass the limitations of structured-only systems, improving triage classification accuracy and enriching contextual understanding of patient acuity (32).

This study builds upon evidence-based AI triage models used in other specialties and addresses gaps in breast care by evaluating a domain-specific clinical decision support system (CDSS) for breast triage. The CDSS augments traditional referral pathways through a structured, data-driven methodology. By integrating comprehensive patient histories, including medical records, family histories, and surgical interventions, it generates nuanced cancer risk assessments and personalised care pathway recommendations, offering alternative scenarios with associated probabilities and confidence levels. This observational study aims to assess the performance of predictive triage data models in clinical practice by benchmarking against actual diagnostic outcomes (33).

Materials and Methods

This retrospective observational study was designed to compare the diagnostic triage predictions generated by specialist physicians with those produced by an ML application, using a standardized set of performance metrics.

The study aimed to address the following research questions:

1. What level of agreement exists between the ML model's predictions and the gold standard?

2. How do the ML model's predictions compare with those of specialist physicians?

3. What is the level of agreement between physicians' predictions and the gold standard?

The study protocol, project number 342655/YD, was approved by the YouDiagnose Ethical Approval Committee under application no. 101/01022020 on February 1, 2020. All methodologies adhered to COPE guidelines and the Declaration of Helsinki. Informed consent was obtained from participants prior to study initiation. No non-anonymized human data or biological samples were used.

The study population comprised female patients aged 18 years and older who presented with breast-related symptoms. Exclusions were applied to males, individuals under 18 years, those with a history of breast cancer, and those with rare conditions such as idiopathic granulomatous mastitis or Mondor's disease. Data were sourced from proprietary industry datasets comprising 348 consecutive, anonymized breast cases. Following a three-step data standardization process, including data cleaning, completeness assessment, and independent review by senior breast consultants, 174 cases were selected for analysis.

All cases were processed using the same three-step data standardization procedure to maintain consistency and reliability, and to ensure compliance with data protection protocols. The methodology is described below, with a detailed workflow provided in Figure 1.

Data Collection: Data from 348 consecutive breast cases were collected from proprietary datasets provided by industry partners. Each case was anonymized and subjected to cleaning procedures consistent with data privacy and confidentiality requirements. A preliminary review excluded 21 cases not meeting the inclusion criteria.

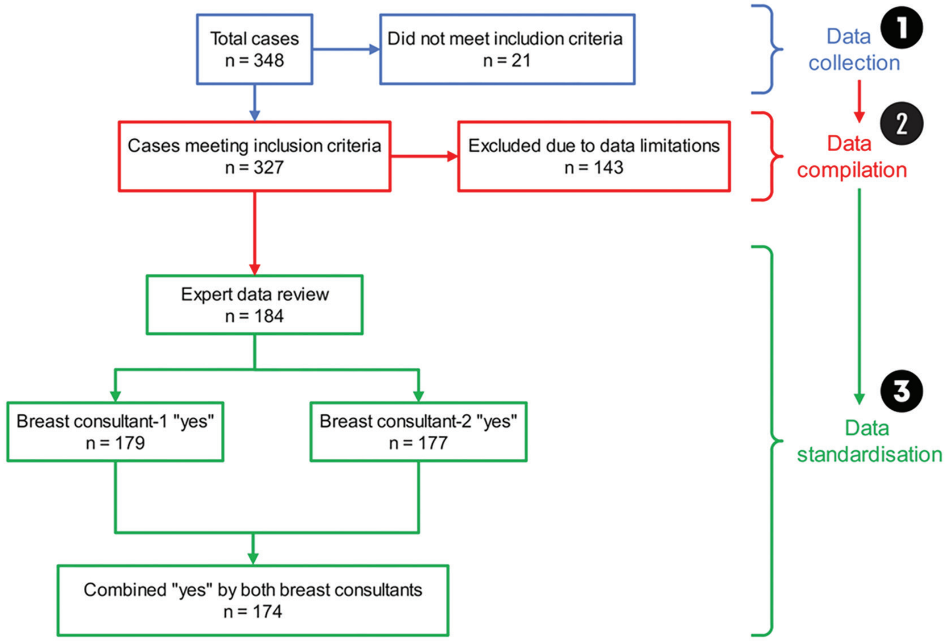
Total cases
n = 348
Did not meet inclusion criteria
n = 21
Data collection
1
Cases meeting inclusion criteria
n = 327
Excluded due to data limitations
n = 143
2
Data compilation
Expert data review
n = 184
Breast consultant-1 "yes"
n = 179
Breast consultant-2 "yes"
n = 177
3
Data standardisation
Combined "yes" by both breast consultants
n = 174

