# A meta-analysis of diagnostic accuracy of machine learning models on mammography in breast cancer classification

**Tengku Muhammad Hanis**
Universiti Sains Malaysia
**Md Asiful Islam**
Universiti Sains Malaysia
**Kamarul Imran Musa** ( drkamarul@usm.my )
Universiti Sains Malaysia

**Research Article**

**Keywords:** AUC, HSROC, digital breast tomosynthesis (DBT), diagnostic radiology, cardiology, ophthalmology, pathology

**Posted Date:** November 5th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-970393/v1

## Abstract

In this meta-analysis, we aimed to estimate the diagnostic accuracy of machine learning models on digital mammograms and tomosynthesis in breast cancer classification and to assess factors affecting its diagnostic accuracy. We searched for related studies in Web of Science, Scopus, PubMed, Google Scholar and Embase. The studies were screened in two stages to exclude unrelated studies and duplicates. Finally, 36 studies containing 68 machine learning models were included in this meta-analysis. The area under the curve (AUC), hierarchical summary receiver operating characteristics (HSROC) curve, pooled sensitivity and pooled specificity were estimated using a bivariate Reitsma model. Overall AUC, pooled sensitivity and pooled specificity were 0.90 (95% CI: 0.85–0.90), 0.83 (95% CI: 0.78–0.87) and 0.84 (95% CI: 0.81–0.87), respectively. Additionally, the three significant covariates identified in this study were country (p=0.003), source (p=0.002) and classifier (p=0.016). Additionally, Deeks' linear regression test indicated that there exists a publication bias in the included studies (p=0.002). Thus, the results should be interpreted with caution.


## Introduction

Breast cancer is the most commonly diagnosed cancer overall and among women worldwide; in fact, it has been identified as the fifth leading cause of cancer-related mortality globally in 2020[1]. It is considered as the most prevalent cancer worldwide[2]. The screening and diagnosis of breast cancer are done using multiple assessments such as breast examination, mammography and biopsy. Mammography, which includes a digital mammogram and digital breast tomosynthesis (DBT), is the standard screening method for breast cancer. The digital mammogram is more commonly used; however, it is found to be less effective in patients with dense breasts and less sensitive to small tumours (tumours with a volume of less than 1 $mm^3$). On the other hand, DBT or the three-dimensional mammogram, which is a more advanced technology of mammography, overcomes these disadvantages. Overall, it provides higher diagnostic accuracy than the two-dimensional mammogram[4]. However, no significant difference was noted between these two technologies when used for screening purposes[5].

Machine learning is expected to improve the area of health care, especially in medical specialisations such as diagnostic radiology, cardiology, ophthalmology and pathology[6]. Factors such as the availability of big medical data and advances in computing technology will help accelerate the use of machine learning in these medical areas. However, in spite of these positive developments, the practical implementation of machine learning in a clinical setting remains to be debatable[7–9]. Issues such as privacy concerns, lack of trust in the technology, machine learning interpretability and unintended bias of the technology are yet to be fully explored[6,10–12]. The use of machine learning on medical images such as mammograms and tomosynthesis mainly aims to improve the diagnostic accuracy in these medical areas. However, machine learning may serve a different purpose based on the varying clinical setting. For example, machine learning may be used as a screening, diagnostic or prognostic tool. These different roles of machine learning will affect how the model is built and deployed; however, most studies do not clearly emphasise the role of their machine learning model with regard to the clinical context and its practical application.

Previous studies of machine learning on medical images associated with breast cancer mostly used digital mammograms[13], while the use of tomosynthesis was not very common. A wide variety of machine learning techniques had been used on these medical images, resulting in a wide range of diagnostic accuracy. Thus, this meta-analysis aims to establish the overall diagnostic accuracy of the machine learning model on digital mammograms and tomosynthesis. This study also aims to assess the factors affecting the diagnostic accuracy of the machine learning model and further perform subgroup analysis.

## Results

**Eligible studies.** In total, 2,897 research articles were identified in the 5 databases, as presented in Figure 1. After the removal of 1,115 duplicates, the remaining 1,782 articles were included in the screening process. A total of 1,346 articles were excluded during the whole screening process. The first screening process excluded 1,157 articles, while the second screening process excluded another 189 papers. Finally, 36 studies containing 68 machine learning models were included in this study.

**Study characteristics.** The main characteristics of the included studies are listed in Supplementary Table S1. The years of publication of the 36 included studies ranged from 2006 to 2020. Eleven studies used primary data from their respective countries, while most studies used secondary databases such as the Mammographic Image Analysis Society (MIAS), mini-MIAS and Digital Database for Screening Mammography (DDSM). Only one study used tomosynthesis images, while the remaining 35 used digital mammogram images. The three most common classifiers were neural network (23.5 %), support vector machine (22.1 %) and deep learning (20.6 %).

**Descriptive statistics.** The study with the highest accuracy was the one carried out by Acharya U et al. in 2008 (98.3 %), while that performed by Kanchanamani et al. in 2016 had the lowest accuracy (48.3 %). The specificity and sensitivity values of each machine learning model are presented in Figure 2. Sensitivity values for machine learning models in this study ranged between 0.03 (95 % CI: 0.00–0.24) and 1 (95 % CI: 0.98–1.00), while specificity values ranged between 0.37 (95 % CI: 0.25–0.50) and 0.98 (95 % CI: 0.93–1.00). In this study, significant differences were observed between sensitivity values ($p < 0.001$) and specificity values ($p < 0.001$) of machine learning models. The pooled diagnostic odds ratio

(DOR) of machine learning models was 28.34 (95 % CI: 17.67–45.45), with the DOR value of each model ranging from 0.90 (95 % CI: 0.44–1.84) to 7513.55 (95 % CI: 445.61–126689.03). Figure 3 presents the DOR values for each machine learning model in this study.

**Overall model.** The pooled area under the curve (AUC) estimated using the bivariate model of Reitsma et al.[14] for overall machine learning models in this study was 0.90 (95 % CI: 0.85–0.90). The HSROC curve plot is presented in Figure 4. Additionally, the pooled sensitivity and pooled specificity values estimated through the same model were 0.83 (95 % CI: 0.78–0.87) and 0.84 (95 % CI: 0.81–0.87), respectively.

**Test for heterogeneity and influential diagnostics.** Based on the HSROC curve plot (Figure 4), there was a moderate deviation of individual models from the curve. The correlation coefficient of the sensitivity and specificity was 0.33. Thus, there was an indication of slight-to-moderate heterogeneity in this study. However, influential diagnostics indicated that there was no influential model in the study. The result of the influential diagnostics is presented in Supplementary File 1.

**Subgroup analysis.** As per our findings, three out of four covariates were found to be significant via a likelihood ratio test; these were country (p = 0.003), source (p = 0.002) and classifier (p = 0.016), while the type of data was not significant (p = 0.121). The detailed result of the likelihood test is presented in Table 1. Thus, the country, source and classifier explained some of the heterogeneity that can be seen in the study. A further subgroup analysis was performed on the three significant covariates. All countries other than the USA and the UK were combined into one group due to the small number of available studies. Subsequently, studies that used data from both the USA and UK were excluded due to a small number of available studies, and those studies did not fit into any other group. Pairwise post hoc comparison of the country subgroup revealed that machine learning models that used data from the USA performed better than models that used data from the other countries in terms of AUC (dAUC = 0.095, 95 % CI: 0.044–0.191). Additionally, for the subgroup analysis of the classifier covariate, three classifiers that were dropped due to a small number of studies were the Gaussian mixture model (GMM), linear discriminant analysis (LDA) and logistic regression. The three significant pairwise comparisons for this subgroup analysis were the neural network and Bayes-based model (dAUC = 0.252, 95 % CI: 0.119–0.379), tree-based model and Bayes-based model (dAUC = 0.249, 95 % CI: 0.073–0.395) and support vector machine and Bayes-based model (dAUC = 0.219, 95 % CI: 0.094–0.350). Lastly, for the subgroup analysis of the source covariate, we dropped studies that used the INbreast database and the MMD database. We have also dropped studies that used both DDSM and MIAS databases and studies with unknown sources of data. Studies that used the MIAS and mini-MIAS databases were further classified into a single group. All pairwise comparisons of the AUC were determined to be not significant in this subgroup analysis. All aforementioned pairwise comparisons were significant after the Bonferroni correction, and there were six non-convergent pairwise comparisons. The results of complete pairwise comparisons for all the three subgroups are presented in Table 2, while Figure 5 delineates the HSROC for the subgroups. The highest AUCs in each subgroup were models with the US data (AUC = 0.94), models that used the DDSM database (AUC = 0.966) and the neural network model (0.938). As shown in Figure 5, models that used the DDSM database performed significantly better than models that used primary data, while other model comparisons were relatively similar to those in Table 2.

