Independent evaluation of a multi-view multi-task convolutional neural network breast cancer classification model using Finnish mammography screening data

Isosalo, A.; Inkinen, S.I.; Turunen, T.; Ipatti, P.S.; Reponen, J. ...

2023-07

Elsevier Ltd.

http://hdl.handle.net/10138/575396

Isosalo, A, Inkinen, S I, Turunen, T, Ipatti, P S, Reponen, J & Nieminen, M T 2023, 'Independent evaluation of a multi-view multi-task convolutional neural network breast cancer classification model using Finnish mammography screening data', Computers in Biology and Medicine, vol. 161, 107023. https://doi.org/10.1016/j.compbiomed.2023.107023

Journal Pre-proof

Independent evaluation of a multi-view multi-task convolutional neural network breast cancer classification model using Finnish mammography screening data

A. Isosalo, S.I. Inkinen, T. Turunen, P.S. Ipatti, J. Reponen, M.T. Nieminen

PII: S0010-4825(23)00488-2
DOI: https://doi.org/10.1016/j.compbiomed.2023.107023
Reference: CBM 107023

To appear in: *Computers in Biology and Medicine*

Received date : 31 December 2022
Revised date : 30 April 2023
Accepted date : 9 May 2023

Please cite this article as: A. Isosalo, S.I. Inkinen, T. Turunen et al., Independent evaluation of a multi-view multi-task convolutional neural network breast cancer classification model using Finnish mammography screening data, *Computers in Biology and Medicine* (2023), doi: https://doi.org/10.1016/j.compbiomed.2023.107023.

Independent evaluation of a multi-view multi-task convolutional neural network breast cancer classification model using Finnish mammography screening data⋆,⋆⋆,⋆⋆⋆

A. Isosalo[a,*], MSc, S.I. Inkinen[a,d], PhD, T. Turunen[b], MD, P.S. Ipatti[b], MD, J. Reponen[a,c], MD, PhD and M.T. Nieminen[a,b,c], PhD

[a]*Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland*
[b]*Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland*
[c]*Medical Research Centre Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland*
[d]*HUS Diagnostic Center, Clinical Physiology and Nuclear Medicine, Helsinki University and Helsinki University Hospital, Helsinki, Finland*

ARTICLE INFO

Keywords:
breast radiology
computer vision
screening
mammography
classification
DICOM

ABSTRACT

Background: Development of deep convolutional neural networks for breast cancer classification has taken significant steps towards clinical adoption. It is though unclear how the models perform for unseen data, and what is required to adapt them to different demographic populations. In this retrospective study, we adopt an openly available pre-trained mammography breast cancer multi-view classification model and evaluate it by utilizing an independent Finnish dataset.

Methods: Transfer learning was used, and the pre-trained model was finetuned with 8,829 examinations from the Finnish dataset (4,321 normal, 362 malignant and 4,146 benign examinations). Holdout dataset with 2,208 examinations from the Finnish dataset (1,082 normal, 70 malignant and 1,056 benign examinations) was used in the evaluation. The performance was also evaluated on a manually annotated malignant suspect subset. Receiver Operating Characteristic (ROC) and Precision-Recall curves were used to performance measures.

Results: The Area Under ROC [95%CI] values for malignancy classification obtained with the finetuned model for the entire holdout set were 0.82 [0.76, 0.87], 0.84 [0.77, 0.89], 0.85 [0.79, 0.90], and 0.83 [0.76, 0.89] for R-MLO, L-MLO, R-CC and L-CC views respectively. Performance on the malignant suspect subset was slightly better. On the auxiliary benign classification task performance remained low.

Conclusions: The results indicate that the model performs well also in an out-of-distribution setting. Finetuning allowed the model to adapt to some of the underlying local demographics. Future research should concentrate to identify breast cancer subgroups adversely affecting performance, as it is a requirement for increasing the model's readiness level for a clinical setting.

Nomenclature

AL	Active Learning
AP	Average Precision
AUPR	Area Under the Precision-Recall curve
AUROC	Area under the Receiver Operating Characteristic curve
BIRADS	Breast Imaging Reporting and Database System
CAD	Computer aided detection
CC	Bilateral craniocaudal
CI	Confidence Interval
CNN	Convolutional Neural Network
DICOM	Digital Imaging and Communication in Medicine
DL	Deep learning
FC	Fully-connected
FFDM	Full-field digital mammography
GPU	Graphics Processing Unit
MIS	Mammographic Information System
MLO	Mediolateral oblique
NYU-BCSD	New York University Breast Cancer Screening Dataset
PACS	Picture Archiving and Communication System
PR	Precision-Recall
QHAdam	Quasi-Hyperbolic Adam
ResNet	Residual neural network
ROC	Receiver Operating Characteristic
SOP	Standard Operating Procedure
UID	Unique Identifier
VOI LUT	Value of interest look-up table

1. Introduction

1.1. Background

Breast cancer is the most common form of cancer among women. There are approximately 2.3 million new female breast cancer incidents per year representing 11.7% of all cancer incidents. [41] One effective means to detect breast cancer is screening mammography. In screening mammography, symptomless women are imaged to detect possible malignant findings in an early stage. Full-field digital mammography (FFDM) is the standard means of imaging for screening purposes [32]. Typically, two projection images, i.e. bilateral craniocaudal (CC) and mediolateral oblique (MLO), are taken from both breasts [3]. These images are then visually interpreted by a radiologist. If a malignant finding is suspected, consensus reading is typically conducted, though practices and conventions can vary.

Several factors affect the interpretation from FFDM. Dense breast tissue can be accounted as one of the factors rendering detection of tumors difficult. Another factor is that observable objects in screening mammograms can vary greatly in size, some of the anomalies having sub-millimeter size. The latter creates high requirements, for example, for image resolution. Overall, the reading is considered to be a demanding task where experience and training have

*The authors received the following financial support for the research, authorship, and publication of this article: This study was funded by the Jane and Aatos Erkko Foundation, Helsinki, Finland, and the Technology Industries of Finland Centennial Foundation, Helsinki, Finland. S.I. Inkinen received funding from the Academy of Finland, Helsinki, Finland (project no. 316899). A. Isosalo received funding from the Jenny and Antti Wihuri Foundation, Helsinki, Finland (grant no. 210099).

**Approval for a register-based study was obtained prior to initiating the study from the City of Oulu (35/2019), Oulu, Finland, and the Northern Ostrobothnia Hospital District (179/2019), Oulu, Finland.

***The CLAIM guidelines [33] were followed when preparing this manuscript (where applicable).

*Corresponding author

antti.isosalo@oulu.fi (A. Isosalo)

ORCID(s): 0000-0002-5335-7535 (A. Isosalo); 0000-0002-9774-8925 (S.I. Inkinen); 0000-0003-2306-3111 (J. Reponen); 0000-0002-2300-2848 (M.T. Nieminen)

a key role [35]. It should be noted that the factors affecting the interpretation done by humans are also present in the automated image analysis.

1.2. Deep learning in mammography-based breast cancer evaluation

Computer aided detection (CAD) systems for breast cancer evaluation from mammography images has been studied extensively for several decades [9, 11, 12, 15]. Currently the focus is gradually shifting from conventional artificial intelligence-based solutions towards deep learning (DL) systems, and more precisely methods utilizing convolutional neural networks (CNN) for image evaluation. In this paradigm, relevant representations are learned directly from the data. [12] Recently applied DL methods for breast cancer detection and classification can be roughly divided into studies requiring pixel-level labels and strong supervision [1], and to those utilizing image-level labels and weak supervision [13, 47, 24, 22]. Pixel-level labels are typically utilized for segmentation purposes aiming, for example, for localization of a lesion. Training such a model involves comparing the predicted segmentation maps against expert annotations. With image-level labels the annotated meaning, whether some particular object category appears in the image, is given to an image as a whole. In this case, the training task deals with focusing on the object category and the patterns distinctive to it. However, it is worth mentioning, that CNN's trained on image-level labels are also applicable for object detection (*e.g.*, [39, 29]).

In a recent paper, Wu et al. [46] propose to perform screening examination classification by combining information simultaneously from several views, namely MLO and CC views. This is one of the few works processing mammograms at their native resolution in addition to the multi-view setting. Their classification method achieved the area under the receiver operating characteristic (AUROC) curve of 0.895 in detecting malignant and AUROC of 0.779 in detecting benign findings in examinations of a screening population. In a later study, Wu et al. [45] demonstrated that modality dropout (*i.e.*, methodology effectively masking out one of the views/modalities completely with a preset probability) has a positive impact on improving a model performance in a multi-view setting. Furthermore, they show that sharing weights between the classifier branches can boost the model performance. Another study approaching breast cancer detection with multi-view setting and learning simultaneously several breast cancer indicators in a multi-task manner was conducted by Kyono et al. [24] Their method was able to reach AUROC of 0.855 and area under the Precision-Recall (AUPR) curve of 0.646 in its main task of classifying samples as malignant or benign with biopsy results as reference. Akselrod-Ballin et al. [2] have concentrated on predicting early breast cancer utilizing machine learning models learning from clinical health records and digital mammograms in a multi-view setting. They demonstrated that their model can significantly reduce false-negative results.