Figure 1. Various stages of data standardisation

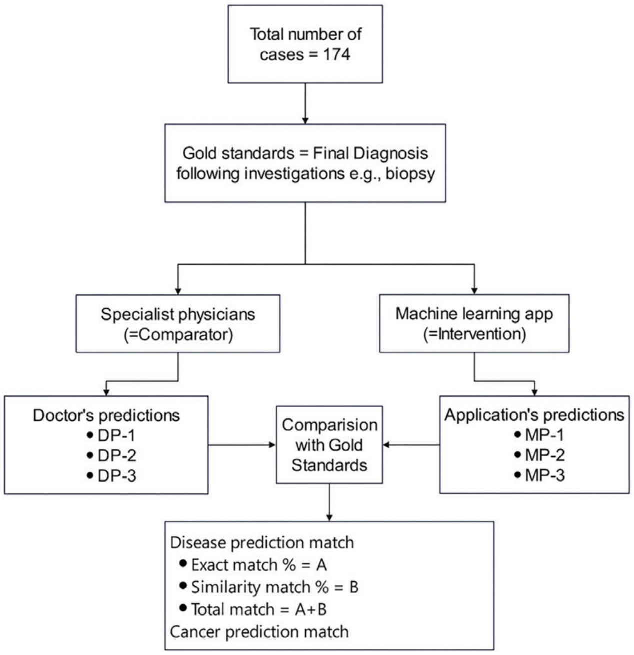
Total number of cases = 174
Gold standards = Final Diagnosis following investigations e.g., biopsy
Specialist physicians (=Comparator)
Machine learning app (=Intervention)
Doctor's predictions
• DP-1
• DP-2
• DP-3
Comparision with Gold Standards
Application's predictions
• MP-1
• MP-2
• MP-3
Disease prediction match
• Exact match % = A
• Similarity match % = B
• Total match = A+B
Cancer prediction match

Figure 2. The study design

DP: Doctor performance; MP: Model performance

Data Compilation: Two physicians and research associates assessed the completeness of key categories, including chief complaints, medical history, diagnostic investigations, interventions, and final diagnoses. Cases with missing data in multiple categories were excluded, removing 143 from the dataset.

Data Review: Two senior breast consultants independently assessed the adequacy of the remaining cases against National Health Service (NHS) standards in the United Kingdom. After this review, 174 cases were approved for inclusion in the final dataset.

The gold standard for diagnostic benchmarking was defined as the final diagnosis established through confirmatory imaging—such as mammography, magnetic resonance imaging (MRI), and ultrasonography—and, when appropriate, biopsy-verified diagnoses.

Seventeen specialist breast surgeons from four NHS Breast Surgery Units were invited to participate, with ten agreeing after three rounds of contact. Eligible physicians had at least five years of specialist experience, were registered with the General Medical Council, and demonstrated proficiency in computer-based systems. All participants received induction training on the study's digital interface and provided differential diagnoses, cancer risk assessments, and urgency ratings for each pseudonymized case under controlled conditions.

The ML predictions were generated using a platform called Open Triage, an open-source software developed and maintained by the Uppsala Centre for Prehospital Research at Uppsala University (34). Building on this platform, we integrated a proprietary breast-specific prediction model designed to support clinical decision-making in breast health. While Open Triage is freely available as open-source software, the final deployed prediction model is a copyrighted product developed by the research team and is not commercially available.

Predictions from both physicians and the ML model were benchmarked against the gold standard, with agreement evaluated using sensitivity, specificity, positive predictive value, negative predictive value, and receiver operating characteristic (ROC) analysis. Diagnostic outcomes were classified as either perfect matches (exact correspondence with the gold standard) or similarity matches (close alignment with minor variations), providing a robust and clinically relevant measure of diagnostic performance. A detailed overview of the study design is provided in Figure 2.

Participating physicians underwent induction training on software navigation and user interface functionality. The study was conducted using laptops connected to secure internet networks under controlled conditions. Physicians accessed pseudonymized patient cases via login credentials and provided differential diagnoses alongside assessments of cancer risk and care urgency.

Predictions made by physicians and the ML model were compared with the gold-standard diagnosis and classified into three outcome types: a perfect match, in which the predicted diagnosis was identical to the gold-standard diagnosis; a similarity match, in which the predicted diagnosis had a closely related pathogenesis, clinical course, and overlapping symptoms; and a non-match, in which the predicted diagnosis did not correspond meaningfully to the gold-standard diagnosis.

For example, a similarity match includes predicting "lactation abscess" when the gold-standard diagnosis is "lactation mastitis". These conditions, while not strictly identical, involve similar underlying mechanisms (inflammation and infection within the lactating breast), present similarly, and require comparable clinical management. This approach acknowledges that certain diseases exist along a spectrum or are commonly conflated due to overlapping features. Categorizing them as similarity matches allows for a meaningful evaluation of both diagnostic accuracy and the clinician's practical reasoning.