Table 1
A likelihood ratio test for bivariate meta-regression models with the null model. *Signifcance at P < 0.05.

| Model | Covariate | $\chi^2$-statistic (df) | P-value |
|---|---|---|---|
| Model 1 | Country | 19.55 (6) | 0.003* |
| Model 2 | Source | 31.10 (12) | 0.002* |
| Model 3 | Type of data | 4.23 (2) | 0.121 |
| Model 4 | Classifier | 30.32 (16) | 0.016* |

Table 2
A post hoc pairwise comparison for covariates country, source of data and classifier. *Signifcance at P < 0.05. **Signifcance after Bonferroni correction. †Non-convergence. ‡mini-MIAS and MIAS databases were combined into a group. §Others: Iran, Portugal, Jordan, China, Korea and Serbia.

| Comparisons | dAUC (95% CI) | P-value |
|---|---|---|
| Country | | |
| USA vs. UK | 0.051 (0.006, 0.127) | 0.035* |
| USA vs. others§ | 0.095 (0.044, 0.191) | 0.001** |
| UK vs. others§ | 0.044 (−0.034, 0.131) | 0.241 |
| Source of data | | |
| Primary data vs. DDSM | _† | _† |
| Primary data vs. MIAS‡ | −0.062 (−0.127, 0.023) | 0.152 |
| DDSM vs. MIAS‡ | _† | _† |
| Classifier | | |
| NN vs. DL | _† | _† |
| NN vs. Tree-based | 0.003 (−0.071, 0.138) | 0.946 |
| NN vs. KNN | 0.157 (0.026, 0.325) | 0.010 |
| NN vs. SVM | 0.033 (−0.034, 0.074) | 0.337 |
| NN vs. Bayes-based | 0.252 (0.119, 0.379) | <0.001** |
| DL vs. Tree-based | −0.016 (−0.122, 0.117) | 0.690 |
| DL vs. KNN | _† | _† |
| DL vs. SVM | _† | _† |
| DL vs. Bayes-based | _† | _† |
| Tree-based vs. KNN | 0.153 (−0.023, 0.333) | 0.082 |
| Tree-based vs. SVM | 0.030 (−0.101, 0.099) | 0.578 |
| Tree-based vs. Bayes-based | 0.249 (0.073, 0.395) | 0.007** |
| KNN vs. SVM | −0.123 (−0.300, −0.004) | 0.044* |
| KNN vs. Bayes-based | 0.096 (−0.121, 0.265) | 0.404 |
| SVM vs. Bayes-based | 0.219 (0.094, 0.350) | <0.001** |

**Publication bias.** Deeks' regression test was performed on the overall models that included all the 68 models from the 36 studies. The test indicated the possibility of publication bias in this study (p = 0.002). Figure 6 shows that Deeks' funnel plot was asymmetrical.

**Quality assessment.** Table 3 shows the quality assessment of the 36 included studies using the updated Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool. Generally, the majority of studies had unclear risk of bias and low applicability concerns. Additionally, several studies with a high risk of bias were observed under the subdomains of 'patient selection' and 'flow and timing' of the risk of bias domain. Most studies used secondary databases and did not explain in detail the data selection process and flow of their studies. Items such as the consecutive or random sampling approach, inappropriate exclusion of the data and the proper interval between the index test and the reference standard were not clearly addressed in most of the included studies. Overall, out of the 36 studies included in the meta-analysis, 2 studies were found to be of poor quality, 9 studies of good quality and 25 studies of moderate quality.

Table 3
Quality asessment of the included studies according to the QUADAS-2 tool.

| Study | Risk of Bias | | | | Applicability | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | Patient Selection | Index Test | Reference Standard | Flow and Timing | Patient Selection | Index Test | Reference Standard | |
| Abdolmaleki2006 | Low | Unclear | Low | Low | Low | Low | Low | Good |
| Acharyau2008 | High | Unclear | Low | Unclear | Low | Low | Low | Good |
| Al-antari2020 | Low | Unclear | Unclear | Low | Unclear | Low | Unclear | Moderate |
| Alfifi2020 | Unclear | Unclear | Unclear | Unclear | Low | Low | Unclear | Moderate |
| Al-hiary2012 | High | Low | Unclear | Unclear | Unclear | Low | Unclear | Moderate |
| Al-masni2018 | Low | Unclear | Low | Unclear | Low | Low | Low | Moderate |
| Bandeira-diniz2018 | High | Low | Low | Unclear | Low | Low | Low | Good |
| Barkana2017 | Unclear | Unclear | Low | Unclear | Unclear | Low | Low | Moderate |
| Biswas2019 | Unclear | Unclear | Unclear | Unclear | Unclear | Low | Unclear | Moderate |
| Cai2019 | Low | Low | Low | Low | Low | Low | Low | Moderate |
| Chen2019a | Low | Unclear | Low | Low | Low | Low | Low | Moderate |
| Chen2019b | Low | Low | Low | Low | Low | Low | Low | Good |
| Danala2018 | Low | Low | Low | Low | Low | Low | Low | Good |
| Daniellopez-cabrera2020 | Unclear | Unclear | Unclear | Unclear | Low | Low | Unclear | Good |
| Fathy2019 | High | Low | Low | Unclear | Low | Low | Low | Poor |
| Girija2019 | Unclear | Low | Unclear | Unclear | Low | Low | Low | Good |
| Jebamony2020 | Unclear | Unclear | Unclear | High | Low | Low | Unclear | Moderate |
| Junior2010 | High | Unclear | Unclear | High | Low | Low | Unclear | Moderate |
| Kanchanamani2016 | Unclear | Unclear | Unclear | Unclear | Low | Low | Unclear | Moderate |
| Kim2018 | Unclear | Low | Low | Low | Low | Low | Low | Moderate |
| Mao2019 | Low | Unclear | Low | Low | Low | Low | Low | Moderate |
| Miao2015 | Unclear | Unclear | Unclear | High | Low | Low | Unclear | Moderate |
| Miao2013 | Low | Low | Unclear | High | Low | Low | Unclear | Moderate |
| Milosevic2015 | Low | Unclear | Unclear | Unclear | Low | Low | Unclear | Moderate |
| Nithya2012 | Unclear | Unclear | Low | Unclear | Low | Low | Low | Moderate |
| Nusantara2016 | Unclear | Low | Unclear | Unclear | Low | Low | Low | Moderate |
| Palantei2017 | High | Unclear | Unclear | Unclear | Low | Low | Unclear | Poor |
| Paramkusham2018 | Unclear | Unclear | Low | Unclear | Low | Low | Low | Moderate |
| Roseline2018 | Unclear | Unclear | Unclear | High | Low | Low | Unclear | Moderate |
| Shah2015 | Unclear | Unclear | Unclear | Unclear | Low | Low | Unclear | Good |
| Shivhare2020 | Unclear | Unclear | Unclear | High | Low | Low | Unclear | Good |
| Singh2018 | Unclear | Unclear | Low | Low | Low | Low | Low | Moderate |
| Venkata2019 | Unclear | Unclear | Unclear | Unclear | Unclear | Low | Unclear | Moderate |
| Wang2017 | High | Unclear | Unclear | Unclear | Low | Low | Unclear | Moderate |
| Wutsqa2017 | High | Unclear | Unclear | Unclear | Low | Low | Unclear | Moderate |
| Yousefi2018 | Unclear | Unclear | Low | Unclear | Low | Low | Low | Moderate |

## Discussion

This study presents the efficacy of machine learning models on digital mammograms and tomosynthesis. According to our findings, machine learning models had a good performance in breast cancer classification using digital mammograms and tomosynthesis, with pooled AUC of 0.9. Several meta-analysis studies that assessed the diagnostic accuracy of machine learning models on MRI in gliomas, prostate cancer and meningioma reported lower AUCs of 0.88, 0.86 and 0.75, respectively[15–17]. The good performance of machine learning used on medical images supported a promising potential to be used in clinical settings, especially as a screening tool and a supplementary diagnostic tool to a radiologist.

