MEDICAL IMAGING—ORIGINAL ARTICLE

Evaluation of deep learning-based artificial intelligence techniques for breast cancer detection on mammograms: Results from a retrospective study using a BreastScreen Victoria dataset

Helen ML Frazer,[1] Alex K Qin,[2] Hong Pan[2] and Peter Brotchie[3]

1 St Vincent's BreastScreen, St Vincent's Hospital Melbourne, Melbourne, Victoria, Australia
2 Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Victoria, Australia
3 Department of Medical Imaging, St Vincent's Hospital Melbourne, Melbourne, Victoria, Australia

HML Frazer MBBS, FRANZCR, MEpi; **AK Qin** PhD; **H Pan** PhD; **P Brotchie** MBBS, FRANZCR, PhD.

Correspondence
Adjunct A/Prof Helen M.L. Frazer, St Vincent's BreastScreen, St Vincent's Hospital Melbourne, 41 Victoria Parade, Melbourne, Vic. 3065, Australia.
Email: helen.frazer@svha.org.au

Conflict of interest: None.

Submitted 9 February 2021; accepted 15 June 2021.

doi:10.1111/1754-9485.13278

Abstract

Introduction: This study aims to evaluate deep learning (DL)-based artificial intelligence (AI) techniques for detecting the presence of breast cancer on a digital mammogram image.
Methods: We evaluated several DL-based AI techniques that employ different approaches and backbone DL models and tested the effect on performance of using different data-processing strategies on a set of digital mammographic images with annotations of pathologically proven breast cancer.
Results: Our evaluation uses the area under curve (AUC) and accuracy (ACC) for performance measurement. The best evaluation result, based on 349 test cases (930 test images), was an AUC of 0.8979 [95% confidence interval (CI) 0.873, 0.923] and ACC of 0.8178 [95% CI 0.785, 0.850]. This was achieved by an AI technique that utilises a certain family of DL models, namely ResNet, as its backbone, combines the global features extracted from the whole mammogram and the local features extracted from the automatically detected cancer and non-cancer local regions in the whole image, and leverages background cropping and text removal, contrast adjustment and more training data.
Conclusion: DL-based AI techniques have shown promising results in retrospective studies for many medical image analysis applications. Our study demonstrates a significant opportunity to boost the performance of such techniques applied to breast cancer detection by exploring different types of approaches, backbone DL models and data-processing strategies. The promising results we have obtained suggest further development of AI reading services could transform breast cancer screening in the future.

Key words: area under curve; artificial intelligence; breast cancer screening; deep learning; mammogram.

Introduction

Breast cancer is the most common cancer and the second most common cause of cancer-related death in Australian women. Approximately one in seven Australian women will be diagnosed with breast cancer in their lifetime.[1]

BreastScreen Australia is a public health programme, offering biennial mammographic screening targeted at women aged 50–74 years (available from age 40). The programme has successfully led to a 41–52% reduction in mortality for screening participants and a 21% reduction in population-level breast cancer mortality in Australia.[2] However, screening has several significant challenges:

- ACCURACY: Interpretation of mammograms by radiologists is subject to human variability and can result in detriment. Despite double reading of each mammogram by two independent radiologists (followed by a third arbitration read if disagreement), 34,001 Australian women (in target age range 50–74 years) in 2018 were recalled for assessment, and later determined not to have breast cancer. These unnecessary assessments create anxiety and may include digital breast tomosynthesis, further mammographic views, ultrasound, clinical examination and biopsy procedures. In addition, 3842 women (in age range 50–69 years screened in 2013–2015) were subsequently diagnosed with breast cancer within 2 years after receiving an 'all-clear' result.[3]

- PERSONALISATION: Screening is predominantly a 'one-size-fits-all' model and not tailored to risk.[4]
- SERVICE AND PARTICIPATION: Service delivery is slow at 14 days for an 'all-clear' mammogram screening result and 28 days for an assessment appointment.[3] Participation remains stubbornly flat at 55%.[3]
- COST: The current reading model is costly and labour-intensive, and screening demand is growing with an ageing population. Additionally, a workforce survey by the Royal Australian and New Zealand College of Radiologists (RANZCR) highlighted the potential undersupply of breast imaging radiologists to support the national breast screening programme in the future.[5]

AI has been held out as a potential solution to these challenges. Recently, AI techniques, particularly those based on DL models typically known as deep neural networks (DNN), have achieved promising results in many medical image analysis applications. Accordingly, there is a growing interest in using such techniques to assist in interpretation of mammograms, for example detecting the presence of breast cancer. Many relevant publications have been produced since June 2019.[6–15] Approaches can be categorised into three types: local, global and global + local as illustrated in Figure 1. Specifically, the local approach uses image patches which correspond to the local regions of cancer or non-cancer in mammographic images to train the DL model, applies the trained model to some patches sampled from a mammographic image to determine whether cancer is present in those patches and accordingly detects the presence of cancer in the whole mammogram. The global approach uses whole mammographic images, and their associated labels that indicate whether cancer is present in an image to train the DL model and applies the trained model to a mammographic image to determine whether cancer is present in that image. The global + local approach uses whole images to train a global DL model and image patches obtained from whole images (manually or automatically) to train a local DL model, and finally combines the global and local models to determine whether cancer is present in the whole image.