Results

Study Population and Baseline Characteristics

A total of 174 cases were included in the final analysis after applying inclusion and exclusion criteria. The cohort's ages ranged from 19 to 75 years (mean 39.4±12.0; median 38; mode 45), indicating a moderately dispersed, approximately bell-shaped distribution that spanned young adulthood to older age brackets and was slightly skewed toward younger individuals. These cases represented 46 distinct breast disease types, including 23 cancer cases and 151 benign conditions, and the gold standard—established through final diagnoses verified by mammograms, MRIs, ultrasounds, and biopsies—served as the reference standard for evaluating diagnostic predictions.

Statistical Approach

To compare the diagnostic performance of specialist physicians and the ML model, a cumulative predictive power approach was employed. This method aggregated all diagnostic predictions—exact matches (A) and similarity matches (B)—from both groups. The combined predictions (A+B) were analysed to evaluate the overall performance of each group. For physicians, predictions were categorized as DP1, DP2, and DP3 (representing their top three differential diagnoses), while for the ML model, predictions were categorized as MP1, MP2, and MP3. Aggregated predictions for both groups were assessed using sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and ROC curve analysis. A comparative analysis of these metrics enabled a direct evaluation of performance (Table 1).

Diagnostic Accuracy

Table 2 summarizes the diagnostic accuracy of both physicians and the ML model. Among the physicians, DP1 achieved moderate performance with 66 exact matches (37.93%) and 58 similarity matches (33.33%), while DP2 and DP3 demonstrated considerably lower accuracy with only 7 and 15 exact matches, respectively, and neither had any similarity matches. In contrast, the ML model's MP1 achieved superior performance with 111 exact matches (63.79%) and 40 similarity matches (22.99%). MP2 and MP3 exhibited lower accuracy, with 23 and 11 exact matches, respectively, and no similarity matches.

When aggregated across all predictions, the ML model outperformed physicians, with an overall accuracy of 100% versus 83.9% for physicians. These findings indicate that the ML model was more effective in achieving both exact and similarity-based matches in this comparative analysis.

Table 1. Displays the years of specialist practice, and the number of cases allocated to each participating physician

Sl no	Username	No of cases	Year of specialist practice	Postgraduate qualifications
1	user291	37	20	MD, FRCS
2	user292	37	15	FRCS
3	user293	13	14	MCh, FRCS
4	user294	11	10	FRCS
5	user295	11	8	FRCS
6	user296	13	9	FRCS
7	user297	12	11	FRCS
8	user298	14	12	FRCS
9	user299	11	9	FRCS
10	user300	15	12	MD, FRCS
174		Total number of cases	Average years of specialist practice	
		12.0		

Table 2. Showing the comparative performance specialist physicians (DP1, DP2, DP3) and the machine learning model (MP1, MP2, MP3)

Variable	Exact match	Exact match accuracy	Similarity match	Similarity match accuracy	Total matches	Total matches accuracy
DP1	66	37.931%	58	33.333%	124	71.264%
DP2	7	4.022%	0	0%	7	4.022%
DP3	15	8.620%	0	0%	15	8.620%
Combined DP	88	50.574%	58	33.333%	146	83.908%
MP1	111	63.793%	40	22.988%	151	86.781%
MP2	23	13.218%	0	0%	23	13.218%
MP3	11	6.321%	0	0%	11	6.321%
Combined MP	145	83.333%	40	22.988%	174	100%

DP: Doctor performance; MP: Model performance

Performance Metrics

Table 3 presents a detailed comparison of performance metrics between the physicians (DP) and the ML model (MP). Key observations include:

- **Sensitivity:** The ML model demonstrated higher sensitivity (0.947) than physicians (0.826), indicating its superior ability to correctly identify true positives.

- **Specificity:** Both groups achieved comparable specificity, with the ML model slightly outperforming physicians (0.854 *vs.* 0.841).

- **PPV:** The ML model achieved a higher PPV (0.500) than physicians (0.442), suggesting greater reliability in predicting positive cases.

- **NPV:** Both groups exhibited high NPV, with the ML model achieving 0.992 compared with 0.969 for physicians.

Table 3. Matching performance metrics: specialist physicians (DP) and the machine learning model (MP)

1. Metric	2. DP	3. MP
4. Sensitivity	5. 0.826	6. 0.947
7. Specificity	8. 0.841	9. 0.854
10. PPV	11. 0.442	12. 0.5
13. NPV	14. 0.969	15. 0.992
16. Accuracy	17. 0.839	18. 0.868

DP: Doctor performance; MP: Model performance

ROC Analysis

The ROC curves for all predictions are illustrated in Figure 3. The dashed diagonal line represents random classification, with an area under the curve (AUC) of 0.50.