Inconsistency among the diagnostic accuracy studies is to be expected[18]. In this meta-analysis, the three covariates that may explain the inconsistency among the studies were country, source and classifier. In terms of country, studies that used data from the USA and the UK had higher AUCs compared to the other countries (others group); however, only a pairwise comparison of the USA and other countries revealed a statistically significant result. This significant result may indicate a difference in characteristics between patients with breast cancer across countries. For example, breast cancer presentation and breast density had been reported to vary across populations[19,20], which, in turn, could affect the diagnostic accuracy of machine learning models. Additionally, this study found that studies that used primary data had lower AUCs compared to studies that used secondary databases. The studies that used primary data may reflect the actual diagnostic accuracy of machine models in real practice as the data was collected specifically for the studies in question. Lastly, this study found that the classifier with the best AUC was the neural network, followed by the tree-based classifier and deep learning. However, the confidence regions of all these three models overlapped with each other (Figure 5), which indicated that none of the machine learning models significantly outperformed the other in terms of breast cancer classification. It is worth noting that one of the findings of this study was that the Bayes-based machine learning model had the lowest AUC (0.69) and performed significantly less than the neural network, tree-based model and support vector machine. Nevertheless, a few studies were dropped in each subgroup analysis due to a small number of studies in that particular group, which limited the pairwise comparison that could be performed in each subgroup analysis.

In this study, we established the good performance of machine learning models on mammography in the classification of breast cancer. We used the bivariate model to estimate the AUC and further applied a bootstrap method to estimate its confidence interval. Furthermore, our meta-analysis included a reasonable number of studies to give a relatively reliable result on the primary outcome and secondary outcomes. However, our study had several limitations. Firstly, we found that our study had a potential publication bias. One of the probable causes was unpublished studies with a low-performance model. Additionally, the overall model in this study had a moderate amount of heterogeneity, and this study included a considerable number of studies that may contribute to both the occurrence of publication bias and the high statistical power of the asymmetry test. As shown in Figure 6, model 10 had a much higher DOR compared to the other models on the right side of the figure; however, removing this model did not have a significant impact on the AUC (Supplementary File 1). Nonetheless, the mechanism of publication bias in diagnostic accuracy studies remains unclear, and a robust assessment of this bias is yet to be proposed[21]. Future meta-analyses may consider including the preprint articles that may be able to reduce the publication bias. Secondly, we only had one study with tomosynthesis, while the rest of the studies were using digital mammograms. Thus, the findings of our study were more inclined towards digital mammograms more than tomosynthesis, although both are considered as mammography technology. Lastly, we limited the language of the included studies to English, which may have increased the risk of bias in our findings.

In conclusion, the performance of machine learning on mammography in breast cancer classification showed promising results, with good sensitivity and specificity values. However, the role of any machine learning technique in the diagnostic pathway should be clearly explained in a diagnostic accuracy study to be efficiently incorporated in the clinical setting. Thus, the limitation of each machine learning model will be apparent to clinicians and other health personnel.

## Methods

**Overview.** This study was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses of diagnostic test accuracy studies (PRISMA-DTA)[22] and Synthesising Evidence from Diagnostic Accuracy Tests (SEDATE)[18] guidelines and recommendations. Both checklists are presented in Supplementary File 2.

**Search strategy**. We searched the Web of Science, Scopus, PubMed, Google Scholar and Embase using predetermined search terms. The search was carried out in August 2020. All search terms for each database are presented in Supplementary File 3.

All the results were imported into Mendeley. Duplicate papers were automatically screened and deleted. Subsequently, a researcher (TMH) manually screened the results again and deleted the remaining duplicates that were not identified using Mendeley. We then divided the screening process into two phases. In the first phase, we applied more lenient selection criteria to screen out the more obvious articles that were not related to our study. A full text of all articles that passed the first phase of selection criteria was downloaded. Additionally, in the second phase, we applied more stringent selection criteria to the articles to fit our study's objectives. Any inconsistency during the selection and extraction process was resolved by discussion and consensus among the researchers.

**Selection criteria.** We divided the screening process into two phases. We mainly screened the titles and abstracts and, if needed, the full text in the first phase. We searched for the following groups of articles in the first phase: (1) articles related to breast cancer prediction or classification; (2) articles that used machine learning models or algorithms; (3) articles written in English; (4) peer-reviewed research articles, proceedings and theses were excluded; (5) articles that used digital mammogram or tomosynthesis data; and (6) articles must at least reported an accuracy value as a performance metrics.

We screened all articles on full text in the second phase of the selection process. We selected the articles based on the following criteria: (1) Articles that focused only on breast cancer classification models. Articles that compared between feature extraction and segmentation methods were excluded. (2) Articles that reported a confusion matrix or at least had reported sufficient data. (3) Articles that had ensembles or hybrid machine learning models as classifiers were excluded. (4) Three-class prediction models were excluded unless a 2 x 2 confusion matrix was reported.

**Data extraction.** We collectively extracted data from included articles into a Microsoft Excel spreadsheet. The extracted variables were as follows: (1) title; (2) first author's last name; (3) year of publication; (4) source of data; (5) country of the data used; (6) type of data; (7) sample size used; (8) classifier; (9) prediction class; (10) accuracy; (11) sensitivity; (12) specificity; and (13) confusion matrix. Additionally, more than one model was extracted from an article if the models used different data, classifiers or prediction classes. However, the model based on accuracy was extracted in the case of articles with relatively similar models.

**Quality assessment.** We used the QUADAS-2[23] tool to assess the quality of the studies that were included in the meta-analysis. The tool consisted of two main domains, that is, the risk of bias and the applicability concerns. Each domain had four and three subdomains, respectively. Each subdomain could be rated as 'no', 'unclear' or 'yes'. The domains were considered a low risk of bias if all subdomains had a rating of 'yes'. However, the domains were considered a high risk of bias if any of the subdomains had one 'no' and no 'yes'. The domains, except for the previous two conditions, were considered an unclear risk of bias. Additionally, we added the overall rating to the QUADAS-2 assessment. We assigned the figures 1, 0 and −1, to low, unclear and high, respectively. Thus, the sum of the overall rating could range from −7 to 7. The overall quality was classified as very poor (−7 to −4), poor (−3 to 0), moderate (1 to 4) and good (5 to 7).

**Outcomes.** The primary outcomes were the overall diagnostic accuracy of the machine learning model in the form of the AUC and the hierarchical summary receiver operating characteristics (HSROC) curve. The secondary outcomes were the result of a likelihood ratio test for variables classifier, country of the data, source of data and type of the data. Variables with a p-value < 0.05 were considered statistically significant and followed up by a post hoc subgroup analysis.

**Statistical analysis.** The statistical analysis was done using R version 4.1.0[24]. The full R code is available on the GitHub website (https://github.com/tengku-hanis/MA-ML-BC). The main R packages used were mada and metafor[25,26]. A continuity correction of 0.5 was applied to the data if there were zero cells in the confusion matrix to avoid statistical artefacts. This approach is the default setting in the mada package. Each machine learning model was summarised by the pooled DOR, sensitivity and specificity. The DOR represents the odds of a positive test result in diseased individuals compared to the odds of a positive result in healthy individuals. Thus, the DOR simply denotes the discriminant ability of the diagnostic test. Additionally, sensitivity represents the ability of the test to correctly identify affected individuals, while specificity reflects the ability of the test to correctly identify healthy individuals among the tested individuals. The pooled sensitivity, pooled specificity, AUC and HSROC curve parameters were estimated using the bivariate model of Reitsma et al.[14] through the mada package. The 95 % confidence interval of the AUC was estimated using a bootstrap method from the dmetatools package[27]. Heterogeneity assessment was done through visual inspection of the HSROC plot and the correlation between sensitivity and specificity. Inconsistency was suspected if the individual studies largely deviated from the HSROC line and the coefficient correlation of sensitivity and specificity was larger than zero[18,28]. The Cochran's Q test and Higgins' $I^2$ statistics were not presented as they were not suitable for heterogeneity assessment in diagnostic test accuracy studies[29].

A likelihood ratio test between bivariate meta-regression models was done to compare a null model and a model with a covariate. Five bivariate meta-regression models were built, including the null model and models with a covariate of country, source, type of data and classifier. The likelihood ratio test with a p-value < 0.05 indicated that the model with a variable was better; thus, the variable was statistically significant. Subsequently, a post hoc subgroup analysis was performed for each significant variable. Pairwise comparisons of the AUC between each model of the subgroups were performed using a bootstrap method in the dmetatools package, and p-values were adjusted using the Bonferroni correction. A p-value below a threshold of 0.05 divided by the number of groups in each subgroup analysis indicated a significant comparison. A non-convergent result indicated that the model did not converge even after 10,000 bootstrap resampling. Any subgroup model with a small number of studies was dropped from the subgroup analysis as the estimates of the AUC and HSROC parameters were not reliable.

An influential diagnostic analysis was performed to assess the overall diagnostic accuracy of the machine learning model using the dmetatools package. The influential diagnostics was done using leave-one-out analysis to estimate the difference in the AUC. Publication bias was evaluated using Deeks' regression test[30]. The approach of Deeks et al. had been considered as the most appropriate one to assess the publication bias in a diagnostic test accuracy study[21]. P-values < 0.10 may indicate the presence of publication bias.