Breast cancer classification from single and dual views has also been proposed in the literature. In a recent study, Shen et al. [39] propose a lesion localization and classification model which can be trained using only image-level

labels. For a better initialization of the model the authors use model parts from a related auxiliary classification task [39]. Their proposed model and architecture has benefits as producing pixel-wise reference contours for training can be a time-consuming task and requires usually special expertise. Single-view models in breast cancer classification have the obvious drawback of not being able to directly benefit from the added information provided by the supporting second view. Chen et al. [7] have approached the malignant classification task with a dual-view architecture. They developed a method for modelling the consistency in global feature representations. In addition, they implemented a method to estimate the relationship of local regions of the views. Their method was able to achieve AUROC of 0.948 on a private dataset [7].

Even though several interesting approaches have already been developed for breast cancer evaluation from mammography screening images, only few works [46, 45, 39, 29, 42, 21] have been reported to exploit the recently published models. Also, the published DL models many times lack evaluation with data from different populations than they have been trained with. Examples of different populations would include images from different manufacturers, different demographics, such as age distribution, and differences in the mammography scoring system (see Willemink et al. [43] for additional biases).

1.3. Objectives

In this retrospective study we have adopted an openly available deep learning model[1], which has been trained on one of the largest FFDM datasets reported in the literature with 1,000,000 mammograms, namely the NYU Breast Cancer Screening Dataset (NYU-BCSD) [47, 46]. Furthermore, we have performed an independent evaluation to assess the models performance by applying transfer learning (see Yosinski et al. [49]) scheme and finetuned the pre-trained model using Finnish mammograqphy screening data. At present, we have only little knowledge on how exclusive the learned correlations are to the original dataset and how to finetune such a model to adapt to differing demographics. What is known in advance is that the Finnish definition of benign differs from the Breast Imaging Reporting and Database System (BIRADS) [27], putting several underlying subclasses under a single category. Similarly, we are lacking information on what kind of richness is required from data used in finetuning and whether single center data from a small catchment area is enough to retain good model performance. To test this, we have created a reference standard that complies with the training data requirements described in [46]. The baseline has been set by running corresponding predictions using the pre-trained weights without finetuning. The benefit of transfer learning has been examined by training the model from random initialization instead of using pre-trained weights. Additional evaluation has been conducted by assessing the model performance on an independent Portuguese INbreast dataset [34]. We have made our annotation tool [20], used in facilitating this study, openly available for the scientific community.

[1] https://github.com/nyukat/breast_cancer_classifier

2. Materials and methods

2.1. Dataset and curation

2.1.1. Finnish dataset

For the secondary use of the existing data, we obtained a permit for a registry-based study from the Northern Ostrobothnia Hospital District (179/2019), Finland, and the City of Oulu (35/2019), Finland, prior to initiating the data extraction. A dataset of 49,634 mammography screening examinations (with 22,739 unique patients) conducted over the 2011-2019 period were extracted. The extracted data originated from two different data sources, namely from the picture and archiving communication system (PACS) and the mammographic information system (MIS). Each examination in the dataset contained digital screening mammograms originating from the PACS in Digital Imaging and Communication in Medicine (DICOM) format. Imaging data was coupled with textual information in a machine-readable form from the MIS. Among the data from the MIS were, for example, screening assessments, patient age, and possible confirmation study results, such as additional imaging results and histology responses. As extraction conditions, examinations with patients who had gone through mastectomy, examinations which had breast implants present, and patients for which the hospital care was ongoing were left out from the data collection. All data was de-identified after linking the images with the MIS data, effectively assigning each patient an anonymous universal identifier. The MIS data was further supplemented with a column to indicate the interval of screening for patients with several examinations in the dataset.

Several exclusions were made according to pre-established constraints on an examination-level to curate the extracted dataset (Figure 1). The dataset was curated to have only examinations with images having Standard Operating Procedure (SOP) Class Unique Identifier (UID) than `1.2.840.10008.5.1.4.1.1.1.2` referring to Digital Mammography X-Ray Image Storage for presentation (Figure 1a). Furthermore, referring to the model architecture (Figure 2), also examinations which did not have all four standard views, namely R-MLO, L-MLO, R-CC and L-CC present, were excluded from the dataset (Figure 1a). The training data was further standardized by limiting to native resolutions of 3062-by-2394 and 2294-by-1914 (Figure 1a). The vast majority of exclusions were done from the pool of normal and benign examinations with no label-preserving (future) screening study or an endpoint of any kind, and having irregular screening interval (Figure 1b). It should be noted that, all malignant suspect examinations, thus having a referral study (*e.g.*, biopsy), with the exception of patients living with breast cancer, were kept in the dataset. The conducted exclusions resulted a total of 11,037 examinations with 4 standard views for the experiments (Table 1).

Regarding the imaging equipment, all digital mammograms in the dataset originated from single manufacturer devices, namely Senograph Essential and Senograph Essential DS (GE Healthcare, Chicago, Illinois, US). The mammograms had a bit depth of either 12 or 14 bits and pixel spacing of either 100 μm or 94 μm. Organ Dose (mean

glandular dose) varied within interval $[0.423, 4.144]$ mGy with median being at 1.063 mGy. Peak kilo voltage output of the X-ray generator ranged within interval $[26.0, 31.0]$ kVp. X-ray tube current varied within interval $[59, 100]$ mA in the dataset. Mo/Mo and Rh/Rh (anode/filter) were used. It should be noted that, the manufacturers represented in the NYU-BCSD [47] does not include GE Healthcare devices, *i.e.*, the data on which the model that we evaluate in this study was originally trained with. Though this will not interfere with the planned finetuning experiments, it may have some effects on the performance of the pre-trained model used as baseline without finetuning.

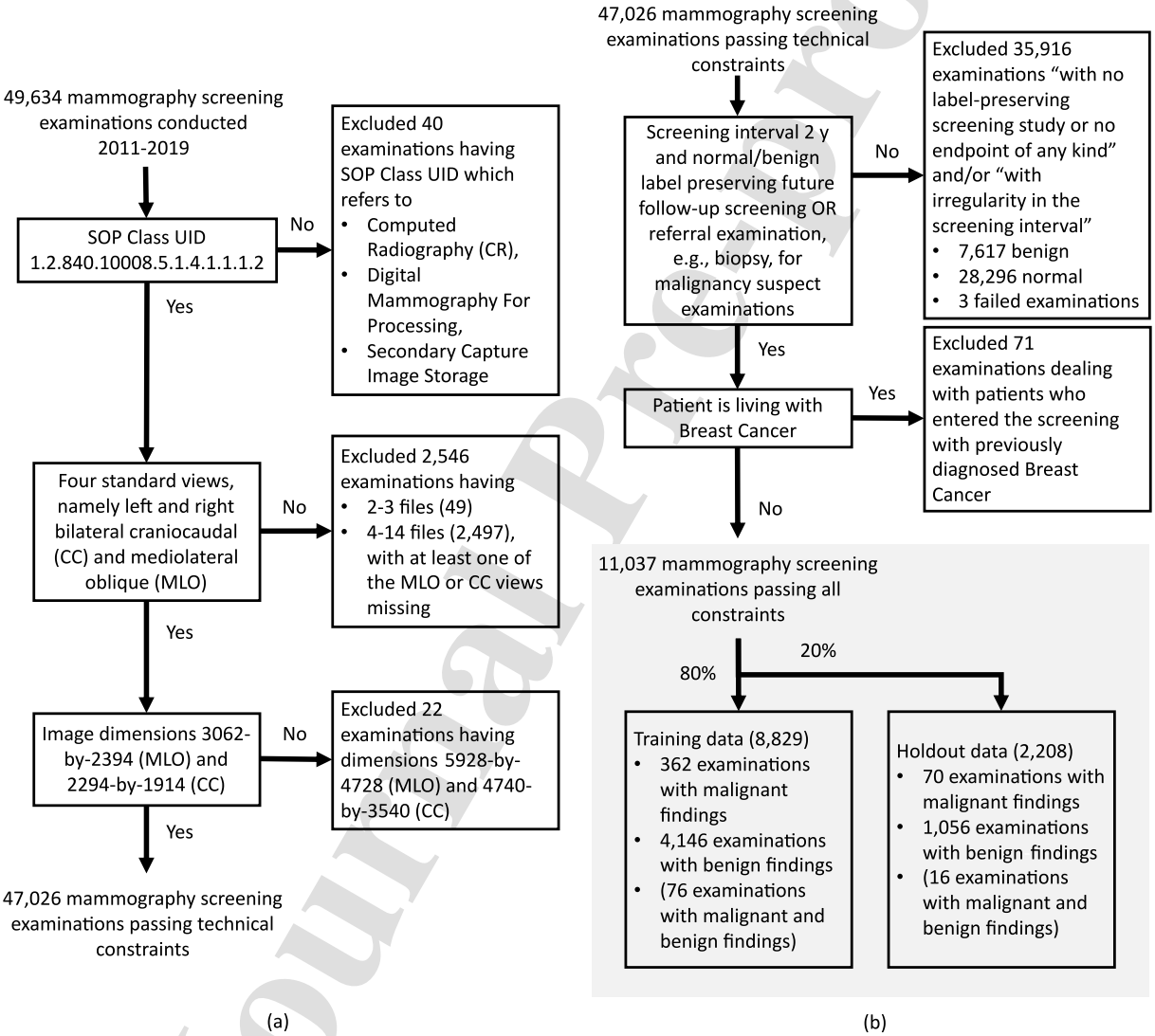


Figure 1: Exclusion flowchart with pre-established (a) technical constraints and (b) clinically oriented constraints posed on the extracted data. The 80-20 split, highlighted with grey, was stratified by using patient identifier and screening assessment result, resulting 8,829 examinations (with 7,189 unique patients) for training and 2,208 examinations (with 2,084 unique patients) for evaluation (holdout subset). SOP stands for Standard Operating Procedure.