Comparatively, the local approach is less impacted by irrelevant image details in the whole image and can reveal the locations of cancers in the whole image. However, it often demands intensive human labour and domain expertise to designate cancer and non-cancer regions in the whole image and a computationally expensive post-processing step to convert the patch-level prediction into the image-level one. The global approach avoids the costly requirement for manually localising regions of interest in the whole image. However, it has to tolerate image details irrelevant to discrimination of cancers and cannot directly localise cancers in the whole image. The global + local approach addresses the drawbacks of the standalone local or global approach and unifies their strengths.

In this study, we examined all three AI approaches by using digital mammograms from the BreastScreen Victoria dataset, which are annotated by breast imaging radiologists with ground truth confirmed by surgical histopathology. Also, we studied the effects of applying different data processing strategies (i.e. enhancing data quality and increasing training data volume) on performance.

Method

We evaluated the local, global and global + local approaches, respectively. Particularly, the two global + local techniques[14,15] we examined mainly differ in how image patches are obtained from the whole image (i.e. manually or automatically) and how global features (from the whole image) and local features (from the image patch) are utilised. We also assessed data processing strategies in terms of data quality enhancement and data volume expansion. Our experiments were performed on Swinburne supercomputer OzSTAR* with a cluster of NVIDIA Tesla P100 GPUs.

Data description

Our study was based on a BreastScreen Victoria dataset with 28,694 digital mammographic images (six mammography machine vendors) from 7498 women with screen-detected breast cancer between January 2014 and December 2017. The mammograms from different vendors may have different photometric interpretation modes, that is monochrome 1 (the lowest pixel value is displayed white) and 2 (the lowest pixel value is displayed black). Mammographic images with breast cancer were annotated by BreastScreen radiologists with circles indicating the sites of biopsy-proven cancer. Typically, each breast for every woman has medio-lateral oblique (MLO) and cranio-caudal (CC) views. To facilitate DL model training and evaluation, all mammograms in the monochrome 1 mode were converted to the monochrome 2 mode by applying a simple linear grayscale inversion. Also, all mammograms were converted from the original DICOM format to the 8-bit PNG format.

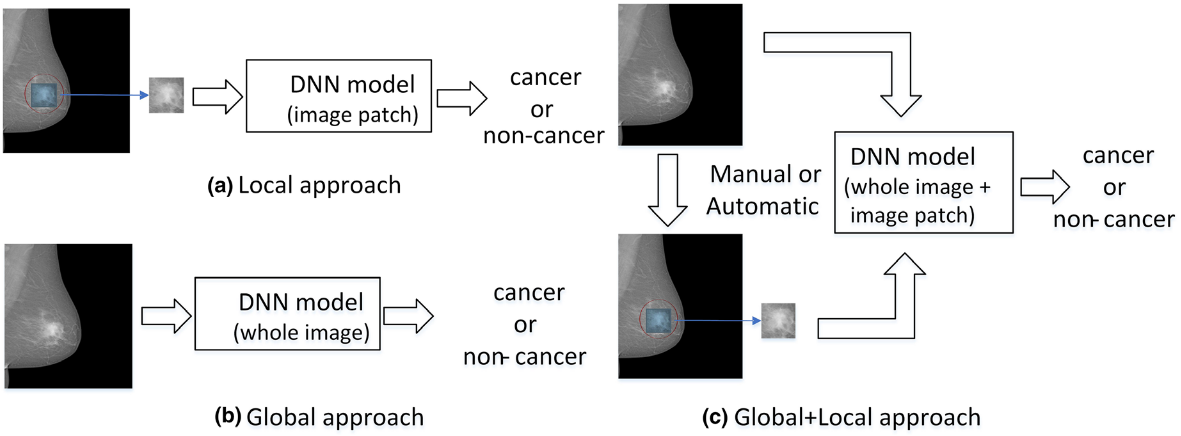


Fig. 1. Illustration of the local, global and global + local approaches.

Assessing AI approaches

Pre-processing data

To assess AI approaches, one set of data is required for training the model and another set of data for testing the trained model. We selected, from the BreastScreen Victoria dataset, 8049 mammograms (3919 with cancers and 4130 without cancers) as the training set and 930 images (434 with cancer and 496 without cancers) as the test set. We only selected the images from those mammography machines that are currently in operational use. Further, our selection of mammograms with cancer was based on the requirements of no breast implants or pacemakers, no annotation conflict from two or more radiologists and only one cancer site on the mammogram. In addition, we generated 8949 image patches with and without cancers from some selected mammograms and established 8375 (3988 with cancer and 4387 without cancer) as the training set and 574 (379 with cancer and 195 without cancer) as the test set for assessing the local approach.

