MAIBAI: Developing a metrological framework for assessment of image-based Artificial Intelligence systems for disease detection

Elizabeth A. Cooke[1,*], Spencer A. Thomas[1], Jessica E. Goldring[1], Alistair Mackenzie[2], Nadia A. S. Smith[2,3] and the MAIBAI consortium

[1]National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK
[2]Royal Surrey NHS Foundation Trust, Guildford, UK
[3]TÜV SÜD UK, Warrington, UK
***elizabeth.cooke@npl.co.uk**

A need for a metrological framework to assess AI models

As image-based artificial intelligence (AI) systems for disease detection become embedded in health systems, there is a growing need for rigorous, standardised methods to evaluate their performance, reliability, and clinical utility across heterogeneous real-world settings. Ensuring compliance with emerging regulatory and technical guidelines is essential for safe and trustworthy deployment.

Metrology, the science of measurement, is the foundation for the MAIBAI project [1] which is creating a comprehensive evaluation framework for diagnostic imaging AI, using breast cancer screening as an exemplar.

Infrastructure and data collection

We use a subset of data from the OPTIMAM Mammography Image Database (OMI-DB) which contains mammography images from breast screening centres in the UK, along with annotated cancer cases and clinical information [2].

A MAIBAI report outlines the characteristics of frameworks for collecting clinical images (e.g., Figure 1) and providing guidelines for their safe and effective use [3]. The report covers ethical considerations, information governance, and necessary infrastructure and data sharing agreements which must be considered as well as the image and clinical data collection.

Accurate data subcategorisation is vital for reliability and traceability in the training and validation of AI models

Clinical and technical features are likely to affect the appearance and interpretation of mammography images and, in turn, affect the output of AI software used to aid clinical decisions. For example, Figure 2 shows the changes in appearance for a breast phantom caused by different imaging systems. Figure 2 also shows the distribution of index of multiple deprivation (IMD) deciles of clients in the MAIBAI dataset. Non-uniform distributions such as this are important to consider when training data and grouping by demographics. Uniform random sampling of data may not be appropriate depending on the specific task.

Table 1 shows the distribution of ethnicities in the MAIBAI dataset. A significant proportion of clients do not have a recorded ethnicity. Lack of data recording may be an issue for an AI model as it is important to train on representative samples of subgroups in the data being classified.

Where possible, AI models should be trained and evaluated using data that includes subcategories based on key clinical and technical features, ensuring increased equitability in the data and coverage of image heterogeneities. Where this is not possible, the subcategories for which the AI model is valid should be clearly defined.

Generative modelling and uncertainty quantification

Generative models can be used for a variety of tasks such as supplementing AI training datasets as a solution to the lack of data in specific groups or reducing the variance between datasets such as between different scanners or populations. These approaches may produce anomalies in the generated data (i.e. hallucinations) requiring methods for assessing the confidence and trustworthiness of the model outputs before their use in downstream tasks.

We use uncertainty quantification to assess trustworthiness of generated data and flag anomalies. For instance, using style transfer to make an images from the Curated Breast Imaging Subset of the Digital Database for Screening Mammography [4] resemble the stylistic content (i.e. patient features) of the VinDr Mammo [5] dataset. Figure 3 shows an example of style transfer and the uncertainty in the resulting image. Hallucinations of structures were found to be more likely to occur in sparser regions of the image patches and these regions correspond with higher uncertainty from the model [6].

Evaluating an AI model using the MAIBAI data and framework

The New York University (NYU) has developed an AI model for breast cancer screening classification. The area under the receiver operating characteristic curves (AUC) for the NYU model was 0.83 for image-wise reading. The software is freely available for download and use [7].

We tested the NYU model without further training on the MAIBAI dataset (Figure 4). We found that there was a difference between the reported AUC of 0.83 for the model and the average found for all cases in this study of 0.74. We have also found that there is a difference in outcomes of the AI between screening centres and equipment manufacturers, in particular for GE systems, which the NYU model was not trained on. **This highlights the need for testing of AI before being deployed, and that there may be a need for re-training or fine-tuning of models.**