Table 1
Implications of the conducted exclusions to the distribution of underlying characteristics of the dataset, exclusions were made according to pre-established constraints, namely Standard Operating Procedure Class for Digital Mammography, view and laterality, image dimensions, screening interval and a requirement for some type of an endpoint, *e.g.*, a label-preserving follow-up study

Characteristic	Passing technical constraints	Passing all constraints
Age group (y)		
49	5,363 (11.40%)	1,664 (15.08%)
50-59	23,272 (49.49%)	5,879 (53.27%)
61-69	18,391 (39.11%)	3,494 (31.65%)
=>70	0 (0%)	0 (0%)
Screening assessment		
Normal	32,140 (68.35%)	3,830 (34.70%)
Benign	11,949 (25.41%)	4,298 (38.94%)
Malignancy cannot be ruled out	2,818 (5.99%)	2,793 (25.31%)
Highly suspicious of malignancy	110 (0.24%)	110 (1.00%)
Malignant	6 (0.01%)	6 (0.05%)
Referral study (*e.g.*, biopsy)		
yes	2,934 (6.24%)	2,909 (26.36%)
no	44,092 (93.76%)	8,128 (73.64%)
Total	47,026 (100%)	11,037 (100%)

2.1.2. Portuguese dataset

As a supplementary evaluation dataset of digital mammograms, 86 examinations with 4 standard views from a well-known Portuguese INbreast dataset [34] were used. Among those were 36 malignant (of which 28 containing also benign findings), 47 benign and 3 normal examinations. All mammograms were reported to originate from single manufacturer device, namely Mammomat Novation (Siemens AG, Munich, Germany), with bit depth of 14 bits and pixel spacing of 70 μm. The anode target and filter materials were not reported. [34] This dataset was used in full for evaluation purposes (see Section 2.8).

2.2. Data post-processing

As a post-processing step, value of interest lookup table (VOI LUT) mapping was performed to standardize the images using Python based pydicom [31] library (version 1.4.2). Moreover, DICOM files in the INbreast dataset were repaired with a proper Siemens VOI LUT, as these were found to be missing. Additionally, the original images with two different native resolutions of 3062-by-2394 and 2294-by-1914 in the Finnish dataset and 3328-by-2560 and 4084-by-3328 in the Portuguese dataset were cropped to remove some of the background (areas not containing useful information for the breast cancer classification) prior to training to reduce the processing time (see [46] for details). Finally, the images were padded to match the dimensions expected by the deep learning model on the fly during model training/finetuning (see Section 2.5).

2.3. Reference standard

The reference standard for each examination was defined as eight (8) binary labels, *i.e.*, four labels indicating malignancy (labelled as 1) or the absence of it (labelled as 0) and four for benignity (labelled as 1) or the absence of it (labelled as 0), similarly as in [46]. As a result, for examinations which had received screening assessment "normal" on both sides, right and left, all eight labels were filled with zeros. Furthermore, labels for those examinations with assessment "benign" for either or both sides, were populated with ones and zeros accordingly. It should be noted that a source of error was introduced here since the MIS data available for this study did not provide information in which of the projections CC or MLO the benign mammographic change was present. Considering the fact that the breast tissue does not fully overlap for the CC and MLO projections, the projections should not, in principle, receive the same label by default.

With several simultaneous malignant and benign findings present in a single screening examination, the most severe findings taking precedence, manual annotations were carried out to refine the labelling. The manual annotations were performed for a subset of examinations (2, 934), having consensus reading assessment "malignancy cannot be ruled out" resulting from the screening (performed by at least two certified radiologists). The examinations were labelled by radiology resident T. T. Moreover, annotations provided necessary information on which of the projections the confirmation studies proven findings were radiologically visible. MATLAB (2020a, Massachusetts, United States) -based in-house mammogram annotation tool [20] (version 1.0) was used for the work. Furthermore, all histology codes present in the MIS were labelled to indicate either malignancy or benignity. These were then compared against malignancy-benignity labels derived from the manual annotations.

For the Portuguese dataset, the reference standard was derived from the BIRADS scores provided with the data. Negative examinations were labelled as "normal", examinations which had received assessment from "benign" to "low suspicion for malignancy" were labelled as "benign". All above or equal to "moderate suspicion for malignancy" were labelled as "malignant".

2.4. Training data sampling

The Finnish data was split into a training set and a holdout set (the latter was kept aside during the method development) having proportions 80% and 20% of the total amount of curated and confirmed samples respectively, *i.e.*, according to the Pareto Principle. The splitting was performed using StratifiedShuffleSplit function from publicly available scikit-learn package [37] (version 0.22.2.post1), which creates a stratified split preserving the percentage of samples for each class. Furthermore, the selection was randomized. Stratification was performed here according to the screening scores (see Section 2.1). In addition, we extended the default ability of the function by introducing an encoded categorical

combining the left and right side screening assessments to ensure findings on both left and right side images are present in the splits, considered crucial for successful training.

For training, the data was split into 5 folds using a K-fold iterator with non-overlapping groups, namely GroupKFold function with anonymous Patient ID as group identifier. The utilized splitting scheme ensured that all examinations from a patient were either in training or testing subset. Due to the small number of malignant examples in the dataset, 362 examinations in total in the training set, we resampled the dataset by oversampling the examples that have malignant finding in the breast on either side. Thus, additional copies were sampled randomly with replacement to increase the number of malignant examples fivefold, i.e., the number of malignant examinations was increased up to 1,810 examinations, where all original examples appear at least once.

During the inference of the holdout set, the data was from the original distribution with the reported exclusions (Figure 1). Here, the inference refers to applying the trained model to unseen data samples for prediction output.

2.5. Deep learning model architecture

A dedicated convolutional neural network (CNN) is used for learning from all four standard views in a multi-view fashion (Figure 2). The architecture is designed to manage high-resolution input without the need for down-scaling of the input mammograms [46]. For each of the views there is a dedicated 22-layer ResNet (ResNet22) with adjusted spatial resolutions, namely depth and width, and stride, in comparison to the ordinary ResNet [16]. The ResNet22 outputs a hidden representation of each of the projections, which are then concatenated view-wise for CC and MLO views. After the concatenation two fully-connected layers are employed in learning the non-linear combinations of the high-level features in the data and interactions between the views. The resulting view-wise activations are converted as probabilities (real values that sum up to 1) using a PyTorch LogSoftmax function [36].

Originally, the model has two variants, namely 1-channel image-only model with a single input and a 3-channel variant with three inputs [46]. In this work, we concentrate on the 3-channel variant. The 3-channel model receives the original mammogram as an input into the first channel. The other two channels receive specific heatmap representations of the original mammogram for possible malignancy and benignity (Figure 3). The heatmaps are generated using an auxiliary patch-level network. Furthermore, the 3-channel model expects all inputs to have dimensions of 2677-by-1942 pixels and 2974-by-1748 pixels for the CC and MLO views respectively (Figure 2). For the experiments, the auxiliary heatmap generating network was adopted as implemented and pre-trained by Wu et al. [46] and its predictions were used as such.

2.6. Model training

For the experiments, the model state dictionary of the Wu et al. [46] 3-channel model variant was loaded and finetuned using our Finnish training subset (Figure 1b). For the optimizer, the Quasi-Hyperbolic Adam [30] was used. Moreover,