When generating image patches without cancer, we randomly selected some mammograms without annotations and sampled the patches from them, where the selected mammograms correspond to the 'normal' mammograms from the same women with histologically proven cancer in the opposite breast. Further, the image patches not within the density range of the patches containing cancer were excluded to avoid the patches not covering the breast tissue. Then, all image patches were visually reviewed to ensure they are all centred on the breast tissue.

We made a 90-10 split of the training set for training and validation and used this setting to perform model selection and hyper-parameter tuning. We applied data pre-processing prior to model training, where the data used for building and evaluating the model were resized, normalised and augmented. Different techniques (to be described below) employed different data pre-processing strategies to seek better generalisation performance. More details about data pre-processing can be referred to in the Supporting Information.

Examining the local approach

We implemented the local approach by using ResNet50[16] as a backbone DL model, pre-trained on the ImageNet[17] dataset (the most widely used large-scale public dataset of natural images) and re-trained on 8375 image patches with cancer and non-cancer labels. The trained model was applied to all image patches centred around each pixel in a mammogram to produce a saliency map that reflects the probability of cancer occurrence at each site on that mammogram where the value of each pixel in the saliency map is the output (in the probability form) of the trained model. The saliency map can be directly provided to radiologists or further processed, for example thresholding the saliency map at a suitable level to convert it to a binary mask which outlines tumour regions on the mammogram.

Examining the global approach

Three DNNs (i.e. Inception-ResNet-V2,[18] EfficientNetB6[19] and NASNetLarge[20]) which have shown state-of-the-art performance in computer vision applications were implemented as the backbone DL models in the global approach. We initialised these models with their pre-training models on ImageNet and re-trained them using the pre-specified hyper-parameters on our training data that were resized to 331 × 331 pixels for NASNetLarge and 768 × 768 pixels for the other two DNNs, respectively. To seek the best classification performance, we applied hyper-parameter tuning to these models with details provided in the Supporting Information.

Examining the global + local approach

We implemented and evaluated two global + local methods[14,15] for breast cancer classification. Specifically, global+local method 1 (GL1)[14] relies on the strongly supervised localisation of abnormal and normal regions for building the local model (i.e. a patch classifier) on a large set of manually obtained image patches, whereas global+local method 2 (GL2)[15] automatically obtains image patches from the saliency map generated by a global model and uses them to train the local model (and thus avoid manual effort) in a weakly supervised manner.

The GL1 follows a two-stage training pipeline. First, a patch classifier is trained on a set of manually obtained image patches with cancer and non-cancer labels. Second, the parameters of the trained patch classifier are used to initialize those of the whole image classifier which is then trained on a set of whole images with image-level annotations. The implementation details of GL1 can be referred to in Shen *et al*.,[14] where the Digital Database for Screening Mammography (DDSM)[21] dataset was used to train the patch and whole image classifiers and the following pre-trained models on DDSM are provided:

- VGG16: VGG16[22] model used as both patch and whole image classifiers,
- ResNet50: ResNet50 model used as both patch and whole image classifiers and
- VGG16_ResNet50_hybrid: VGG16 model used for the patch classifier and ResNet50 model used for the whole image classifier.

We directly applied the pre-trained patch classifier[†] on DDSM to initialise the whole image classifier and then re-trained it using the default data pre-processing operations and hyper-parameter settings suggested in Shen *et al*.[14] on our training images resized to 1152 × 896 to enable the use of the pre-trained model.

The GL2 has three modules, that is global, local and fusion modules. First, the global module is trained on the whole image to generate a saliency map that reflects the coarse locations of malignant lesions. Next, several highly possible malignant regions revealed by the saliency map are extracted automatically as image patches which are then used to train the local module. Finally, the fusion module aggregates the global information (generated by the global module) and the local details (generated by the local module) to make a prediction on the existence of malignant lesions in a mammogram. The details of GL2 can be referred to in Shen *et al*.[15]

Several models pre-trained on the NYU breast cancer screening dataset[23] by using different hyper-parameter settings are available. We explored three of such pre-trained models[‡] with ResNet22 and ResNet18[16] used in global and local modules, corresponding to three different hyper-parameter settings, that is GL2-ResNet22/18-Setting 1, GL2-ResNet22/18-Setting 2 and GL2-ResNet22/18-Setting 3. To re-train them on our own training data, we resized images to 2944 × 1920 (to enable the use of the pre-trained model), employed the pre-trained model to initialise the model to be re-trained, and used the default data pre-processing operations and hyper-parameter settings suggested in Shen *et al*.[15]

Assessing data processing strategies

Data quality enhancement

We assessed background cropping with text removal (BCTR) and contrast adjustment (CA) for enhancing data quality. BCTR aims to improve image quality by removing text and background areas which do not contain breast regions by using image segmentation techniques.[24] CA aims to improve image quality by transforming the intensity scales (i.e. the range of intensity values) of the breast regions of all mammograms into a uniform one (determined by selected mammograms that contain breast regions with visually good appearance) and thus the same contrast. Notably, CA is not used to enhance the visual quality of mammograms to improve readability for radiologists but used as a data normalisation operation for training the DL model. In this study, we use CA to reduce the variations of intensity scales across the breast regions of different mammograms to support the AI model learning tissue structure properties alongside intensity properties. More details about BCTR and CA can be referred to in the Supporting Information.