- Among physicians' predictions, DP1 achieved the highest AUC (0.83), significantly outperforming DP2 (0.50) and DP3 (0.54).

- For the ML model, MP1 demonstrated superior predictive power with an AUC of 0.91, while MP2 and MP3 performed no better than random guessing (AUC = 0.50).

- Combined predictions for both groups mirrored their highest-performing individual predictor: DP1 for physicians (AUC = 0.83) and MP1 for the ML model (AUC = 0.91).

Statistical comparison of AUCs was performed using DeLong's test for correlated ROC curves, as all predictions were made on the same dataset. The differences between top-performing and lower-performing models were statistically significant ($p<0.001$ for all pairwise comparisons).

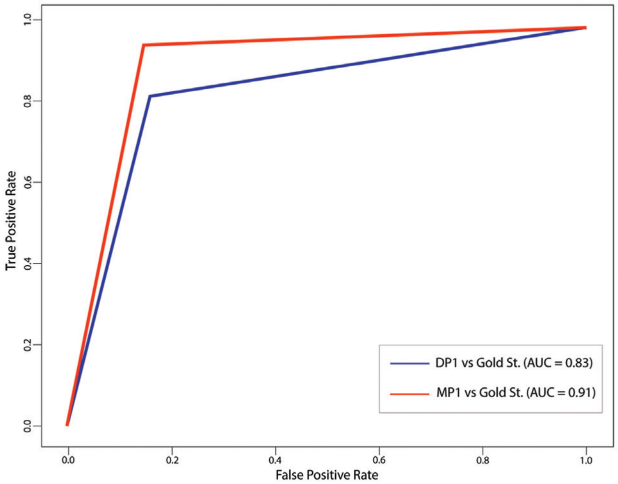

Figure 3. ROC curve all predictions

ROC: Receiver operating characteristic

Data Analysis

The Wilcoxon Signed-Rank test was employed to compare physician- and ML-generated predictions for agreement with the gold-standard diagnosis.

- Although a test statistic of 0.0 would normally not correspond to a small p-value, the reported p-value of 1.52e-10 indicates a statistically significant difference between physician and ML predictions.

- These findings suggest that observed differences in diagnostic accuracy are unlikely to be due to random variation.

Confusion Matrix Analysis

Confusion matrices comparing diagnostic classifications by physicians and the ML model are presented in Figure 4.

- For benign cases, the ML model correctly classified 130 true negatives and produced 22 false positives, while physicians correctly identified 128 true negatives and recorded 24 false positives.

- For cancer cases, The ML model demonstrated superior sensitivity with no false negatives, correctly identifying all 22 cases as true positives. Physicians correctly identified 19 cancer cases but misclassified 3 cancer cases as false negatives.

The findings demonstrate that the ML model has superior sensitivity and precision in identifying cancer cases. However, due to the limited sample size, the difference in cancer detection performance between the ML model and specialist physicians did not reach statistical significance (p>0.05), indicating that the ML model's diagnostic capability is comparable to that of specialist physicians.

Key Observations

1. The ML model consistently and significantly outperformed physicians in diagnostic accuracy and across all performance metrics.

2. The ROC analysis confirmed that the top-performing prediction in each group dominated the overall predictive ability.

3. The ML model demonstrated superior time efficiency, processing cases almost instantaneously, compared with the several minutes per case required by physicians.

4. While both methods showed high accuracy in identifying benign cases, the ML model exhibited enhanced sensitivity in detecting cancer cases.

5. Despite these advantages, no statistically significant difference in cancer identification was observed, owing to sample size limitations (p>0.05). This suggests that in classifying cases as benign or malignant, the ML model performs comparably to industry comparators and demonstrates noninferiority relative to physicians.

Discussion and Conclusion

This study demonstrates that ML models, when applied to real-world clinical datasets in an experimental setup, exhibit diagnostic performance that is comparable to or superior to those of specialist physicians across several metrics, highlighting their effectiveness in breast triage and warranting further evaluation in live, patient-facing clinical settings. This discussion contextualizes the findings, outlines their clinical implications, and presents limitations alongside conclusions. It also highlights existing studies on AI/ML-led triage in breast surgery and identifies directions for future research.

Diagnostic Performance

Our analysis demonstrated that the ML model outperformed physicians on key diagnostic metrics, including sensitivity (94.7% *vs.* 82.6%), PPV (50.0% *vs.* 44.2%), and overall diagnostic accuracy (100.0% *vs.* 83.9%). While the ML model correctly identified all cancer cases (i.e., produced no false negatives), physicians misclassified three cancer cases as false negatives. However, statistical analysis revealed no significant difference in cancer identification performance between the ML model and physicians due to sample size limitations (p>0.05). This indicates that the ML model performs at a level comparable to industry standards and demonstrates diagnostic capability that is non-inferior to that of specialist physicians. Importantly, these findings highlight the potential of the ML model to complement physicians in early cancer triage predictions and to assist in reducing the likelihood of missed diagnostic opportunities that occur in referral pathways, a persistent challenge in breast health triage.