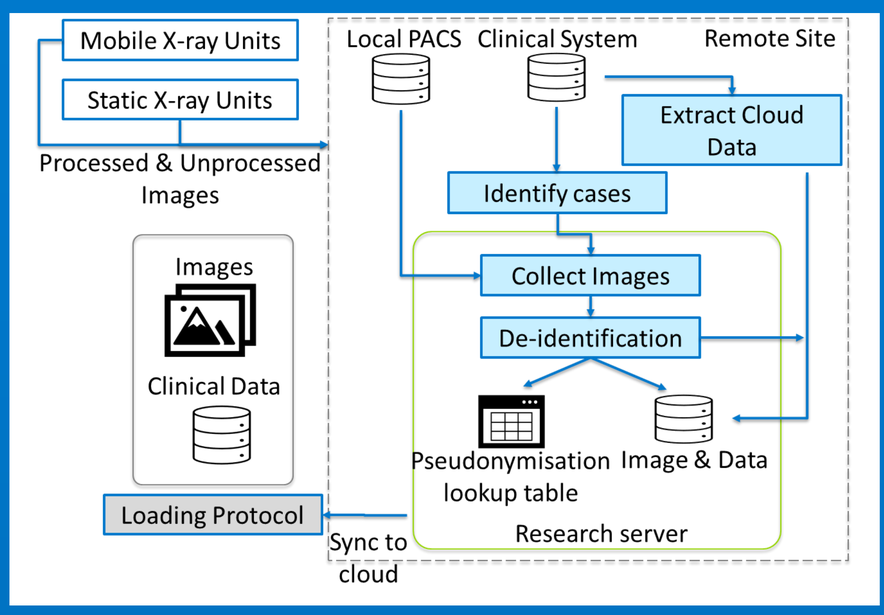

Figure 1: Example of the image collection process for mammography images in the UK Breast Screening Programme [2].

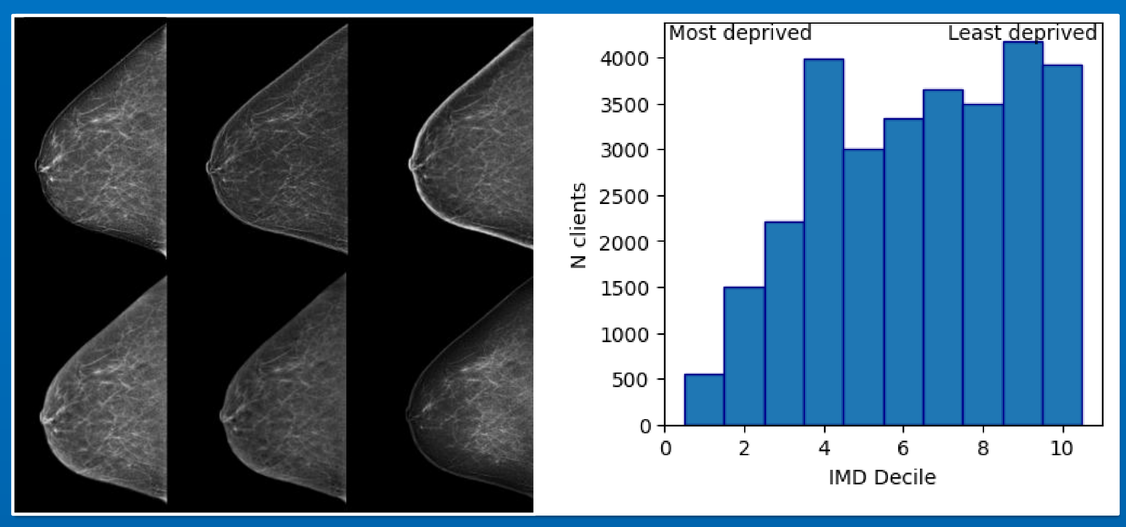

Figure 2 (above): Left: breast phantom acquired on six different scanners. Right: distribution of clients in the MAIBAI dataset across index of multiple deprivation (IMD) deciles. Total clients = 29,795, missing clients = 10,095.

Ethnic Category	%
White	54.6
Not recorded	34.4
Asian/Asian British	4.1
Black/Black British/Caribbean/African	2.5
Not stated	2.3
Other	1.3
Mixed/Multiple	< 1

Table 1 (left): Distribution of clients in the MAIBAI dataset across ethnic categories.