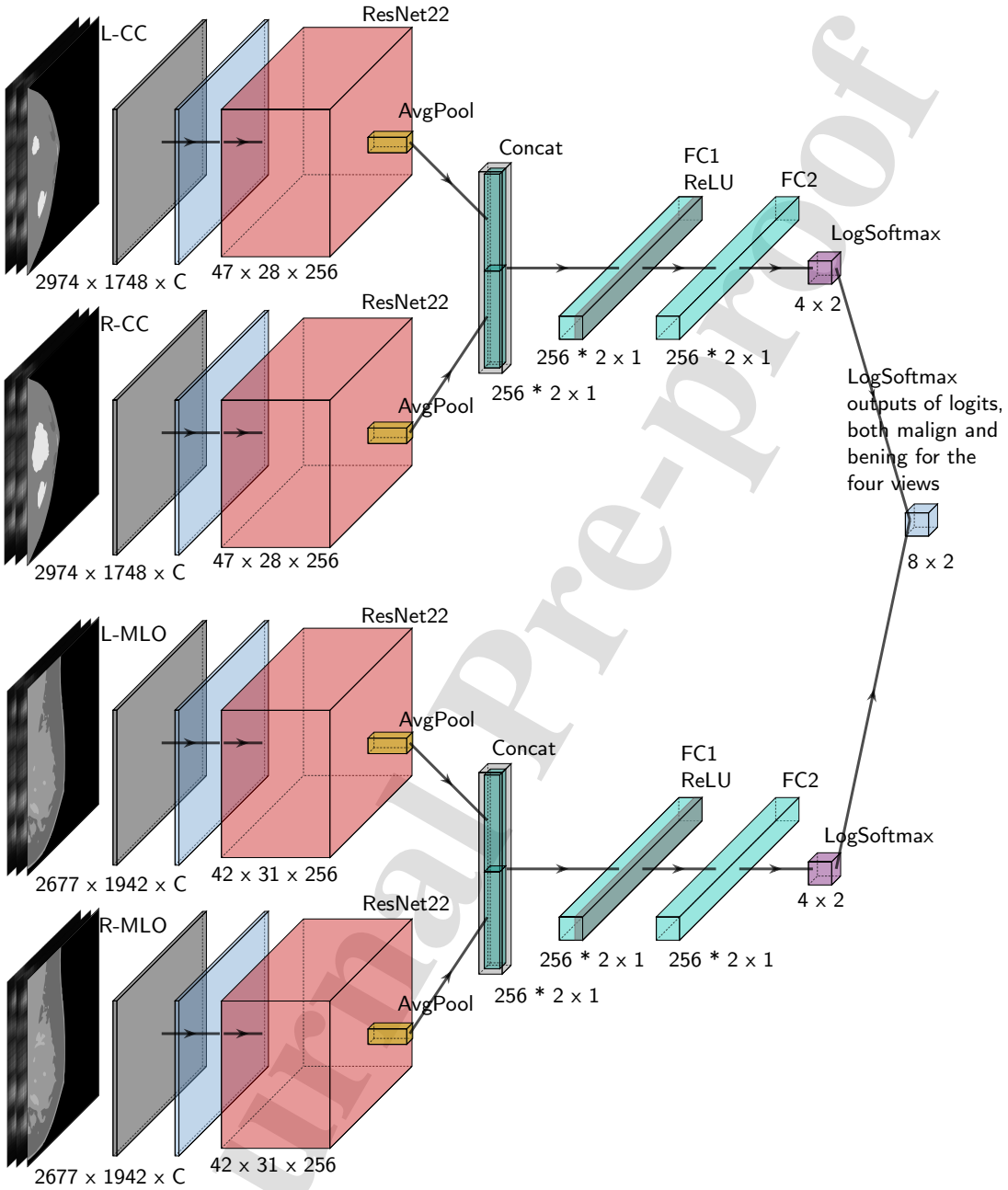


Figure 2: Multi-view model architecture. Images are inputted into the model so that the left side view representatives are flipped to have the same orientation as the right side view representatives. Gaussian noise with a standard deviation of 0.01 is added to the images prior ResNet22 block (depicted with a blue cuboid). For each of the views there is a dedicated ResNet22 with adjusted spatial resolutions. Number of channels is doubled by each ResNet block of the ResNet22 resulting a H-by-W-by-256 tensor as output, where $H = 42$ and $W = 31$, and $H = 47$ and $W = 28$ for the CC and MLO views respectively. Stride of two (2) is used to reduce the feature map size. ResNet has ReLU [23] and Batch Normalization [18] layers included. Fully-connected layers compile the information extracted by the preceding layers. Illustration was made using PlotNeuralNet tool [19].

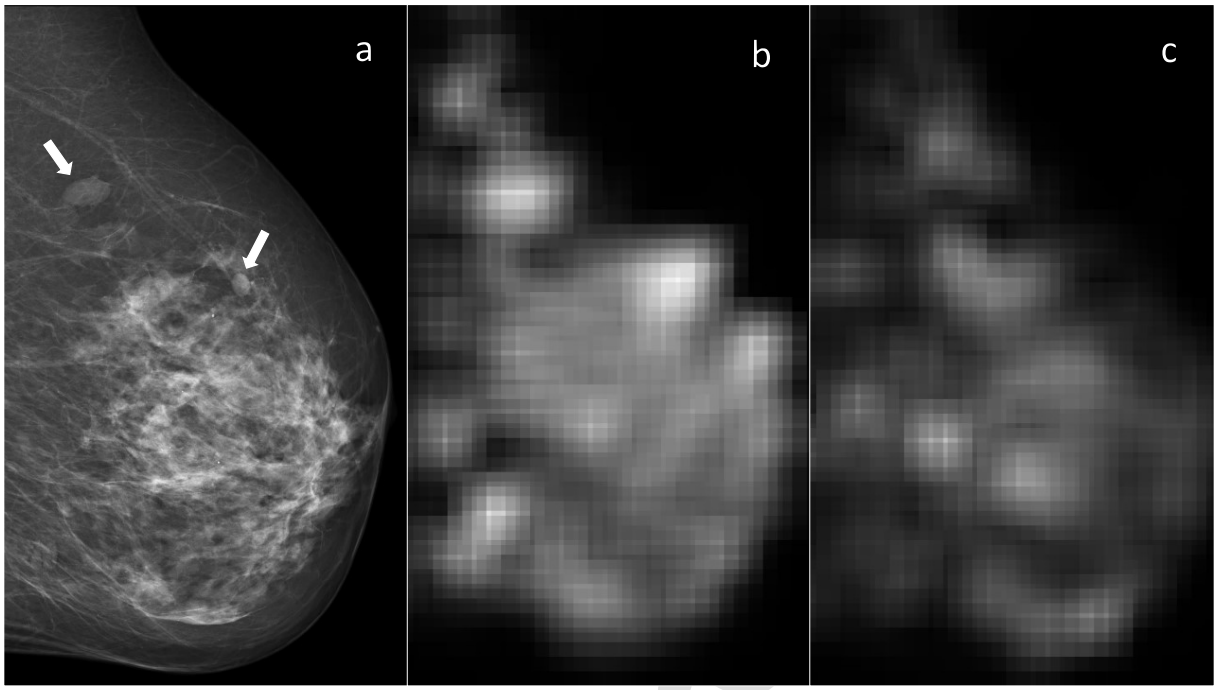


Figure 3: 3-channel model L-CC-branch inputs: (a) original VOI LUT mapped image, (b) malignant and (c) benign heatmaps. Arrows in the leftmost image point out two benign masses. The auxiliary patch-classifier responsible for the heatmap generation was used without finetuning. Mammogram courtesy of Breast Research Group, INESC Porto, Portugal.

parameter values $\nu_1 = 0.7$ and $\nu_2 = 1$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon =$ 1e-8 from the Ma et al. [30] were chosen for the experiments. Separate learning rates, 1e-4 and 1e-7, were applied to model base parameters and output layer parameters correspondingly. Optimizer weight decay was set as 1e-5. Furthermore, no learning rate scheduler was used. Based on the quantity and how we chose to split our data, we further chose to finetune the whole model and not just the fully-connected layers. The base layers of the model were refrained from training (frozen) during the first 5 epochs to prevent fully-connected layers from corrupting the pre-trained weights of the convolutional layers during backpropagation. Thus, the errors from the new task were not propagated to the frozen weights [49] during backward pass in the early stages of finetuning.

To comply with the Wu et al. [46] model definition, we used the PyTorch Negative Log-Likelihood Loss [36] (calculating the mean loss), which follows from the use of LogSoftmax in the model. During finetuning, training loss was calculated based on the eight labels defined in Section 2.3. In this work, we computed the loss separately for four CC and four MLO view outputs (logits related to the tasks predicting benignity and malignancy) and summed those to get a loss for a batch.

Following Wu et al. [46] implementation, image pixels were standardized in-place by subtracting the image mean and dividing with the image standard deviation. Augmentations used in Wu et al. [46], namely adding variation to window size and location, were adopted as such. This process included padding the images to match the

dimensions expected by the pre-trained deep learning model. Further details can be found in [46]. Additionally, Cutout augmentation [8] was used in randomly masking out input image sections during training for regularization. Moreover, Cutout was performed randomly with probability of 0.3. Cut out size was 500-by-500 pixels, and number of masked out sections was 3. Augmentations were performed on the fly during finetuning. Furthermore, no augmentations were used during validation stage.

We used training batch size of 8 (twice the size used in [46], making the training more well-conditioned) and validation batch size of 12. Number threads was 10. Fold training time was set to 70 epochs and the best performing model was saved whenever the validation loss was less than with the previously saved model checkpoint. To guarantee the reproducibility of the experiments, our training pipeline was implemented to store the previously mentioned parameters, such as the choice of optimizer and its parameters, learning rate, batch-size, model choice, and cross-validation split indices.

Furthermore, the model finetuning and evaluation was carried out on a single 4,608-core Turing architecture Titan RTX (NVIDIA, Santa Clara, California, US) graphics processing unit (GPU) with 24 GB of memory. In our implementation, finetuning of all network layers reserved approximately 22 GB of GPU memory. Completing a 5-fold training to produce models for the method evaluation phase took approximately 14 days using our hardware.