Increasing the volume of training data

We studied the effect of increasing the training data volume on performance. Specifically, we selected additional previously unused mammographic images with and without cancers from the original dataset and added them into the original training set to form an expanded training set of 12,531 (cancer: 5888 and non-cancer: 6643) mammograms. We maintained the same test set for a fair comparison.

Results

Evaluation of AI approaches

Our implemented local approach was trained for 100 epochs. The trained model was tested on 574 image patches, producing promising patch classification results with AUC of 0.9765 [95% CI 0.9770, 0.9761] and ACC of 0.9442 [95% CI 0.9446, 0.9437], where a bootstrapping method was used to obtain the 95% confidence interval (CI) for AUC and ACC with details in the Supporting Information. The saliency maps generated by applying the trained model to some image patches are illustrated in Figure 2, reflecting the probability of cancer occurrence at each site within the patch.

Table 1 reports the classification results in terms of AUC and ACC for the global approach implemented with three different DL models (including Inception-ResNet-V2, EfficientNetB6 and NASNetLarge) and the two global + local approaches GL1 and GL2 implemented with different DL models (GL1) or the same DL model under different hyper-parameter settings (GL2). It can be observed that

- With the global approach, the best AUC and ACC results are achieved by Inception-ResNet-V2. After performing hyper-parameter tuning, the performances of Inception-ResNet-V2, EfficientNetB6 and NASNetLarge are all improved. Due to space limit, the saliency maps generated by applying the re-trained models with hyper-parameter tuning to some mammograms are provided in the Supporting Information.
- With the global + local approach, directly applying pre-trained models obtained from DDSM and NYU datasets to our data leads to reduced performance. We think that the poor performance is due to the discrepancy in data properties (e.g. intensity scales) between DDSM and NYU datasets and ours. This is verified by the fact that re-training pre-trained models on our data significantly improves the performance as shown in Table 1. For GL1, GL1-VGG16 achieves the best AUC and ACC results. For GL2, GL2-ResNet22/18-Setting1 outperforms the other models in terms of the highest AUC and the competitive ACC.
- The global + local approach GL2 which automatically generates saliency maps and extracts suspicious image patches around malignant regions from the mammogram, demonstrates the most promising results. Example saliency maps and image patches generated by GL2 are provided in the Supporting Information.

Figure 3 plots the receiver operating characteristic (ROC) curves of the best performing models for the global approach (i.e. Inception-ResNet-V2, EfficientNetB6 and NASNetLarge) and the global+local approaches GL1 and GL2 (i.e. GL1-VGG16 and GL2-ResNet22/18-Setting1).

Evaluation of data quality enhancement

We assessed the effect of data quality enhancement in terms of BCTR and CA by comparing the performances of the five best performing models in Figure 3 with and without applying these two operations. The results are reported in Table 2.

It can be observed that applying BCTR leads to performance improvement for GL1-VGG16 and GL2-ResNet22/18-Setting1 but degradation for Inception-ResNet-V2, EfficientNetB6 and NASNetLarge. In fact, BCTR results in smaller images with changed sizes which, after being resized to a square shape required by the three DL models (to enable using the pre-trained model), causes ratio aspect distortion and thus performance degradation. This is not the case for GL1 and GL2 approaches because their DL models do not require square-shaped input images. Applying CA consistently improves performance across all tested cases because CA reduces the discrepancy of the intensity scales of breast regions. GL2-ResNet22/18-Setting1 with BCTR and CA achieves the best performance among all tested cases.

Evaluation of increasing training data volume

To evaluate the effect of increasing the volume of training data on performance, we used an expanded training set and repeated the experiment described in the previous section. The results are reported in Table 3. It can be observed that using additional training data leads to consistent performance improvement (over those reported