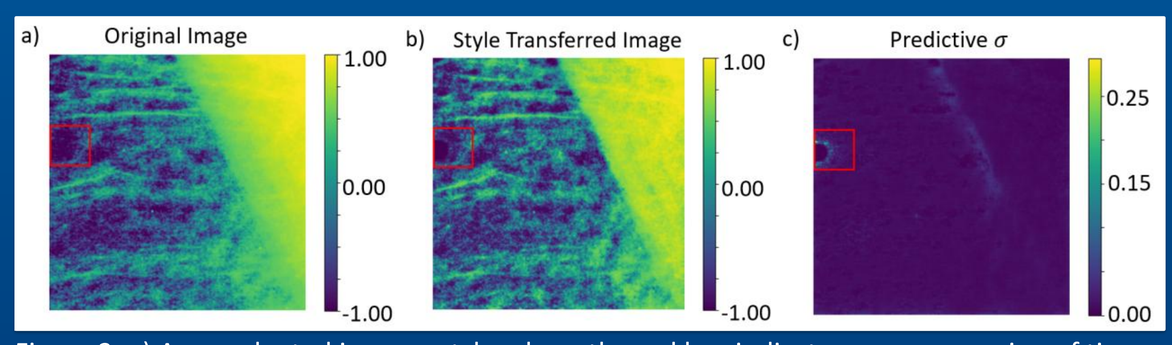

Figure 3: a) An unadapted image patch, where the red box indicates a sparser region of tissue. b) The style-transferred (cycleGAN) counterpart to a), that contains hallucinated features (tissue void) in this sparse region. c) the pixel-wise predictive standard deviation (σ). The hallucinated region has correspondingly high σ.

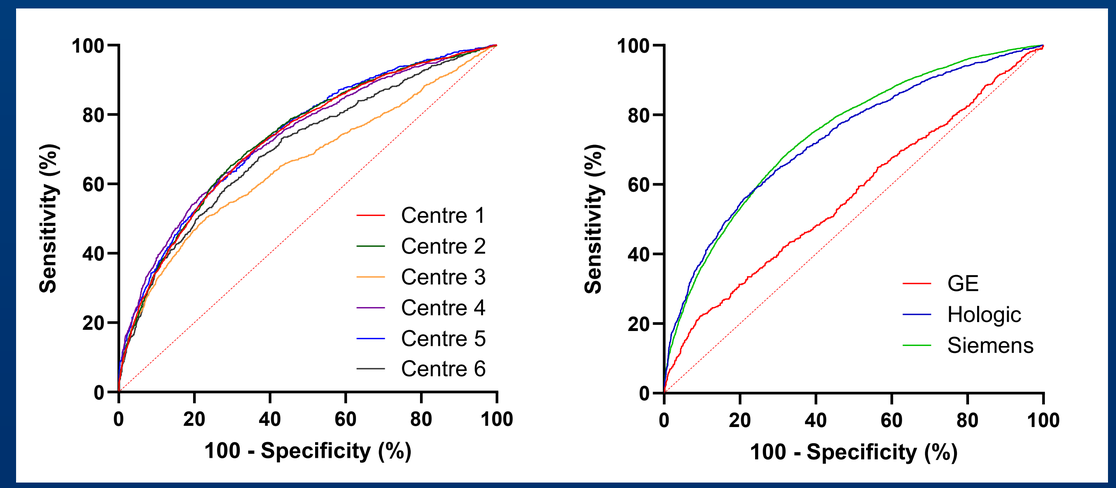

Figure 4: Receiver operating characteristic curves for (left) different screening centres and (right) different mammography equipment manufacturers.

[1] https://www.maibaiproject.eu/
[2] M. D. Halling-Brown et al., “OPTIMAM Mammography Image Database: A Large-Scale Resource of Mammography Images and Clinical Data,” Radiol. Artif. Intell., vol. 3, no. 1, p. e200103, Jan. 2021, doi: 10.1148/ryai.2020200103.
[3] A. Mackenzie et al., “Technical specification of a framework for the collection of clinical images and data”, arXiv:2508.03723
[4] R. Sawyer-Lee, et al. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm). 2016.
[5] Hieu T Nguyen, et al. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data, 10(1):277, 2023.
[6] C. Bench et al. “Trustworthy image-to-image translation: evaluating uncertainty calibration in unpaired training scenarios”, arxiv/2501.17570
[7] Wu, N. et al., “Deep Neural Networks Improve Radiologists’ Performance in Breast Cancer Screening,” IEEE Trans. Med. Imaging 39(4), 1184–1194 (2020).

Scan me to visit the MAIBAI project website

www.npl.co.uk

MCP-001, Issue 1, July 2025