# Data Availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

# Code Availability

The full R code is available on the GitHub website (https://github.com/tengku-hanis/MA-ML-BC).

## Declarations

## Data Availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files).

## Code Availability

The full R code is available on the GitHub website (https://github.com/tengku-hanis/MA-ML-BC).

## Acknowledgements

This study and its publication are supported by the School of Medical Sciences, Universiti Sains Malaysia, and the Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education, Malaysia (FRGS/1/2019/SKK03/USM/02/1). The funders had no role in the design of the study, data collection, data analysis, interpretation of the data or writing of the manuscript.

## Author contributions

K.I.M., T.M.H. and M.A.I. design and download the relevant papers, T.M.H. extracted the data and performed the meta-analysis, K.I.M., T.M.H. and M.A.I. wrote the manuscript.

## Competing interests

The authors declare no competing interests.

## References

1. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).

2. World Health Organization. Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

3. Wang, L. Early diagnosis of breast cancer. *Sensors* **17**, 1572 (2017).

4. Gilbert, F. J. & Pinker-Domening, K. Diagnosis and staging of breast cancer: when and how to use mammography, tomosynthesis, ultrasound, contrast-enhanced mammography, and magnetic resonance imaging. in *Diseases of the Chest, Breast, Heart and Vessels 2019–2022 Diagnostic and Interventional Imaging* (eds. Hodler, J., Kubik-Huch, R. A. & Von Schulthess, G. K.) 155–166 (2019). doi:10.1007/978-3-030-11149-6.

5. Hofvind, S. *et al.* Two-view digital breast tomosynthesis versus digital mammography in a population-based breast cancer screening programme (To-Be): a randomised, controlled trial. *Lancet Oncol.* **20**, 795–805 (2019).

6. Ahuja, A. S. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* **7**, e7702 (2019).

7. Abdullah, R. & Fakieh, B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. *J. Med. Internet Res.* **22**, 1–8 (2020).

8. Doraiswamy, P. M., Blease, C. & Bodner, K. Artificial intelligence and the future of psychiatry: Insights from a global physician survey. *Artif. Intell. Med.* **102**, 101753 (2020).

9. Blease, C. *et al.* Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *J. Med. Internet Res.* **21**, 1–10 (2019).

10. Meskó, B. & Görög, M. A short guide for medical professionals in the era of artificial intelligence. *npj Digit. Med.* **3**, 126 (2020).

11. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).

12. Asan, O., Bayrak, A. E. & Choudhury, A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J. Med. Internet Res.* **22**, 1–7 (2020).

13. Yassin, N. I. R., Omran, S., El Houby, E. M. F. & Allam, H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. *Comput. Methods Programs Biomed.* **156**, 25–45 (2018).

14. Reitsma, J. B. *et al.* Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J. Clin. Epidemiol.* **58**, 982–990 (2005).

15. Cuocolo, R. *et al.* Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. *Eur. Radiol.* **30**, 6877–6887 (2020).

16. van Kempen, E. J. *et al.* Accuracy of machine learning algorithms for the classification of molecular features of gliomas on MRI: a systematic literature review and meta-analysis. *Cancers (Basel).* **13**, 2606 (2021).

17. Ugga, L. *et al.* Meningioma MRI radiomics and machine learning: systematic review, quality score assessment, and meta-analysis. *Neuroradiology* (2021) doi:10.1007/s00234-021-02668-0.

18. Sotiriadis, A., Papatheodorou, S. I. & Martins, W. P. Synthesizing evidence from diagnostic accuracy tests: the SEDATE guideline. *Ultrasound Obstet. Gynecol.* **47**, 386–395 (2016).

19. Tehranifar, P., Rodriguez, C. B., April-Sanders, A. K., Desperito, E. & Schmitt, K. M. Migration history, language acculturation, and mammographic breast density. *Cancer Epidemiol. Biomarkers Prev.* **27**, 566–574 (2018).

20. Vieira, R., Biller, G., Uemura, G., Ruiz, C. & Curado, M. Breast cancer screening in developing countries. *Clinics* **72**, 244–253 (2017).

21. van Enst, W. A., Ochodo, E., Scholten, R. J., Hooft, L. & Leeflang, M. M. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC Med. Res. Methodol.* **14**, 70 (2014).

22. McInnes, M. D. F. *et al.* Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies. *JAMA* **319**, 388 (2018).

23. Reitsma, J. B. *et al.* QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **155**, 529–536 (2011).

24. R Core Team. R: a language and environment for statistical computing. (2021).

25. Doebler, P. MADA: meta-analysis of diagnostic accuracy. (2020).

26. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).

27. Noma, H., Matsushima, Y. & Ishii, R. Confidence interval for the AUC of SROC curve and some related methods using bootstrap for meta-analysis of diagnostic accuracy studies. *Commun. Stat. Case Stud. Data Anal. Appl.* 1–15 (2021) doi:10.1080/23737484.2021.1894408.

28. Shim, S. R., Kim, S.-J. & Lee, J. Diagnostic test accuracy: application and practice using R software. *Epidemiol. Health* **41**, 1–8 (2019).

29. Lee, J., Kim, K. W., Choi, S. H., Huh, J. & Park, S. H. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: a practical review for clinical researchers-Part II. Statistical methods of meta-analysis. *Korean J. Radiol.* **16**, 1188 (2015).

30. Deeks, J. J., Macaskill, P. & Irwig, L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J. Clin. Epidemiol.* **58**, 882–893 (2005).

31. Abdolmaleki, P., Guiti, M. & Tahmasebi, M. Neural network analysis of breast cancer from mammographic evaluation. *Iran. J. Radiol.* **3**, 155–162 (2006).

32. Acharya U, R., Ng, E. Y. K., Chang, Y. H., Yang, J. & Kaw, G. J. L. Computer-based identification of breast cancer using digitized mammograms. *J. Med. Syst.* **32**, 499–507 (2008).

33. Al-Antari, M. A., Han, S.-M. & Kim, T.-S. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms. *Comput. Methods Programs Biomed.* **196**, 105584 (2020).

34. Alfifi, M., Shady, M., Bataineh, S. & Mezher, M. Enhanced artificial intelligence system for diagnosing and predicting breast cancer using deep learning. *Int. J. Adv. Comput. Sci. Appl.* **11**, 498–513 (2020).

35. Al-Hiary, H., Alhadidi, B. & Braik, M. An implemented approach for potentially breast cancer detection using extracted features and artificial neural networks. *Comput. Informatics* **31**, 225–244 (2012).

36. Al-masni, M. A. *et al.* Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. *Comput. Methods Programs Biomed.* **157**, 85–94 (2018).

37. Bandeira Diniz, J. O. *et al.* Detection of mass regions in mammograms by bilateral analysis adapted to breast density using similarity indexes and convolutional neural networks. *Comput. Methods Programs Biomed.* **156**, 191–207 (2018).

38. Barkana, B. D. & Saricicek, I. Classification of breast masses in mammograms using 2D homomorphic transform features and supervised classifiers. *J. Med. Imaging Heal. Informatics* **7**, 1566–1571 (2017).

39. Biswas, R., Roy, S. & Biswas, A. Mammogram classification using curvelet coefficients and gray level co-occurrence matrix for detection of breast cancer. *Int. J. Innov. Technol. Explor. Eng.* **8**, 4819–4824 (2019).

40. Cai, H. *et al.* Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. *Comput. Math. Methods Med.* **2019**, 2717454 (2019).

41. Chen, S. *et al.* A new application of multimodality radiomics improves diagnostic accuracy of nonpalpable breast lesions in patients with microcalcifications-only in mammography. *Med. Sci. Monit.* **25**, 9786–9793 (2019).

42. Chen, X. *et al.* Applying a new quantitative image analysis scheme based on global mammographic features to assist diagnosis of breast cancer. *Comput. Methods Programs Biomed.* **179**, 104995 (2019).

43. Danala, G. *et al.* Classification of breast masses using a computer-aided diagnosis scheme of contrast enhanced digital mammograms. *Ann. Biomed. Eng.* **46**, 1419–1431 (2018).

44. Daniel López-Cabrera, J. *et al.* Classification of Breast Cancer from Digital Mammography Using Deep Learning. *Intel. Artif.* **23**, 56–66 (2020).

45. Fathy, W. E. & Ghoneim, A. S. A deep learning approach for breast cancer mass detection. *Int. J. Adv. Comput. Sci. Appl.* **10**, 175–182 (2019).