2.7. Method evaluation

The method was evaluated using the Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve representing classification performance at various thresholds and the corresponding AUROC and Average Precision (AP) metrics and their 95% Confidence Intervals (CI) (values ranging from 0 to 1) estimated from 2,000 bootstrap samples. Each of the standard views received their own ROC and PR curves in our evaluation—for both malignant and benign classification task. Youden index [50], and more precisely its maximum value, was used to find ROC curve cut-off points where the absolute value of the difference between sensitivity and specificity is at its minimum. Sensitivity and Specificity were calculated in those cut-off points for convenience. In addition, we reported the maximum F1 score for each standard view and classification task (50 evenly spaced values were calculated over a specified interval of [0, 1]). Furthermore, the index of the maximum F1 score among the calculated ones, was used for finding the corresponding values for Precision and Recall. Furthermore, we averaged the predictions of the five models resulting from the 5-fold cross-validation prior computing the listed evaluation metrics.

2.8. Test populations

Three test populations were chosen to study our research questions and perform model evaluation. First, the annotated all-malignant suspect subset (S1) of the Finnish holdout data (Figure 1) was used for the inference. Second, the whole Finnish holdout subset (S2) of data, where also samples of "normal" and "benign" which were assumed not to contain

any malignant characteristics, was used. Third, completely unseen examinations of the Portuguese data (S3) was used to study the training/finetuning effects. Wu et al. [46] pre-trained model weights were used here as a baseline in all experiments.

3. Results

In the following we present the ROC and PR curves for model predictions for each standard view. Youden index, Sensitivity, Specificity, F1 score, Precision and Recall values for the malignant and the benign classification tasks are given in the Appendix A (Table A.1 and Table A.2).

3.1. Evaluation using the Finnish data

3.1.1. Annotated malignant suspect holdout subset (S1)

Finetuning only slightly improved the malignant classification in terms of AUROC and AP (Figure 4a). The proportion of positive malignant samples (differs for different projections) in the annotated malignant suspect holdout dataset were 0.09 and 0.09 for the L-MLO and L-CC and 0.09 and 0.10 for the R-MLO and R-CC projections respectively as depicted by the dotted lines in the PR curves. For the benign classification the finetuning appears to have been somewhat harmful (Figure 4b) and the benign detection in this subset lowered surprisingly for the CC view. There was minimal improvement on the left side when looking at the PR curve representing the MLO view (Figure 4b). The proportion of positive benign samples in the annotated malignant suspect holdout dataset were 0.27 and 0.30 for the L-MLO and L-CC and 0.27 and 0.28 for the R-MLO and R-CC projections, respectively.

3.1.2. Entire holdout subset (S2)

Adding exams which have received assessment "normal" or "benign" in the screening phase introduced false positive predictions which is depicted in the PR curves (Figure 5a). The suboptimal performance depicted by the baseline L-CC and R-CC curves for the benign classification seem to have been slightly rectified with the finetuning (PR curve in Figure 5b). The proportion of positive malignant samples in the holdout dataset were 0.02 and 0.02 for the L-MLO and L-CC and 0.02 and 0.02 for the R-MLO and R-CC projections respectively, and the proportion of positive benign samples 0.33 and 0.33 for the L-MLO and L-CC and 0.31 and 0.34 for the R-MLO and R-CC projections.

3.2. Evaluation using the Portuguese data (S3)

The malignant classification performance on Portuguese evaluation data was not greatly affected by the finetuning (Figure 6), which was performed solely using the Finnish data. The proportion of positive malignant samples for the Portuguese data were 0.23 and 0.23 for the L-MLO and L-CC and 0.20 and 0.19 for the R-MLO and R-CC projections respectively. For benign classification, there is a noticeable improvement seen in the CC projection ROC and PR curves,

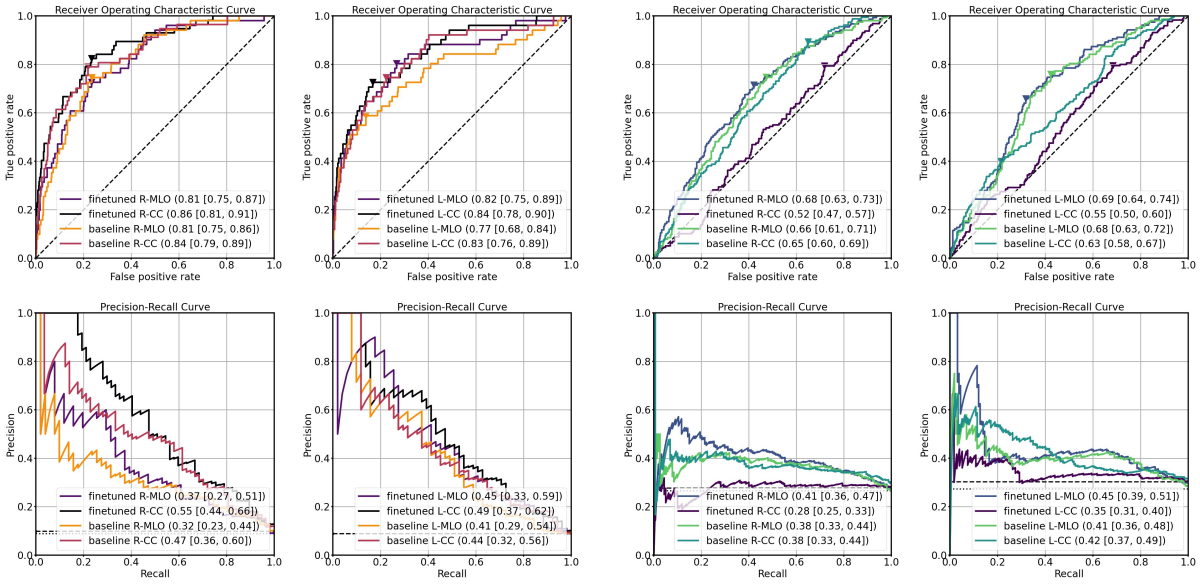


Figure 4: Results for annotated malignant suspect subset (S1) of the Finnish holdout data. Here the pre-trained model without finetuning as baseline is compared to the model finetuned on our Finnish data. Results are depicted laterality-wise. In the subfigures, depicted are the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves for the (a) malignant and (b) benign classification tasks. Dotted diagonal line in ROC curve represents a classifier which gives random results. The dotted vertical lines in the PR curve plot resemble a classifier which gives random results, which is different for each projection due to the amount of examples per side and view. In the PR curve the dotted line also shows the proportion of positive samples in the used holdout dataset. AUROC and AP and their 95% CI are shown in the legend for ROC and PR curves correspondingly. cut-off points, marked with an asterisk in the ROC curves are achieved via Youden index, its maximum value.

slightly surpassing the uninformative classifier, though with poor early retrieval for L-CC (Figure 6b). The proportion of positive benign samples were 0.58 and 0.59 for the L-MLO and L-CC and 0.67 and 0.69 for the R-MLO and R-CC projections respectively.

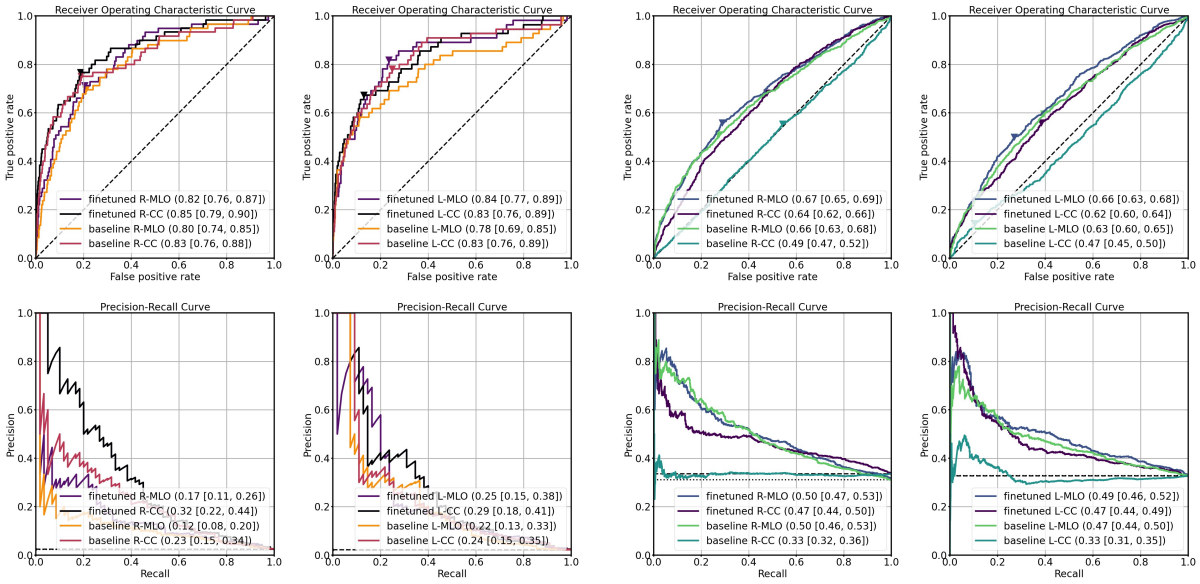
3.3. Training from random initialization

Considering our relatively large dataset, (re-)training the models from random initialization resulted in a rather poor performance for both S1 and S2 when compared to the baseline and the finetuned models (Figure 7 and Figure 8).