Both physicians and the ML model exhibited high specificity and NPV, with comparable results between the two groups (specificity: 0.854 *vs.* 0.841; NPV: 0.992 *vs.* 0.969). While both approaches demonstrated strong performance in ruling out benign cases, the ML model, as previously noted, showed a marginally higher sensitivity in detecting malignancies. However, this advantage did not reach statistical significance. These results underscore the potential for integrating ML models alongside clinical expertise to enhance diagnostic efficiency and support decision-making in breast care pathways.

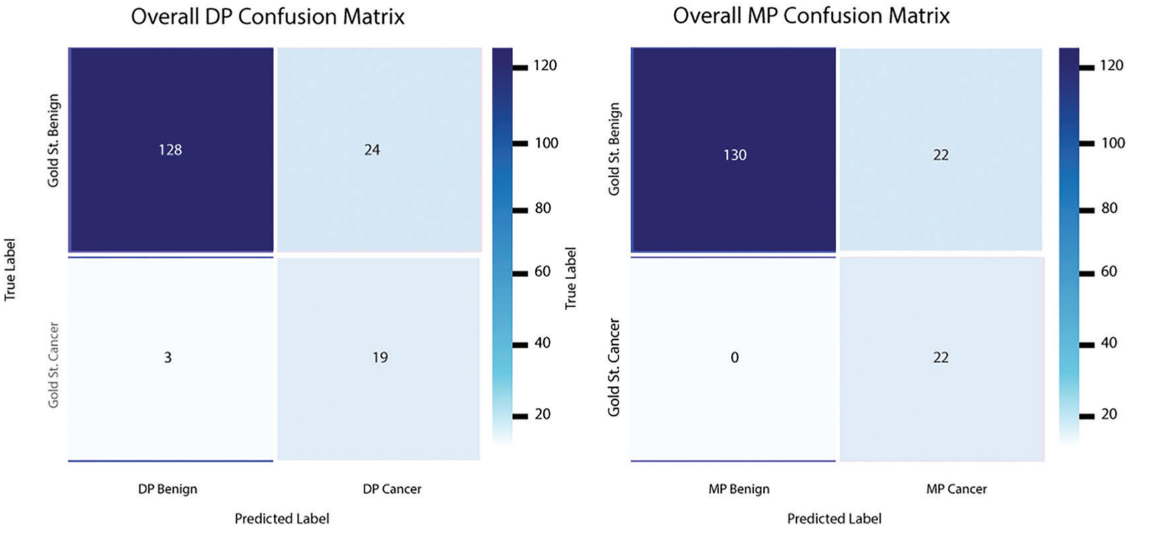

Figure 4. Confusion matrices from DP and MP

DP: Doctor performance; MP: Model performance

Efficiency and Time Savings

The time required for diagnosis differed markedly between groups. Physicians required 789 minutes to review all cases, whereas the ML model processed the entire dataset in 0.3215 seconds. Figure 5 demonstrates this efficiency gap, showing the ML model's superior speed without compromising diagnostic accuracy.

Integrating ML into triage can reduce manual workload, ease referral bottlenecks, and improve resource use, particularly in high-volume environments. Further research is needed to evaluate the practical time and cost benefits of combining ML with clinical expertise.

Implications for Clinical Practice

This study underscores the value of integrating ML analytics with physician expertise to create a hybrid triaging model that leverages the strengths of each. Physicians offer intuition, empathy, and judgment in complex cases, while ML systems contribute consistent risk assessments and pattern recognition, thereby reducing cognitive burden and standardizing decision-making.

This partnership can address inefficiencies commonly observed in manual triage by enabling timelier and more targeted patient management. ML models can assist in pre-screening referrals and identifying potentially high-risk cases, helping clinicians allocate their expertise more efficiently across all levels of care—from routine assessments to complex diagnostic decisions—while maintaining oversight and accountability at every step.

The findings confirm that ML tools can effectively support breast health triage, matching or exceeding the diagnostic accuracy of specialists, while significantly reducing time demands. When guided by patient-centred care principles, this data-driven collaboration has the potential to transform triage across healthcare systems—advancing equity, efficiency, and clinical precision.

Supporting Evidence from Contemporary Literature

The findings of this study align with substantial evidence from contemporary literature supporting the clinical utility and effectiveness of AI-assisted breast triaging systems. Mazo et al. (35) conducted a systematic review of CDSS in breast cancer care, demonstrating that such systems significantly assist healthcare staff with clinical decision-making while improving care quality and minimizing costs. This systematic review validates the conceptual framework underlying ML model implementation in breast care pathways, providing robust evidence that automated decision support systems can enhance healthcare efficiency while maintaining patient-centred care.