46. Girija, O. K. & Sudheep Elayiodm, M. Hybrid method of local binary pattern and classification tree for early breast cancer detection by mammogram classification. *Int. J. Recent Technol. Eng.* **8**, 139–145 (2019).

47. Jebamony, J. & Jacob, D. Classification of benign and malignant breast masses on mammograms for large datasets using core vector machines. *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.* **16**, 703–710 (2020).

48. Junior, G. B., Martins, L. D. O., Silva, A. C. & Paiva, A. C. Comparison of support vector machines and bayesian neural networks performance for breast tissues using geostatistical functions in mammographic images. *Int. J. Comput. Intell. Appl.* **09**, 271–288 (2010).

49. Kanchanamani, M. & Perumal, V. Performance evaluation and comparative analysis of various machine learning techniques for diagnosis of breast cancer. *Biomed. Res.* **27**, 623–631 (2016).

50. Kim, E.-K. E.-K. *et al.* Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. *Sci. Rep.* **8**, 2762 (2018).

51. Mao, N. *et al.* Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study. *J. Am. Coll. Radiol.* **16**, 485–491 (2019).

52. Miao, J. H., Miao, K. H. & Miao, G. J. Breast cancer biopsy predictions based on mammographic diagnosis using support vector machine learning. *Cyber Journals Multidiscip. Journals Sci. Technol. J. Sel. Areas Bioinforma.* **5**, (2015).

53. Miao, K. H. & Miao, G. J. Mammographic diagnosis for breast cancer biopsy predictions using neural network classification model and receiver operating characteristic (ROC) curve evaluation. *J. Sel. Area Bioinforma.* (2013).

54. Milosevic, M., Jankovic, D. & Peulic, A. Comparative analysis of breast cancer detection in mammograms and thermograms. *Biomed. Tech.* **60**, 49–56 (2015).

55. Nithya, R. & Santhi, B. Breast cancer diagnosis in digital mammogram using statistical features and neural network. *Res. J. Appl. Sci. Eng. Technol.* **4**, 5480–5483 (2012).

56. Nusantara, A. C., Purwanti, E. & Soelistiono, S. Classification of digital mammogram based on nearest-neighbor method for breast cancer detection. *Int. J. Technol.* **1**, 71–77 (2016).

57. Palantei, E., Amaliah, A. & Amirullah, I. Breast cancer detection in mammogram images exploiting GLCM, GA features and SVM algorithms. *J. Telecommun. Electron. Comput. Eng.* **9**, 113–117 (2017).

58. Paramkusham, S., Rao, K. M. M., Prabhakar Rao, B. V. V. S. N. & Sharma, S. Application of TAR signature for breast mass analysis. *Biomed. Res.* **29**, 2030–2034 (2018).

59. Roseline, R. & Manikandan, S. Determination of breast cancer using knn cluster technique. *Indian J. Public Heal. Res. Dev.* **9**, 418–423 (2018).

60. Shah, H. Automatic classification of breast masses for diagnosis of breast cancer in digital mammograms using neural network. *Int. J. Sci. Technol. Eng.* **1**, (2015).

61. Shivhare, E. & Saxena, V. (Nigam). Breast cancer diagnosis from mammographic images using optimized feature selection and neural network architecture. *Int. J. Imaging Syst. Technol.* ima.22467 (2020) doi:10.1002/ima.22467.

62. Singh, L. & Jaffery, Z. A. Computer-aided diagnosis of breast cancer in digital mammograms. *Int. J. Biomed. Eng. Technol.* **27**, 233–246 (2018).

63. Venkata, M. D. & Lingamgunta, S. Triple-modality breast cancer diagnosis and analysis in middle aged women by logistic regression. *Int. J. Innov. Technol. Explor. Eng.* **8**, 555–562 (2019).

64. Wang, S. *et al*. Abnormal breast detection in mammogram images by feed-forward neural network trained by jaya algorithm. *Fundam. Informaticae* **151**, 191–211 (2017).

65. Wutsqa, D. U. & Setiadi, R. P. Point operation to enhance the performance of fuzzy neural network model for breast cancer classification. *J. Eng. Appl. Sci.* **12**, 4405–4410 (2017).

66. Yousefi, M., Krzyżak, A. & Suen, C. Y. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. *Comput. Biol. Med.* **96**, 283–293 (2018).

## Supplemental Table

Supplementary Table S1. Characteristics of included studies.