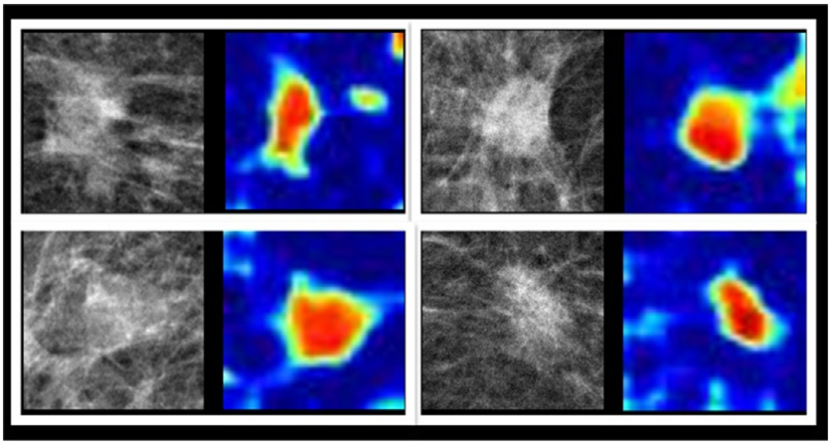
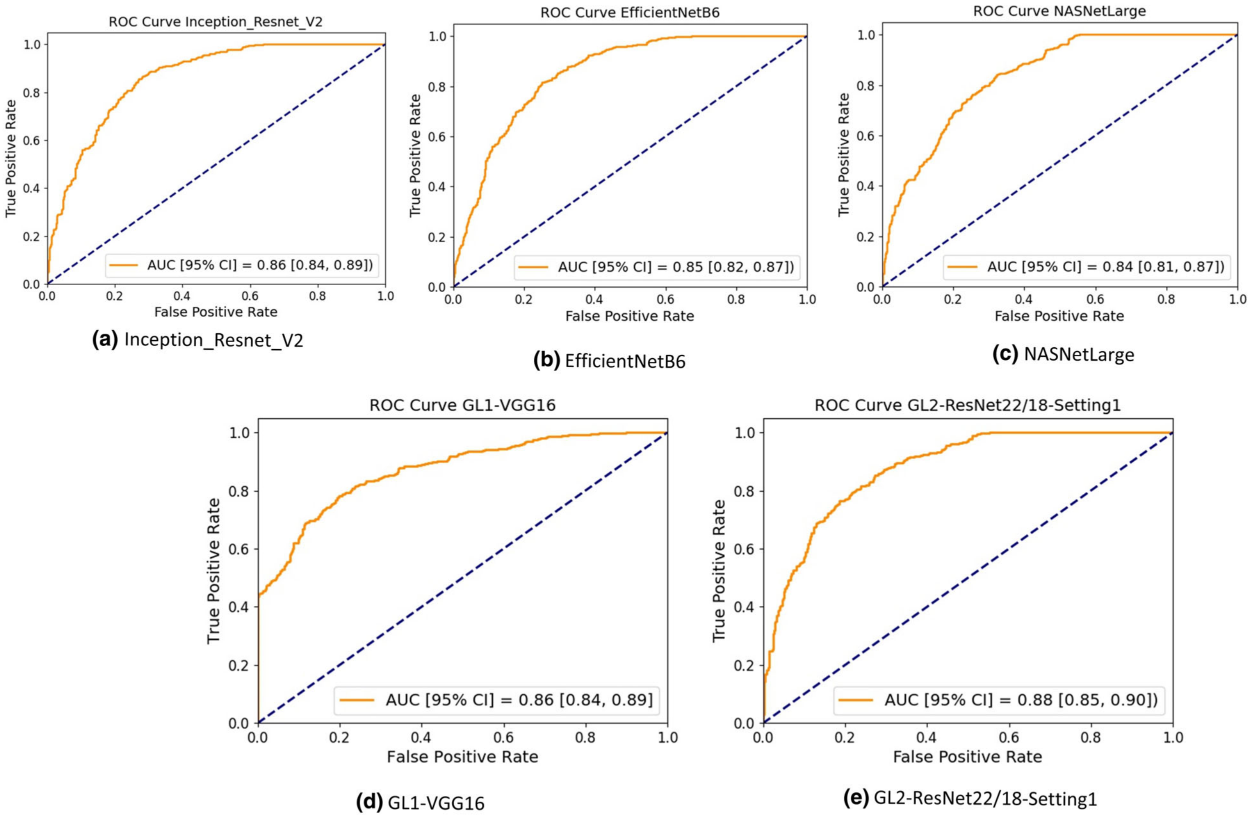


Fig. 2. Saliency maps (second and fourth columns) generated by the local approach applied to example image patches (first and third columns).

Table 1. Comparison of the classification performances of using the re-trained models with (w/) and without (w/o) hyper-parameter tuning in the global approach and using pre-trained and re-trained models in two global + local approaches, where the best AUC and ACC in the global approach and two global + local approaches are highlighted in bold