4. Discussion

4.1. Immediate findings and comparison to existing work

Model performance on the malignant classification task was satisfactory on the Finnish holdout data, and especially on the annotated malignant suspect subset of the holdout data. The AUROC values for malignancy classification obtained with pre-trained model weights initialization were 0.81, 0.82, 0.86, and 0.84 for R-MLO, L-MLO, R-CC



(a) (b)

Figure 5: Results for the entire Finnish holdout subset (S2) with also normal and benign examinations without suspicion of malignancy in addition to examinations suspected of malignancy. Here the pre-trained model without finetuning as baseline is compared to the model finetuned on our Finnish data. Results are depicted laterality-wise. In the subfigures, depicted are the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves for the (a) malignant and (b) benign classification tasks. Dotted diagonal line in ROC curve represents a classifier which gives random results. The dotted vertical lines in the PR curve plot resemble a classifier which gives random results, which is different for each projection due to the amount of examples per side and view. In the PR curve the dotted line also shows the proportion of positive samples in the used holdout dataset. AUROC and AP and their 95% CI are shown in the legend for ROC and PR curves correspondingly. Cut-off points, marked with an asterisk in the ROC curves are achieved via Youden index, its maximum value.

and L-CC views respectively for the malignant suspect subset (S1), whereas the average AUROC obtained by Wu et al. [46] for their view-wise model was 0.843 for their biopsied subpopulation. The malignancy classification AUROC values for the entire holdout subset (S2) were 0.82, 0.84, 0.85, and 0.83 for R-MLO, L-MLO, R-CC and L-CC views respectively, whereas the average AUROC obtained by Wu et al. [46] was 0.886 for their screening population. Even though AUROC is considered prevalence invariant direct comparison should be taken with caution. The results with the Portuguese dataset suggest that the finetuning which was conducted solely with the Finnish data does not harm the models generalizing ability to unseen examinations from the INbreast dataset.

Model performance on the auxiliary benign classification task remained low. This is not surprising, since the Finnish class "benign" has more within class variation (see Section 4.2) than the BIRADS "benign", the latter in essence defined as "0% probability of malignancy". The AUROC values for benign classification obtained with pre-trained model weights initialization were 0.68, 0.69, 0.52, and 0.55 for R-MLO, L-MLO, R-CC and L-CC views respectively for the malignant suspect subset (S1), whereas the average AUROC obtained by Wu et al. [46] for their view-wise

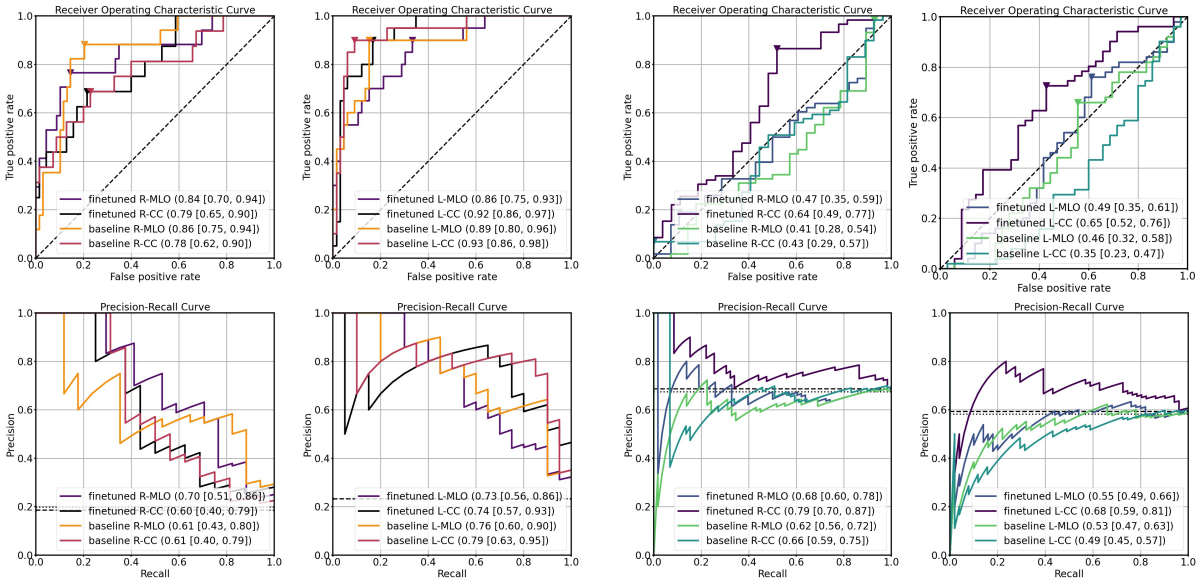


Figure 6: Results for the independent Portuguese INbreast data (S3). Here the pre-trained model without finetuning as baseline is compared to the model finetuned on our Finnish data. Results are depicted laterality-wise. In the subfigures, depicted are the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves for the (a) malignant and (b) benign classification tasks. Dotted diagonal line in ROC curve represents a classifier which gives random results. The dotted vertical lines in the PR curve plot resemble a classifier which gives random results, which is different for each projection due to the amount of examples per side and view. In the PR curve the dotted line also shows the proportion of positive samples in the used holdout dataset. AUROC and AP and their 95% CI are shown in the legend for ROC and PR curves correspondingly. Cut-off points, marked with an asterisk in the ROC curves are achieved via Youden index, its maximum value.

model was 0.690 for their biopsied subpopulation. The AUROC values for benign classification for the entire holdout subset (S2) were 0.67, 0.66, 0.64, and 0.62 for R-MLO, L-MLO, R-CC and L-CC views respectively, whereas the average AUROC obtained by Wu et al. [46] for their view-wise model was 0.747 for their screening population for a comparison. Although the benign classification results were not encouraging, learning several tasks simultaneously can be important for the overall learning task [6] (see also Section 4.3). It is fair to note, that our choice not to finetune the auxiliary patch classifier producing the two heatmap inputs for the 3-channel model may have a role regarding the weak benign classification results. Finetuning the patch classifier would have though required us to also annotate some of the non-malignant suspect examinations. What we can say with high certainty is that the malignant heatmap generation would not have benefited the finetuning with our sample amounts, if not to cause adverse effects.

Apart from the comparisons performed against the Wu et al. [46] results on their private dataset, comparison to other methods found in the literature was difficult, as there are only few studies experimenting with a multi-view architecture. Moreover, many of the methods presented in the literature concentrate only on malignant classification or the experimental setup differs in some other regard. Kyono et al. [24] have studied an interesting architecture

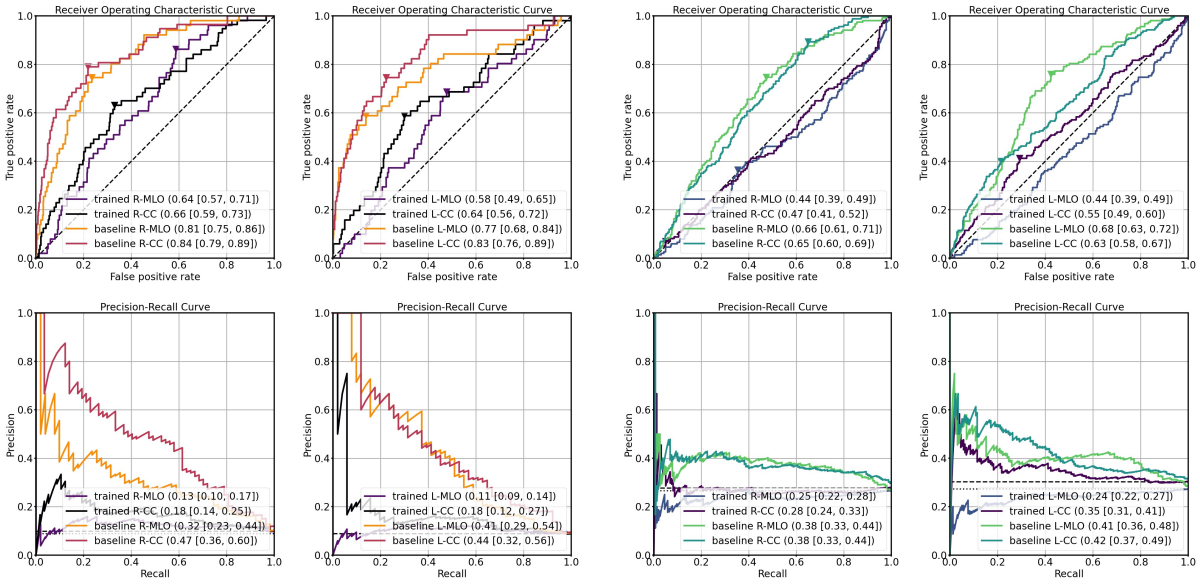
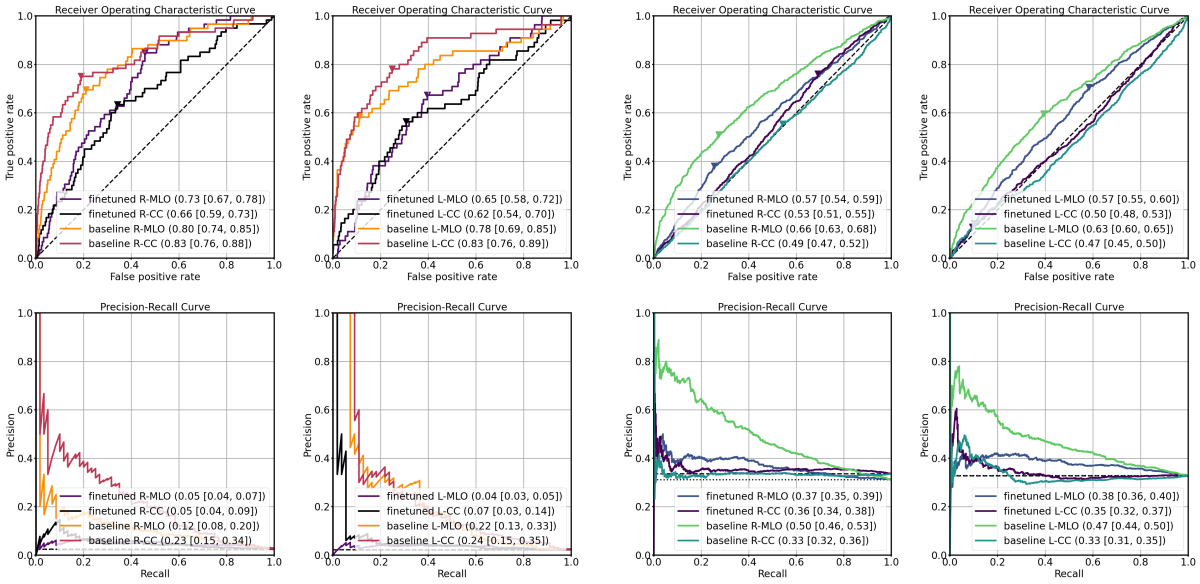