| Study | ID | Country | Source | Type of data | Classifier | Prediction class | TP | TN | FP | FN | Accuracy |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Abdolmaleki2006[31] | 1 | Iran | Primary data | DM | NN | Benign – Malignant | 16 | 14 | 8 | 2 | 0.75 |
| Acharyau2008[32] | 2 | USA | DDSM | DM | NN | Normal – Benign – Malignant | 55 | 28 | 2 | 5 | 0.97 |
| | 3 | USA | DDSM | DM | GMM | Normal – Benign – Malignant | 57 | 29 | 1 | 3 | 0.98 |
| Al-antari2020[33] | 4 | USA | DDSM | DM | DL | Benign – Malignant | 59 | 59 | 1 | 1 | 0.98 |
| | 5 | Portugal | INbreast | DM | DL | Benign – Malignant | 14 | 6 | 2 | 0 | 0.95 |
| Alfifi2020[34] | 6 | UK | MIAS | DM | DL | Normal – Benign – Malignant | 124 | 66 | 7 | 3 | 0.95 |
| | 7 | UK | MIAS | DM | Tree-based | Normal – Benign – Malignant | 102 | 54 | 29 | 15 | 0.78 |
| | 8 | UK | MIAS | DM | KNN | Normal – Benign – Malignant | 99 | 50 | 32 | 19 | 0.74 |
| Al-hiary2012[35] | 9 | Jordan | Primary data | DM | NN | Normal – Cancer | 14 | 15 | 1 | 2 | 0.91 |
| Al-masni2018[36] | 10 | USA | DDSM | DM | NN | Benign – Malignant | 240 | 226 | 14 | 0 | 0.97 |
| Andeira-diniz2018[37] | 11 | USA | DDSM | DM | DL | Non-mass – Mass | 2418 | 4306 | 442 | 225 | 0.91 |
| | 12 | USA | DDSM | DM | DL | Non-mass – Mass | 1774 | 5615 | 210 | 188 | 0.95 |
| Barkana2017[38] | 13 | USA | DDSM | DM | NN | Benign – Malignant | 325 | 270 | 70 | 57 | 0.82 |
| | 14 | USA | DDSM | DM | SVM | Benign – Malignant | 318 | 278 | 62 | 64 | 0.83 |
| Biswas2019[39] | 15 | UK | MIAS | DM | NN | Normal – Abnormal | 32 | 12 | 3 | 1 | 0.92 |
| Cai2019[40] | 16 | China | Primary data | DM | SVM | Benign – Malignant | 48 | 39 | 6 | 6 | 0.89 |
| Chen2019a[41] | 17 | China | Primary data | DM | Tree-based | Benign – Malignant | 31 | 30 | 11 | 9 | 0.75 |
| Chen2019b[42] | 18 | USA | Primary data | DM | SVM | Benign – Malignant | 102 | 104 | 37 | 32 | 0.75 |
| | 19 | USA | Primary data | DM | SVM | Benign – Malignant | 103 | 114 | 27 | 31 | 0.79 |
| Danala2018[43] | 20 | USA | Primary data | DM | DL | Benign – Malignant | 63 | 24 | 9 | 15 | 0.78 |
| | 21 | USA | Primary data | DM | DL | Benign – Malignant | 55 | 21 | 12 | 23 | 0.68 |
| Daniellopez-cabrera2020[44] | 22 | UK | mini-MIAS | DM | DL | Normal – Abnormal | 31 | 101 | 2 | 4 | 0.97 |
| | 23 | UK | mini-MIAS | DM | DL | Benign – Malignant | 14 | 28 | 3 | 1 | 0.91 |
| Fathy2019[45] | 24 | USA | DDSM | DM | DL | Normal – Abnormal | 389 | 325 | 71 | 1 | 0.91 |
| Girija2019[46] | 25 | UK | mini-MIAS | DM | Tree-based | Normal – Abnormal | 266 | 48 | 4 | 4 | 0.98 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 26 | UK | mini-MIAS | DM | Tree-based | Benign – Malignant | 200 | 55 | 6 | 9 | 0.94 |
| Jebamony2020[47] | 27 | UK | mini-MIAS | DM | NN | Benign – Malignant | 33 | 41 | 12 | 5 | 0.85 |
| | 28 | UK | mini-MIAS | DM | SVM | Benign – Malignant | 37 | 49 | 4 | 1 | 0.96 |
| Junior2010[48] | 29 | UK | mini-MIAS | DM | NN | Normal – Abnormal | 16 | 69 | 5 | 18 | 0.79 |
| | 30 | UK | mini-MIAS | DM | SVM | Normal – Abnormal | 20 | 80 | 1 | 7 | 0.93 |
| Kanchanamani2016[49] | 31 | UK | MIAS | DM | SVM | Normal – Abnormal | 46 | 120 | 24 | 0 | 0.87 |
| | 32 | UK | MIAS | DM | Bayes-based | Normal – Abnormal | 30 | 94 | 50 | 16 | 0.65 |
| | 33 | UK | MIAS | DM | DL | Normal – Abnormal | 23 | 101 | 43 | 23 | 0.65 |
| | 34 | UK | MIAS | DM | KNN | Normal – Abnormal | 28 | 112 | 32 | 18 | 0.74 |
| | 35 | UK | MIAS | DM | LDA | Normal – Abnormal | 28 | 112 | 32 | 18 | 0.74 |
| | 36 | UK | MIAS | DM | SVM | Benign – Malignant | 58 | 53 | 2 | 7 | 0.93 |
| | 37 | UK | MIAS | DM | Bayes-based | Benign – Malignant | 50 | 20 | 35 | 15 | 0.58 |
| | 38 | UK | MIAS | DM | DL | Benign – Malignant | 29 | 29 | 26 | 36 | 0.48 |
| | 39 | UK | MIAS | DM | KNN | Benign – Malignant | 41 | 25 | 30 | 24 | 0.55 |
| | 40 | UK | MIAS | DM | LDA | Benign – Malignant | 38 | 33 | 22 | 27 | 0.59 |
| Kim2018[50] | 41 | Korea | Primary data | DM | DL | Normal – Abnormal | 471 | 548 | 71 | 148 | 0.82 |
| Mao2019[51] | 42 | China | Primary data | DM | SVM | Benign – Malignant | 13 | 14 | 1 | 7 | 0.80 |
| | 43 | China | Primary data | DM | Logistic | Benign – Malignant | 17 | 14 | 1 | 3 | 0.89 |
| | 44 | China | Primary data | DM | KNN | Benign – Malignant | 8 | 14 | 1 | 12 | 0.83 |
| | 45 | China | Primary data | DM | Bayes-based | Benign – Malignant | 9 | 13 | 2 | 11 | 0.78 |
| Miao2015[52] | 46 | USA | MMD | DM | SVM | Benign – Malignant | 381 | 399 | 28 | 22 | 0.94 |
| Miao2013[53] | 47 | USA | MMD | DM | NN | Benign – Malignant | 360 | 384 | 43 | 43 | 0.90 |
| Milosevic2015[54] | 48 | UK | MIAS | DM | SVM | Normal – Abnormal | 23 | 163 | 24 | 90 | 0.62 |
| | 49 | UK | MIAS | DM | KNN | Normal – Abnormal | 44 | 138 | 49 | 69 | 0.61 |
| | 50 | UK | MIAS | DM | Bayes-based | Normal – Abnormal | 53 | 113 | 74 | 60 | 0.55 |
| | 51 | Serbia | Primary data | DM | SVM | Normal – Abnormal | 121 | 130 | 20 | 29 | 0.84 |
| | 52 | Serbia | Primary data | DM | KNN | Normal – Abnormal | 84 | 79 | 71 | 66 | 0.54 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 53 | Serbia | Primary data | DM | Bayes-based | Normal – Abnormal | 114 | 118 | 32 | 36 | 0.77 |
| Nithya2012[55] | 54 | USA | DDSM | DM | NN | Normal – Abnormal | 23 | 24 | 2 | 1 | 0.94 |
| Nusantara2016[56] | 55 | UK | MIAS | DM | KNN | Normal – Abnormal | 10 | 20 | 0 | 1 | 0.97 |
| Palantei2017[57] | 56 | UK | MIAS | DM | SVM | Normal – Abnormal | 9 | 21 | 4 | 0 | 0.88 |
| Paramkusham2018[58] | 57 | USA | DDSM | DM | SVM | Benign – Malignant | 10 | 10 | 1 | 1 | 0.91 |
| Roseline2018[59] | 58 | UK | MIAS | DM | KNN | Benign – Malignant | 49 | 60 | 4 | 2 | 0.95 |
| Shah2015[60] | 59 | UK | MIAS | DM | NN | Normal – Abnormal | 54 | 49 | 2 | 3 | 0.95 |
| | 60 | UK | MIAS | DM | NN | Benign – Malignant | 24 | 22 | 2 | 6 | 0.85 |
| Shivhare2020[61] | 61 | USA, UK | DDSM, MIAS | DM | NN | Benign – Malignant | 12 | 16 | 2 | 3 | 0.85 |
| | 62 | USA, UK | DDSM, MIAS | DM | DL | Benign – Malignant | 1 | 17 | 1 | 14 | 0.55 |
| | 63 | USA, UK | DDSM, MIAS | DM | SVM | Benign – Malignant | 0 | 18 | 0 | 15 | 0.55 |
| Singh2018[62] | 64 | UK | MIAS | DM | NN | Benign – Malignant | 25 | 14 | 1 | 2 | 0.93 |
| Venkata2019[63] | 65 | NA | NA | DM | Logistic regression | Benign – Malignant | 14 | 14 | 1 | 1 | 0.93 |
| Wang2017[64] | 66 | UK | mini-MIAS | DM | NN | Normal – Abnormal | 92 | 92 | 8 | 8 | 0.92 |
| Wutsqa2017[65] | 67 | UK | MIAS | DM | NN | Normal – Abnormal | 14 | 8 | 0 | 2 | 0.92 |
| Yousefi2018[66] | 68 | USA | Primary data | Tomosynthesis | Tree-based | Benign – Malignant | 11 | 13 | 2 | 2 | 0.87 |

DM, digital mammogram; NN, neural network; GMM, Gaussian mixture model; DL, deep learning; KNN, K-nearest neighbour; SVM, support vector machine; LDA, linear discriminant analysis; NA, not available; TP, true positive; TN, true negative; FP, false positive; FN, false negative

## Figures

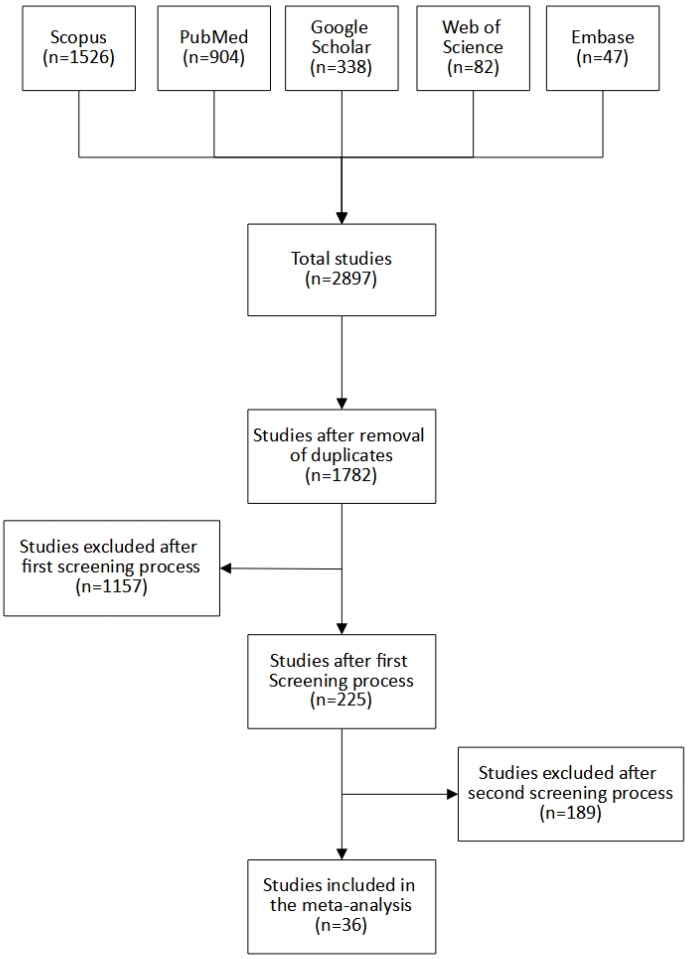


**Figure 1**

Flow diagram of the study selection process.