The DENSE trial validation studies provide particularly compelling evidence for the real-world effectiveness of AI-assisted breast triaging. These studies demonstrated that combined computer-aided triaging and computer-aided diagnosis systems dismissed 32.7% of normal examinations, while correctly identifying 46.3% of benign lesions without missing any malignant cases, yielding significantly fewer false-positive referrals (a 48.6% reduction) compared to radiological reading alone (36). Similarly, the MASAI trial showed that AI-supported screening detected 29% more cancers than traditional screening methods while reducing mammogram reading workload by 44% (37). These findings directly parallel our results, which show superior ML performance in sensitivity (0.947 *vs.* 0.826) and dramatic efficiency gains (0.3215 seconds per case *vs.* 4.5 minutes per case), reinforcing the clinical validity of AI-assisted breast triaging approaches.

Furthermore, Arıbal (38) provides a crucial perspective on the future of breast radiology, emphasizing that AI integration will transform radiologists into specialized clinicians collaborating with AI systems rather than being replaced by them, a view that aligns with our discussion of hybrid approaches that combine physician expertise with ML precision. The emphasis by Oren et al. (2020) on shifting

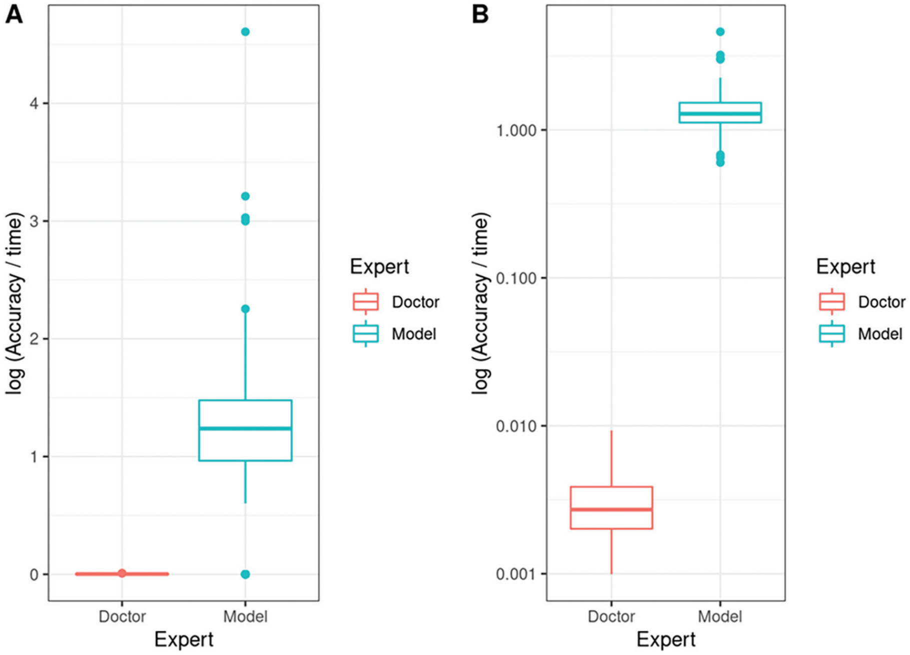

Figure 5. Comparative analysis of log accuracy between doctor and model

AI research focus from radiographic pathological data to clinically meaningful endpoints such as survival, symptoms, and treatment necessity is particularly relevant to our study's emphasis on practical clinical outcomes and diagnostic accuracy metrics that directly impact patient care decisions. These converging lines of evidence from multiple research settings strengthen the foundation for implementing ML-based triaging systems in clinical practice.

Study Limitations

This study, while yielding promising outcomes, is subject to several noteworthy limitations. The relatively small sample size of 174 cases restricts the statistical power of the findings and may hinder their generalizability to broader populations. Furthermore, the research was confined to breast health triaging and employed standardized datasets, which may not adequately represent the variability and complexity encountered in other clinical domains.

The study design also relied on a controlled environment where physicians made decisions based on pseudonymized case details. Such conditions may not fully capture the nuances of real-world clinical practice, where additional contextual and patient-specific factors are at play. Moreover, differences in physician training, experience, or familiarity with digital tools could have influenced performance outcomes. There remains a risk of bias, particularly given the potential for overinterpretation of performance metrics when comparing multiple variables.

However, validation in larger, more diverse cohorts is essential to confirm these findings and explore their generalizability. Future research should incorporate real-world clinical data across multiple specialties and evaluate the long-term impact of ML-assisted triaging on patient outcomes.