Models in the global approach	Re-trained model w/o hyper-parameter tuning		Re-trained model w/ hyper-parameter tuning	
	AUC [95% CI]	ACC [95% CI]	AUC [95% CI]	ACC [95% CI]
Inception-ResNet-V2	0.8398 [0.812, 0.869]	0.7333 [0.696, 0.752]	**0.8625** [0.836, 0.890]	**0.7572** [0.725, 0.790]
EfficientNetB6	0.8394 [0.808, 0.869]	0.7387 [0.713, 0.770]	0.8459 [0.818, 0.874]	0.7540 [0.719, 0.786]
NASNetLarge	0.8264 [0.794, 0.858]	0.7430 [0.717, 0.775]	0.8393 [0.807, 0.868]	0.7479 [0.717, 0.784]
Models in two global + local approaches	Pre-trained model		Re-trained model	
	AUC [95% CI]	ACC [95% CI]	AUC [95% CI]	ACC [95% CI]
Global + local approach 1 (GL1)				
GL1-ResNet50	0.6236 [0.584, 0.659]	0.5677 [0.539, 0.588]	0.7779 [0.744, 0.817]	0.6871 [0.653, 0.725]
GL1-VGG16	0.6087 [0.564 - 0.650]	0.5613 [0.522 - 0.598]	**0.8636** [0.837, 0.891]	**0.7580** [0.729, 0.793]
GL1-VGG16_ResNet50_hybrid	0.6273 [0.585, 0.668]	0.5845 [0.579, 0.656]	0.8009 [0.766, 0.834]	0.6935 [0.654, 0.728]
Global + local approach 2 (GL2)				
GL2-ResNet22/18-Setting1	0.5638 [0.520, 0.609]	0.5452 [0.507 - 0.584]	**0.8758** [0.850, 0.900]	0.7790 [0.746, 0.811]
GL2-ResNet22/18-Setting2	0.6128 [0.572, 0.655]	0.5634 [0.527, 0.602]	0.8719 [0.843, 0.897]	**0.7806** [0.747, 0.811]
GL2-ResNet22/18-Setting3	0.5774 [0.534, 0.621]	0.5398 [0.502, 0.579]	0.841 [0.812, 0.870]	0.745 [0.712, 0.774]



(a) Inception_Resnet_V2

(b) EfficientNetB6

(c) NASNetLarge

(d) GL1-VGG16

(e) GL2-ResNet22/18-Setting1

Fig. 3. Receiver Operating Characteristic (ROC) curves of the best models for Inception-ResNet-V2, EfficientNetB6, NASNetLarge, GL1, and GL2.

in Table 2) across all tested cases. Notably, the overall best performance with AUC of 0.8979, 95% CI [0.873, 0.923] and ACC of 0.8178, 95% CI [0.785, 0.850], is achieved by GL2-ResNet22/18-Setting1.

Discussion

Medical images, including mammograms, are unique with high resolution and often very small, subtle regions of

Table 2. Comparison of the classification performances of five DL-based AI techniques with (w/) and without (w/o) data quality enhancement, where the overall best AUC and ACC are highlighted in bold

DL-based AI techniques	w/o BCTR		w/ BCTR	
	AUC [95% CI]	ACC [95% CI]	AUC [95% CI]	ACC [95% CI]
Inception-ResNet-V2				
w/o CA	0.8625 [0.836, 0.890]	0.7572 [0.725, 0.790]	0.8496 [0.819, 0.878]	0.7543 [0.721, 0.787]
w/ CA	0.8694 [0.841, 0.894]	0.7726 [0.743, 0.804]	0.8650 [0.838, 0.891]	0.7664 [0.734, 0.797]
EfficientNetB6				
w/o CA	0.8459 [0.818, 0.874]	0.7540 [0.719, 0.786]	0.8427 [0.812, 0.871]	0.7389 [0.704, 0.773]
w/ CA	0.8521 [0.821, 0.880]	0.7638 [0.732, 0.798]	0.8467 [0.814, 0.878]	0.7498 [0.714, 0.784]
NASNetLarge				
w/o CA	0.8393 [0.807, 0.868]	0.7479 [0.717, 0.784]	0.8366 [0.803, 0.864]	0.7380 [0.702, 0.770]
w/ CA	0.8481 [0.816, 0.876]	0.7557 [0.723, 0.789]	0.8435 [0.811, 0.870]	0.7483 [0.713, 0.781]
GL1-VGG16				
w/o CA	0.8636 [0.837, 0.891]	0.7580 [0.729, 0.793]	0.8714 [0.846, 0.899]	0.7667 [0.734, 0.799]
w/ CA	0.8745 [0.850, 0.900]	0.7785 [0.747, 0.810]	0.8757 [0.852, 0.903]	0.7796 [0.748, 0.811]
GL2-ResNet22/18-Setting1				
w/o CA	0.8758 [0.850, 0.900]	0.7790 [0.746, 0.811]	0.8783 [0.851, 0.903]	0.7806 [0.747, 0.814]
w/ CA	0.8810 [0.855, 0.905]	0.7946 [0.765, 0.826]	**0.8877** [0.865, 0.911]	**0.8011** [0.770, 0.833]

Table 3. Comparison of the classification performances of the five DL-based AI techniques trained on the expanded training set with (w/) and without (w/o) data quality enhancement, where the overall best AUC and ACC are highlighted in bold