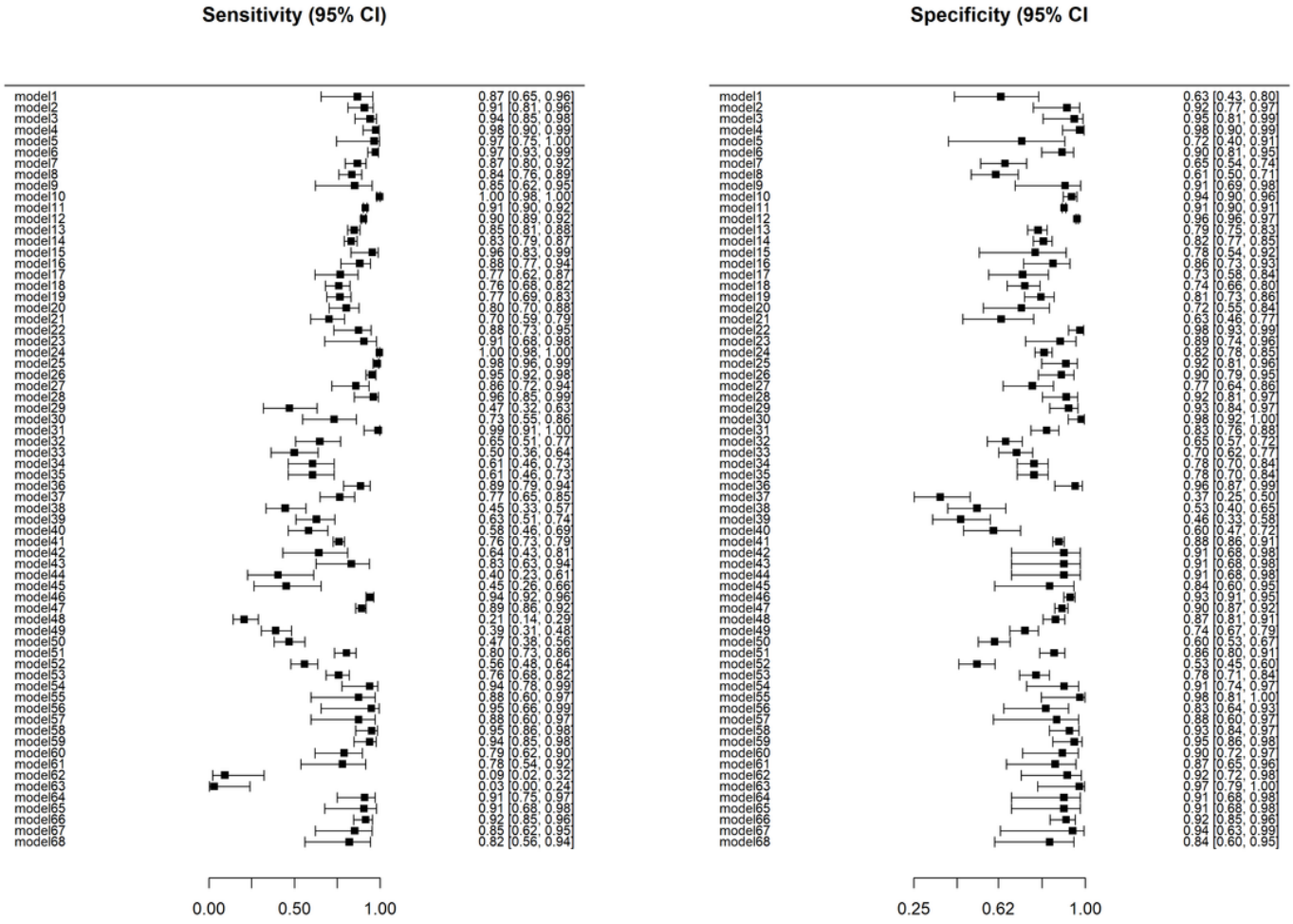


**Figure 2**

Sensitivity and specificity of machine learning models in the study.

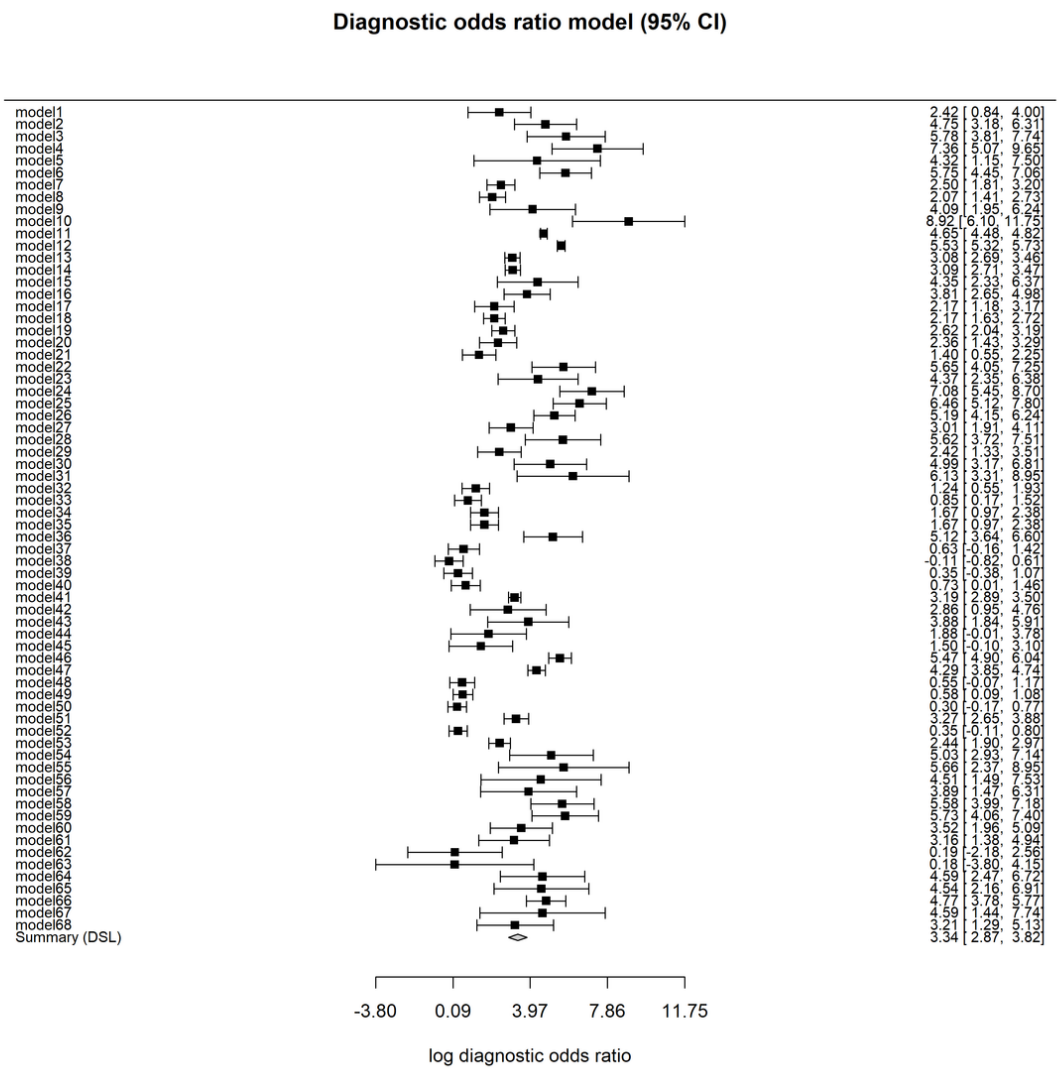


**Figure 3**

Diagnostic odds ratio of machine learning models in the study.