Ethics

Ethics Committee Approval: The study protocol, project number 342655/YD, was approved by the YouDiagnose Ethical Approval Committee under application no. 101/01022020 on February 1, 2020. All methodologies adhered to COPE guidelines and the Declaration of Helsinki.

Informed Consent: Informed consent was obtained from participants prior to study initiation.

Footnotes

Authorship Contributions

Surgical and Medical Practices: A.M., N.K.; Concept: A.M., N.K., H.D.; Design: A.M., N.K., H.D.; Data Collection or Processing: A.M.; Analysis or Interpretation: A.M., N.K.; Literature Search: A.M.; Writing: A.M.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. van Wegen ME, Fransen LFC, Thijssen WAMH, Alexandridis G, de Groot B. The association between urgency level and hospital admission, mortality and resource utilization in three emergency department triage systems: an observational multicenter study. Scand J Trauma Resusc Emerg Med. 2025; 33: 72. (PMID: 40312391) [Crossref]
2. NHS England. Digital, General practice. 2025 [cited 2025 Jul 10]. NHS England » Digitally enabled triage. [Crossref]
3. NHS England » Faster diagnostic pathways: implementing a timed breast cancer diagnostic pathway: guidance for local health and social care systems [Internet]. [cited 2025 Jul 10]. [Crossref]
4. Breast screening: programme overview - GOV.UK [Internet]. [cited 2025 Jul 10]. [Crossref]
5. Friedewald SM, Sieniek M, Jansen S, Mahvar F, Kohlberger T, Schacht D, et al. Triaging mammography with artificial intelligence: an implementation study. Breast Cancer Res Treat. 2025; 211: 1-10. (PMID: 39881074) [Crossref]
6. Miglioretti DL, Bissell MCS, Kerlikowske K, Buist DSM, Cummings SR, Henderson LM, et al. Assessment of a risk-based approach for triaging mammography examinations during periods of reduced capacity. JAMA Netw Open. 2021; 4: e211974. (PMID: 33764423) [Crossref]
7. Seker ME, Koyluoglu YO, Ozaydin AN, Gurdal SO, Ozcinar B, Cabioglu N, et al. Diagnostic capabilities of artificial intelligence as an additional reader in a breast cancer screening program. Eur Radiol. 2024; 34: 6145-6157. (PMID: 38388718) [Crossref]
8. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health. 2020; 2: e468-e474. (PMID: 33328114) [Crossref]
9. Ginsburg O, Yip CH, Brooks A, Cabanes A, Caleffi M, Dunstan Yataco JA, et al. Breast cancer early detection: A phased approach to implementation. Cancer. 2020; 126(Suppl 10): 2379-2393. (PMID: 32348566) [Crossref]
10. NHS Digital. Breast (symptomatic) referrals - NHS e-Referral Service [Internet]. 2024 [cited 2025 Jul 10]. [Crossref]
11. Laws S, Spiller K, Glew C. Evaluation of a pilot of a community virtual triage for breast symptoms outside of usual primary or secondary care pathways. Ann R Coll Surg Engl. 2024; 106: 596-600. (PMID: 38404244) [Crossref]
12. MacNeill F, Lead C, Irvine T, Clinical Advisor S. Breast surgery GIRFT programme national specialty Report. 2021. [Crossref]
13. Shetty G, Datta U, Rea I, Rai S, Hwang MJ, Hoar F, et al. Rapid implementation of triaging system for assessment of breast referrals from primary care centres during the COVID-19 pandemic. Ann R Coll Surg Engl. 2021; 103: 576-582. (PMID: 34464568) [Crossref]
14. Farzandipour M, Nabovati E, Sharif R. The effectiveness of tele-triage during the COVID-19 pandemic: a systematic review and narrative synthesis. J Telemed Telecare. 2024; 30: 1367-1375. (PMID: 36683438) [Crossref]
15. Llewelyn H. Evidenced based practice should reduce overdiagnosis and overtreatment. BMJ. 2012; 344: e4296. (PMID: 22734095) [Crossref]
16. Moynihan R, Henry D, Moons KG. Using evidence to combat overdiagnosis and overtreatment: evaluating treatments, tests, and disease definitions in the time of too much. PLoS Med. 2014; 11: e1001655. (PMID: 24983872) [Crossref]
17. Kulkarni SS, Dewitt B, Fischhoff B, Rosengart MR, Angus DC, Saul M, et al. Defining the representativeness heuristic in trauma triage: a retrospective observational cohort study. PLoS One. 2019; 14: e0212201. (PMID: 30735553) [Crossref]
18. Mohan D, Farris C, Angus DC, Fischhoff B, Rosengart MR, Yealy DM, et al. Understanding the role of heuristics in physician non-compliance with trauma triage guidelines. J Am Coll Surg. 2014; 219: S111. [Crossref]
19. Mohan D, Fischhoff B, Angus DC, Rosengart MR, Wallace DJ, Yealy DM, et al. Serious games may improve physician heuristics in trauma triage. Proc Natl Acad Sci U S A. 2018; 115: 9204-9209. (PMID: 30150397) [Crossref]
20. Iacobucci G. NHS 111 sends more and more callers to emergency departments, analysis shows. BMJ. 2017; j965. [Crossref]