DL-based AI techniques	w/o BCTR		w/ BCTR	
	AUC [95% CI]	ACC [95% CI]	AUC [95% CI]	ACC [95% CI]
Inception-ResNet-V2				
w/o CA	0.8670 [0.839, 0.892]	0.7823 [0.746, 0.808]	0.8598 [0.834, 0.885]	0.7759 [0.744, 0.805]
w/ CA	0.8767 [0.850, 0.905]	0.7838 [0.749, 0.819]	0.8666 [0.846, 0.900]	0.7780 [0.747, 0.814]
EfficientNetB6				
w/o CA	0.8611 [0.832, 0.888]	0.7605 [0.730, 0.798]	0.8553 [0.818, 0.883]	0.7465 [0.714, 0.785]
w/ CA	0.8636 [0.833, 0.893]	0.7684 [0.735, 0.800]	0.8587 [0.830, 0.889]	0.7568 [0.722, 0.790]
NASNetLarge				
w/o CA	0.8549 [0.827, 0.882]	0.7564 [0.724, 0.788]	0.8431 [0.812, 0.871]	0.7463 [0.711, 0.780]
w/ CA	0.8599 [0.830, 0.888]	0.7615 [0.728, 0.795]	0.8522 [0.823, 0.880]	0.7553 [0.721, 0.788]
GL1-VGG16				
w/o CA	0.8698 [0.845, 0.897]	0.7774 [0.749, 0.813]	0.8720 [0.847, 0.901]	0.7783 [0.750, 0.815]
w/ CA	0.8773 [0.858, 0.910]	0.7877 [0.752, 0.822]	0.8787 [0.852, 0.905]	0.7888 [0.754, 0.830]
GL2-ResNet22/18-Setting1				
w/o CA	0.8817 [0.857, 0.906]	0.7860 [0.756, 0.817]	0.8872 [0.861, 0.910]	0.7946 [0.765, 0.823]
w/ CA	0.8878 [0.861, 0.912]	0.7968 [0.767, 0.830]	**0.8979** [0.873, 0.923]	**0.8178** [0.785, 0.850]

interest. In breast screening programmes, radiologists need to use their visual perception and acuity to examine a whole image and explore smaller regions of interest to conclude whether they view a mammogram as suspicious for breast cancer. Significant variation can be observed in the reporting of images such that the current BreastScreen Australia standard is double reading of every mammogram followed by a third arbitration read if needed.[3,25]

A range of international studies[6–15] has applied AI techniques to assist in reading mammograms and seeking to address the unique challenges arising therein. Our study employed a high-quality, multi-vendor dataset of digital mammograms with cancer being annotated (and confirmed on surgical histopathology) and explored a range of DL-based AI techniques with different types of approaches and backbone DL models as well as different data processing strategies. Specifically, we focused on global and global + local approaches and studied the effects of enhancing the data quality and increasing the training data volume on performance. The global + local approach with ResNet22 and ResNet18 used as its global and local DL modules, respectively, when BCTR and CA are applied and the training data volume is increased, produces the overall best result with AUC of 0.8979, [95% CI 0.873, 0.923] and ACC of 0.8178 [95% CI 0.785, 0.850].

There are challenges with drawing conclusions from comparisons of AUC and ACC values unless the same data are used for training and testing. However, these results are similar to those reported in recently published studies on screening data, for example McKinney *et al*.[8] reported the best AUC of 0.889 on a UK screening dataset. It shows the potential to replace the second reader with an AI reader in screening programmes (non-inferior) and the use of an AI reader in single reader diagnostic services like in the USA (superior).[8]

Whilst the potential to see such AI services become part of screening programmes in the future is clear, significant technical and clinical challenges still need to be overcome. Furthermore, the ethical, legal and social implications of introducing AI into healthcare systems need to be further developed.

Technical

There remains opportunity for future work to improve performance. Some key opportunities are described below:

- The DL models used in either the global or the local approaches are large-scale, leading to expensive computational cost and preventing use of a large batch size for training (due to memory limit) to enable DL models to converge more quickly than using a small batch size. Therefore, it is highly desirable to study the effect of using lightweight DL models.
- For GL2, we re-trained the pre-trained DL models based on NYU data by using our own training data, where CA was applied to our training data to reduce its difference from NYU data in terms of image properties. More data enhancement strategies including better CA methods need further exploration to make the best use of pre-trained models and thus boost performance.
- In the current implementation of GL1 and GL2, the default hyper-parameter settings as suggested in Shen *et al*.[14] and Shen *et al*.[15] were used. Tuning hyper-parameters in GL1 and GL2, given the availability of sufficient computing resources, may lead to performance improvement.

Clinical

The current evaluation of AI techniques needs to shift from simply understanding AUC and ACC performance to clinically meaningful outcomes in a population screening programme. A scoping review of AI for early detection of breast cancer[26] highlighted few studies that report comparative estimates for AI and radiologists. Key focus areas for our future work included as follows:

- Demonstrating comparative performance in low cancer prevalent 'real-world' retrospective, prospective feasibility studies and randomised controlled trials,
- Reducing unnecessary recalls and interval cancers,
- Distinguishing prognostically significant breast cancers,
- Understanding the performance of AI techniques in various risk cohorts (covariate-adjusted AUC) such as screening round, age, density and family history alongside opportunities for risk prediction enabling personalisation of screening pathways,
- Explainability and reporting of AI reading results,
- Generalisability in other screening jurisdictions in Australia and globally.