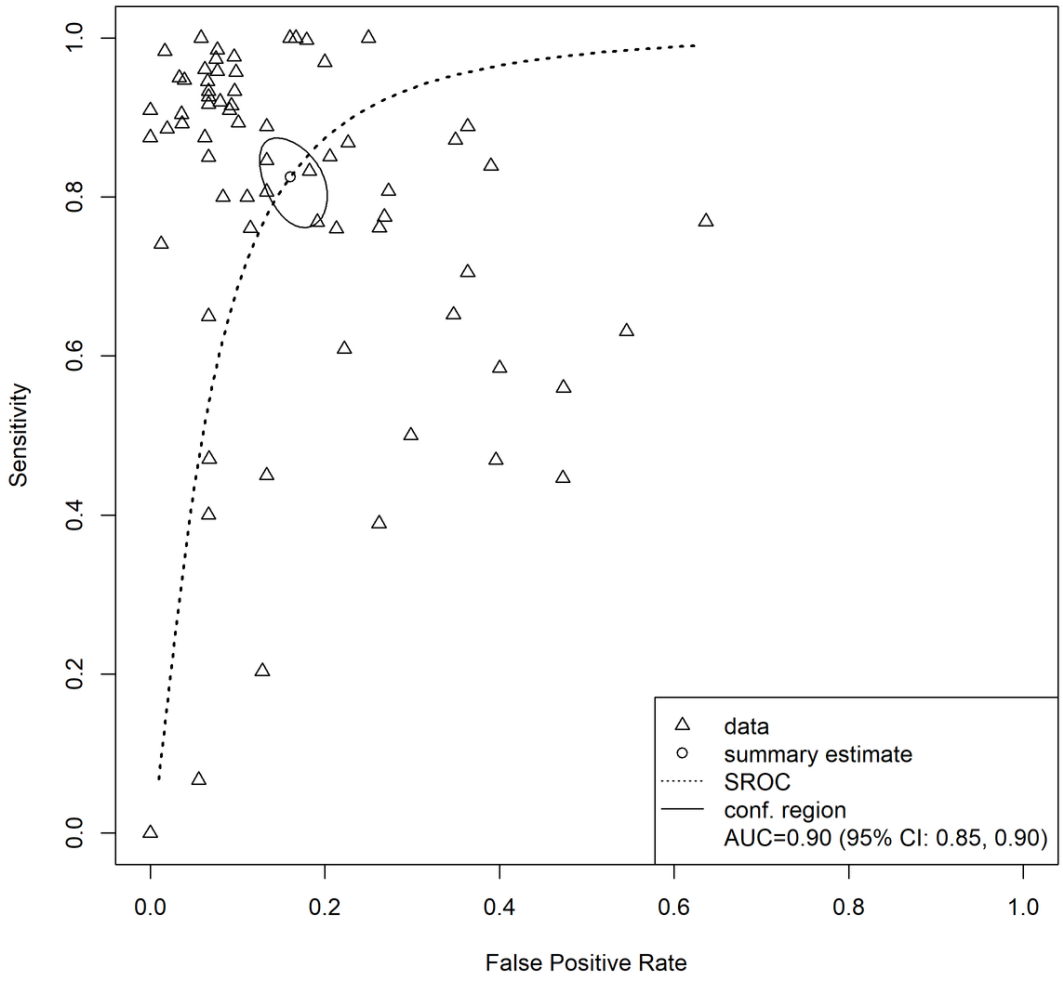


**Figure 4**

Hierarchical summary receiver operating characteristics (HSROC) curve for overall machine learning models in the study.

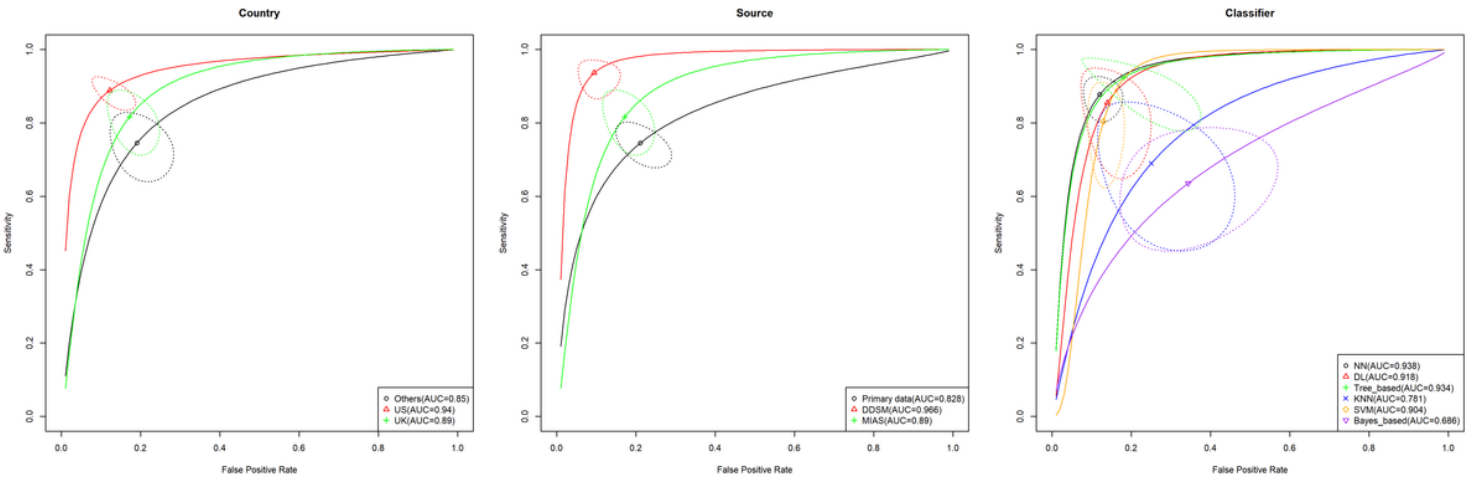


**Figure 5**

Hierarchical summary receiver operating characteristics (HSROC) curve for each subgroup analysis in the study.

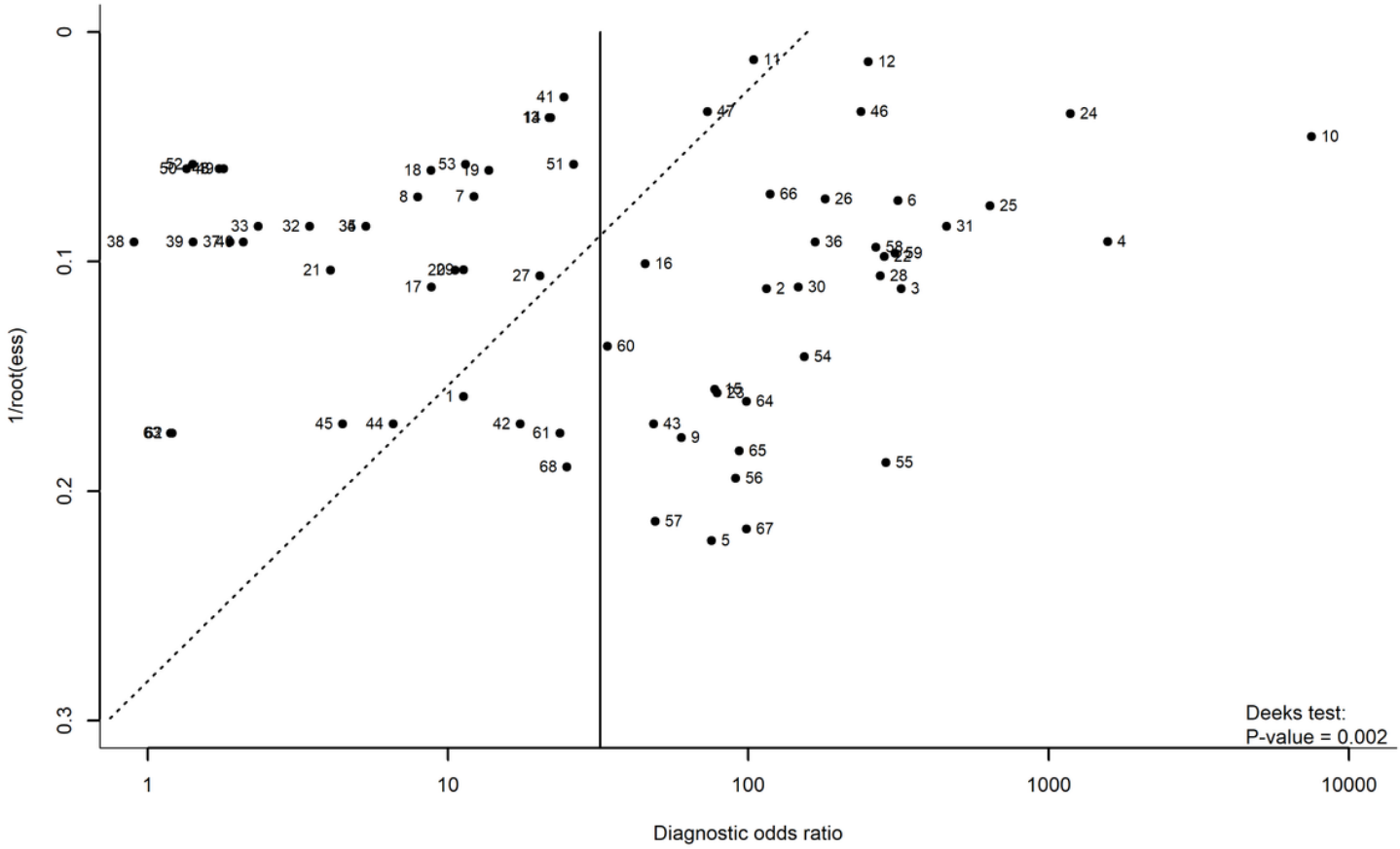


**Figure 6**

Deeks' funnel plot.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supp1influentialdiagnostic.docx
- Supp2.pdf
- Supp3Searchterms.docx