Figure 7: Results for annotated malignant suspect subset (S1) of the Finnish holdout data. Here the pre-trained model without finetuning as baseline is compared to the model trained from random initialization on our Finnish data. Results are depicted laterality-wise. In the subfigures, depicted are the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves for the (a) malignant and (b) benign classification tasks. Dotted diagonal line in ROC curve represents a classifier which gives random results. The dotted vertical lines in the PR curve plot resemble a classifier which gives random results, which is different for each projection due to the amount of examples per side and view. In the PR curve the dotted line also shows the proportion of positive samples in the used holdout dataset. AUROC and AP and their 95% CI are shown in the legend for ROC and PR curves correspondingly. Cut-off points, marked with an asterisk in the ROC curves are achieved via Youden index, its maximum value.

performing multi-task learning of diagnosis, imaging phenotype which the authors have named as sign, level of suspicion, conspicuity referring to the radiological occultness, breast density, and patient age, while combining also non-imaging features, an idea worth further research. Their method achieved AUROC of 0.855 in classifying samples as malignant or benign (treating benign as an equivalent to normal) on their private dataset, collected from six screening centers in UK. They also reported an AUPR of 0.646 for the same task, but the exact proportions of positive samples were not provided. In comparison to the dataset used in this study, their dataset of 7,060 examinations is particularly rich, with more than 1,000 malignant examinations [24, 14] and resembles to some extent our all-malignant suspect subset (S1), and the AUROC value is of the same magnitude as in our study. AUPR is influenced by class prevalence and direct comparison to results obtained with other datasets should be avoided. A recent study by Chen et al. [7] utilizing dual-view architecture with modestly downscaled CC and MLO views as an input. Their method uses weak image-level labels similarly to the method evaluated in this study. Their method yielded the highest malignancy classification AUROC [95% CI] among the works we have came across, 0.948 [0.937, 0.953] on their training dataset of 139,034 examinations, of which 5,901 cancerous, and 0.926 [0.922, 0.930] on their holdout dataset of 1,691,654



(a) (b)

Figure 8: Results for the entire Finnish holdout subset (S2) with also normal and benign examinations without suspicion of malignancy in addition to examinations suspected of malignancy. Here the pre-trained model without finetuning as baseline is compared to the model trained from random initialization on our Finnish data. Results are depicted laterality-wise. In the subfigures, depicted are the Receiver Operating Characteristic Curve (ROC) and Precision-Recall (PR) curves for the (a) malignant and (b) benign classification tasks. Dotted diagonal line in ROC curve represents a classifier which gives random results. The dotted vertical lines in the PR curve plot resemble a classifier which gives random results, which is different for each projection due to the amount of examples per side and view. In the PR curve the dotted line also shows the proportion of positive samples in the used holdout dataset. AUROC and AP and their 95% CI are shown in the legend for ROC and PR curves correspondingly. Cut-off points, marked with an asterisk in the ROC curves are achieved via Youden index, its maximum value.

examinations, of which 5,232 cancerous. For a subset of Portuguese INbreast dataset Chen et al. [7] report AUROC [95% CI] 0.994 [0.985, 1.000] and AUPR [95% CI] 0.986 [0.966, 1.000] for malignant classification. Even though we have used a different subset (86 examinations with 36 malignant) of the Portuguese dataset for evaluation than Chen et al. (31 examinations with 15 malignant), a similar trend in the results can be observed. Comparison of various studies presenting multi-view models for breast cancer evaluation cited by this paper can be found in Table 2.

Some conclusions of the clinical relevance of a method can be drawn by comparing to human-level performance. A recent study utilizing 2,512,577 mammography screening examinations from 146 screening centers in the US reported Sensitivity for the initial screening assessment 0.85 [0.848, 0.86.] and Specificity 0.918 [0.910, 0.974] [40]. Another study benchmarking mean screening performance reported Sensitivity [95%CI] of 0.869 [0.863, 0.876] and Specificity [95%CI] 0.889 [0.888, 0.889] [25]. As we can see from the malignant classification results (Appendix A, Table A.1), the evaluated method was not able to retain this level on our private dataset

Comparison performed using various metrics (Appendix A, Table A.1 and Table A.2) mainly confirmed what was already seen in the ROC and PR curves. Literature suggests that for well calibrated model outputs the optimal threshold would be approximately half of the optimal F1 value [28]. This seems to hold for our experiments even though the model outputs have not been calibrated. Furthermore, high F1 score usually guarantees good results also when there is an inherit imbalance between the positive and negative class, such as the breast cancer screening scenario where the proportion of malignant samples is very low (*e.g.*, for our private holdout dataset this proportion was 0.02). With a low F1 score there is no telling which type of error the model is suffering. It should be noted, that using the highest F1 score for finding the corresponding values for Precision and Recall might not be the best choice for this application, but it was the only reproducible way to determine the index that can be offered here.

Successfully training and finetuning a multi-view models is still an open research question. Studies which have experimented on the same model as we have, have reported that the model has a tendency of ignoring a less predictive modality (i.e., one of the views MLO or CC) [46, 45] and our results support these findings. Recent research has offered a so-called greedy learner hypothesis [44] as an explanation, *i.e.*, that the model learns from the modality from which it is fastest to learn. Suboptimal view-wise performance which can not be explained by noisy data (*e.g.*, quality issues with the medical record or errors in annotations) is seen on the CC view. Our observation on the other hand was that a careful choice of optimizer may alleviate this phenomenon (see also Section 4.3). Slight differences in cut-off points (see for example Figure 5a) though remain.

We note that in contrast to Wu et al. [46] the results were achieved without test-time augmentation during inference, thus there can be slight decrease in the performance metrics as such method can produce more favorable (for the model) variations of the unseen input. Our aim was to study the effects of the training to the model as such without post-training techniques for improved performance. There is only little research on the design choices of test-time augmentation [38]. The baseline results, *i.e.*, without finetuning, indicate that the patterns in breast tissue appear alike even for different vendors in full-field digital mammography and that the model performs well also in an out-of-distribution setting.

Lastly, in our experiments optimizing the whole feature space, i.e., updating the weights of the whole network resulted the best outcome. Our finding is contrary to Wu et al. [45] as they found in their experiments that it is better to keep the base of the model (other than the FC layers) frozen. Moreover, finetuning was best done using the entire training subset.