Ethical, legal and social

Finally, the ethical, legal and social implications need to be examined and addressed for successful translation into the healthcare setting.[27] Here, our future work is focused on

- Quality management of AI techniques to prevent bias and drift, and to ensure stated clinical performance achieved by them,
- Clinician and client acceptance, experience and workflow,
- Legal obligations.

In conclusion, using AI in screening programmes has major transformative possibilities. Our study revealed the potential of applying DL-based AI techniques to mammogram reading. The challenge is now shifting towards testing prospectively, demonstrating clinically meaningful outcomes and addressing ethical, legal and social implications. This is now the focus of our 'Transforming Breast Cancer Screening with Artificial Intelligence' or 'BRAIx' research programme.

Acknowledgements

The authors would like to thank St Vincent's BreastScreen, St Vincent's Hospital Melbourne, BreastScreen Victoria, Professor Gustavo Carneiro (Director of Medical Machine Learning at the Australian Institute of Machine Learning, University of Adelaide) and Katrina Kunicki (AI Innovation and Project Manager, BRAIx program, St Vincent's Institute of Medical Research) for their encouragement and assistance. The study was funded by the Medical Device Partnering Program (MDPP).

Notes

*https://supercomputing.swin.edu.au/

[†] We used the open source code and pre-trained patch classifiers from https://github.com/lishen/end2end-all-conv to re-train the GL1 model on our data.

[‡] We used the open source code and pre-trained models from https://github.com/nyukat/GMIC to re-train the GL2 model on our data.

References

1. Australian Institute of Health and Welfare 2019. Cancer in Australia 2019. Cancer series no. 119. Cat. no. CAN 123. Canberra: AIHW.
2. Nickson C, Velentzis LS, Brennan P, Mann GB, Houssami N. Improving breast cancer screening in Australia: a public health perspective. *Health Res Pract* 2019; **29**: e2921911.
3. Australian Institute of Health and Welfare 2020. BreastScreen Australia monitoring report 2020. Cancer series no. 129. Cat. no. CAN 135. Canberra: AIHW.
4. Cancer Council Australia. Optimising early detection of breast cancer in Australia. Sydney: Cancer Council Australia. [Cited 30 May 2021.] Available from URL: https://www.cancer.org.au/about-us/policy-and-advocacy/early-detection-policy/breast-cancer-screening/optimising-early-detection
5. 2016 Workforce Survey Report Australia Faculty of Clinical Radiology Royal Australian and New Zealand College of Radiologists, 2018.
6. Akselrod-Ballin A, Chorev M, Shoshan Y *et al*. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* 2019; **292**: 331–42.
7. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019; **292**: 60–6.
8. McKinney SM, Sieniek M, Godbole V *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577**: 89–94.
9. Kim H-E, Kim HH, Han B-K *et al*. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancer Digit Health* 2020; **2**: e138–48.
10. Schaffter T, Buist D, Lee C *et al*. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. *JAMA Netw Open* 2020; **3**: e200265.
11. Rodriguez-Ruiz A, Lang K, Gubern-Merida A *et al*. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019; **111**: 916–22.
12. Lotter W, Diab AR, Haslam B *et al*. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. *Nat Med* 2021; **27**: 244–9.
13. Wu N, Phang J, Park J *et al*. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging* 2020; **39**: 1184–94.
14. Shen L, Margolies L, Rothstein J, Fluder E, McBride R, Sieh W. Deep learning to improve breast cancer detection on screening mammography. *Sci Rep* 2019; **9**: 12495.
15. Shen Y, Wu N, Phang J *et al*. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Med Image Anal* 2021; **68**: 1–17.
16. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016; pp. 770–8.
17. Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009; pp. 248–55.
18. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017; pp. 4278–84.
19. Tan M, Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: *International Conference on Machine Learning (ICML)*, 2019; pp. 6105–14.
20. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
21. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data* 2017; **4**: 170177.
22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
23. Wu N, Phang J, Park J *et al*. The NYU Breast Cancer Screening Dataset v1.0. Technical Report, 2019.
24. Gonzalez RC, Woods RE. *Digital Image Processing*, 4th edn. New York, NY, USA: Pearson, 2018.
25. Brennan PC, Ganesan A, Eckstein MP, Ekpo E. Benefits of independent double reading in digital mammography. *Acad Radiol* 2019; **26**: 717–23.
26. Houssami N, Kirkpatrick-Jones G, Noguchi N, Lee CI. Artificial Intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices* 2019; **16**: 351–62.
27. Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. *Breast* 2020; **49**: 25–32.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Supplementary material from the evaluation of deep learning models for detection of breast cancer on mammography.