21. McKinstry B, Campbell J, Salisbury C. Telephone first consultations in primary care. BMJ. 2017; j4345. **[Crossref]**

22. Gellert GA, Kuszczyński K, Marcjasz N, Jaszczak J, Price T, Orzechowski PM. A comparative performance analysis of live clinical triage using rules-based triage protocols versus artificial intelligence-based automated virtual triage. J Hosp Adm [Internet]. 2023 [cited 2025 Jul 10]; 13: 8. **[Crossref]**

23. NHS Digital, NHS England, National Disease Registration Service (NDRS). Urgent suspected cancer referrals: referral, conversion and detection rates by geography, April 2022 to March 2023 [Internet]. 2024 [cited 2024 Nov 9]. **[Crossref]**

24. Meechan GT, Collins JP, Moss-Morris RE, Petrie KJ. Who is not reassured following benign diagnosis of breast symptoms? Psychooncology. 2005; 14: 239-246. (PMID: 15386770) **[Crossref]**

25. Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. Am J Sports Med. 2014; 42: 2371-2376. (PMID: 25073597) **[Crossref]**

26. Yao LH, Leung KC, Tsai CL, Huang CH, Fu LC. A novel deep learning-based system for triage in the emergency department using electronic medical records: retrospective cohort study. J Med Internet Res. 2021; 23: e27008. (PMID: 34958305) **[Crossref]**

27. Jones K, Lennon E, McCathie K, Millar A, Isles C, McFadyen A, et al. Teledermatology to reduce face-to-face appointments in general practice during the COVID-19 pandemic: a quality improvement project. BMJ Open Qual. 2022; 11: e001789. (PMID: 35618315) **[Crossref]**

28. Brewer F, Kentley J, Kalsi D, Mullarkey D, Thomas L, Morgan H, et al. BT10 (P025) community lesion imaging clinics for tele-triage of suspected skin cancer: implementation of a low-cost model in an NHS trust. Br J Dermatol [Internet]. 2024 [cited 2025 Jul 10]; 191(Suppl 1): i193-i94. **[Crossref]**

29. Yi PH, Singh D, Harvey SC, Hager GD, Mullen LA. DeepCAT: deep computer-aided triage of screening mammography. J Digit Imaging. 2021; 34: 27-35. (PMID: 33432446) **[Crossref]**

30. Jones K, Lennon E, McCathie K, Millar A, Isles C, McFadyen A, et al. Teledermatology to reduce face-to-face appointments in general practice during the COVID-19 pandemic: a quality improvement project. BMJ Open Qual. 2022; 11: e001789. (PMID: 35618315) **[Crossref]**

31. Buchard A, Bouvier B, Prando G, Beard R, Livieratos M, Busbridge D, et al. Learning medical triage from clinicians using Deep Q-Learning. 2020. **[Crossref]**

32. Porto BM. Improving triage performance in emergency departments using machine learning and natural language processing: a systematic review. BMC Emerg Med. 2024; 24: 219. (PMID: 39558255) **[Crossref]**

33. Lambe KA, O'Reilly G, Kelly BD, Curristan S. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. BMJ Qual Saf. 2016; 25: 808-820. (PMID: 26873253) **[Crossref]**

34. Spangler D, Hermansson T, Smekal D, Blomberg H. A validation of machine learning-based risk scores in the prehospital setting. PLoS One. 2019; 14: e0226518. (PMID: 31834920) **[Crossref]**

35. Mazo C, Kearns C, Mooney C, Gallagher WM. Clinical decision support systems in breast cancer: a systematic review. Cancers (Basel). 2020; 12: 369. (PMID: 32041094) **[Crossref]**

36. Verburg E, van Gils CH, van der Velden BHM, Bakker MF, Pijnappel RM, Veldhuis WB, et al. Validation of combined deep learning triaging and computer-aided diagnosis in 2901 breast MRI examinations from the second screening round of the dense tissue and early breast neoplasm screening trial. Invest Radiol. 2023; 58: 293-298. (PMID: 36256783) **[Crossref]**

37. Lång K, Josefsson V, Larsson AM, Larsson S, Högberg C, Sartor H, et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. Lancet Oncol. 2023; 24: 936-944. (PMID: 37541274) **[Crossref]**

38. Arıbal E. Future of breast radiology. Eur J Breast Health. 2023; 19: 262-266. (PMID: 37795010) **[Crossref]**