4.2. Reference standard considerations

Assigning reference labels should reflect the patient records as well as what is actually seen in the images. It is for example important to notice that MLO and CC views should not automatically receive the same image level cancer label, as the views represent partially different tissue areas. This is rarely mentioned in studies regarding

Table 2
Comparison of various studies utilizing multi-view models for breast cancer evaluation

Study	Method	Dataset used	Experimental setup	Performance metrics
Kyono et al. (2019) [24]	InceptionResNetV2-based multi-task CNN; pre-trained weights from ImageNet classification	NHS Breast Screening Program (NHSBSP) dataset originating from six screening centers [14]; 7,060 examinations, with more than 1,000 malignant examinations	Amount of 2,000 examinations were used for 10-fold cross-validation training of the multi-view multi-task model; multi-task CNN's was trained using a randomly partitioned subset of the remaining examinations with 75-25 spliting scheme; radiological assessments, density estimates, age at examination, pathology outcomes from core biopsy or cancerous breast tissue removal were used in creating the reference standard	Malignant classification AUROC of 0.795
Wu et al. (2020b) [46]	ResNet-based 22-layer CNN; pre-trained weights from BIRADS classification; several multi-view model variations receiving four standard views as input	NYU Breast Cancer Screening Dataset (NYU-BCSD) [47]; 229,426 screening examinations in total; 985 mammograms with biopsy confirmed malignant and 5,556 with benign findings, 234 mammograms with both benign and malignant findings; image level labels	Data was split into training, validation, and testing sets with 80-10-10 splitting scheme [47]; all 4,844 biopsy confirmed examinations were used each epoch and among the examinations not having biopsy taken randomly sampled subset of 4,844 examinations were drawn each epoch during model training; pathology reports were used in creating the reference standard	Malignant classification (view-wise model) AUROC of 0.843 for biopsied subpopulation and 0.886 for screening population; benign classification (view-wise model) AUROC of 0.690 for biopsied subpopulation and AUROC of 0.747 for screening population; ensembling can boost the performance
Wu et al. (2020a) [45]	ResNet-based 22-layer CNN; several dual-view model variations used in conjunction with different regularization methods receiving CC and MLO views as input	NYU Breast Cancer Screening Dataset (NYU-BCSD) [47]; 229,426 screening examinations in total; 985 mammograms with biopsy confirmed malignant and 5,556 with benign findings; image level labels; CC and MLO views share the same label	Training procedure similar to Wu et al. (2020b); pathology reports were used in creating the reference standard	Classification with weight sharing model AUROC of 0.879 for the test set; ensembling can boost the performance
Chen et al. (2022) [7]	EfficientNet-b0-based CNN; pre-trained weights from ImageNet classification; dedicated global consistency, local co-occurence and fusion modules trained end-to-end	Annotated Digital Mammograms and Associated Non-Imaging data (ADMANI) datasets originating from several screening centers [10]; 139,034 examinations (ADMANI1) with 5,901 cancerous cases having malignant findings for training; 1,691,654 examinations (ADMANI2) with 5,232 cancerous cases for evaluation; image level labels	Data was split into training, validation, and testing sets with 80-10-10 splitting scheme for training on ADMANI1, additional evaluation on ADMANI2; pathology outcomes from biopsy and surgery, or reporting of interval cancer were used in creating the reference standard	Malignant classification AUROC of 0.926 on ADMANI2
Isosalo et al. (current study)	ResNet-based 22-layer CNN; pre-trained weights from NYU; multi-view model receiving four standard views as input	Oulu Dataset of Screening Mammography originating from a single screening center; 49,634 screening examinations in total; 11,037 screening examinations curated for experiments; image level labels	Data was split into training and holdout sets with 80-20 splitting scheme; 5-fold cross-validation was used for finetuning/training the model; 8,829 examinations with 362 examinations with malignant findings, 4,146 with benign findings (76 examinations having also malignant findings) for training; holdout set of 2,208 examinations with 70 malignant examinations, 1,056 examinations with benign findings (16 examinations having also malignant findings) for evaluation; mammographic information system data including screening assessments, patient age, and possible confirmation study results, such as additional imaging results and histology responses, were used in creating the reference standard	Malignant classification AUROC values of 0.82, 0.84, 0.85, and 0.83 for the holdout set for R-MLO, L-MLO, R-CC and L-CC views respectively; benign classification AUROC values of 0.67, 0.66, 0.64, and 0.62 for R-MLO, L-MLO, R-CC and L-CC views respectively for the holdout set

screening mammography data. Radiologically occult findings [17] present a slightly different problem. Occult findings are present, but not visible for the human annotator. Also in this work, several such cases were identified among the malignant suspect examinations. Moreover, excluding these would have resulted a significant reduction in our malignant samples. Therefore, it was decided to consistently label the views with radiologically occult lesions according to histological response retrieved from the MIS. An identical label was assigned to CC and MLO views in these cases, introducing a potential error. However, we observed this choice to have a favorable overall effect to the model training.

Large and partially uncharted area are the "normal" representing the majority of examinations and the second largest group of "benign" in the Finnish screening assessment scale. It is not uncommon that breasts assessed as "normal" contain some mammographic changes. Among these can be, for example, certain calcifications and steatonecrosis. Also, some of the clearly benign findings such as fibroadenomas, calcification which have remained unchanged for a long period of time can end up assigned as "normal" in the screening. Several benign findings, *e.g.*, circumscribed masses with macrocalcifications, masses of fatty or mixed density, and calcification of the arteries (vascular breast calcifications), do have visually distinctive benign morphology. However, for example, benign cysts and lymph nodes have visual characteristics to those considered as malignant (see for example [4]). Thus, such breast patterns can complicate the efforts of training a successful breast cancer classifier. Furthermore, based on our data, screening assessment "malignancy cannot be ruled out" in the Finnish scale is used to expresses nonspecific findings that require further assessment with other methods such as ultrasound. There is broad variation within this category, and what is more, variation which does not necessarily have a common denominator. Radiological and pathological information should nonetheless be in concordance. Moreover, breasts that end up having benign finding in the follow-up examination, assessed as "malignancy cannot be ruled out" during screening, have visual characteristics pointing towards malignant. Breasts receiving assessment "highly suspicious of malignancy" or "malignant" in the Finnish scale are clear to decide upon for a human eye. Therefore, as an implication, we see that category "malignancy cannot be ruled out" representatives having nonspecific findings should be looked in detail in further studies.

4.3. Training considerations

Favorable model initialization can be seen important in training CNN's. Wu et al. [46] reused model parts from pre-training on BIRADS classification, a task similar to the one described in Geras et al. [13]. The authors justify this with small amount (N = 5,832) of biopsied samples. Their pre-training on BIRADS resulted in more robust model and in the end better classification performance than training from random initialization. In our study, we had far fewer samples with pathological-anatomical diagnosis (N=1,012), and thus when training from randomly initialized weights the performance remained notably low. Moreover, in this type of studies, we should look closely also at the underlying breast cancer subtypes, parenchymal patterns (see for example Li et al. [26]). In our case, among our 1,012 examinations

with pathological-anatomical diagnosis there were 63 different subtypes and furthermore their visual manifestations in the mammograms. With some subtypes this leaves us with only few samples to learn from. Therefore, initializing model weights from a strong pre-trained model is preferred. Moreover, as depicted in the results presented in Section 3.3, training from random initialization results in suboptimal performance also in our case.

Alongside with suitable model initialization, optimization method can have a key role. The quasi-hyperbolic Adam variant, proposed by Ma et al. [30], was chosen as an optimizer in our implementation because of its appealing properties related to improved stability of the training and potential improvement on the generalization ability of the model. It would be of interest to invest time searching optimal parameters for the optimizer and experiment with $\nu_2 < 1$, which might further improve stability of the training [30] and allow the reduction of training time with potentially better convergence. There was some indication that multi-view model finetuning might benefit from QHAdam in comparison to Adam (data not shown).

Multi-task learning can be hypothesized to induce learning bias [6] and therefore steer the model to learn more meaningful representations. In essence this happens through learning more versatile patterns and therefore achieve better generalization to unseen data. Wu et al. [46, 45] regard the multi-task learning scenario of predicting both malignancy and benignity of the sample having also an important regularizing effect.

Another way of regularizing a model training is to include augmentations. In their original work, Wu et al. [46] resorted only to augmentations which modify the image size and crop location. We added Cutout augmentation by DeVries and Taylor [8], to steer the model to take into account also subtle signals instead of building on the presence of the strongest visual cues for classification. As a future work, the method could benefit from additional augmentations. With high resolution medical images training this model already reserves a significant amount of computational resources. Therefore, we did not experiment any additional augmentations at this point.

4.4. Dataset sampling considerations

With the multi-view setting it is difficult to fully balance the dataset when it comes to lesions appearing in just some of the projections. To achieve a good side-wise balance with good number of representatives of different breast cancer subtypes within cross-validation fold would require a very large dataset, larger than what we have in our disposal.

In order to train an unbiased model, the model should be taught with data representing all different patterns appearing in screening examinations. Therefore, bias is introduced by excluding large amounts of data. The largest exclusions in our work concern the group of examinations with irregular screening interval and missing later label preserving follow-up. The requirement for irregular screening interval can be relaxed, but the requirement for the follow-up is considered important requirement to have.

4.5. Future work

Budd et al. [5] have hypothesized that for a clinically credible artificial intelligence active collaboration with human and machine might be required. Yan et al. have proposed an active learning (AL) model for breast cancer evaluation in dual-view mammogram setting [48], highly compatible with the model studied in this work. Their model utilized the expected consistency between the model predictions to MLO and CC views as AL criteria. If we could devise a fix to the issues arising from the fact that there is only partial tissue overlap between MLO and CC views, there would be a mechanism for annotation efficient training. Finally, when AL is coupled with a well-performing pre-trained model such as the Wu et al. [46], efficient model finetuning in terms of manual annotations could be possible also for a smaller screening centers. Moreover, AL may play an important role in the future in identifying the challenging subgroups and other ambiguities of the screening mammography datasets.

5. Conclusions

In this retrospective study we adopted an openly available multi-view deep learning model and performed an independent evaluation to assess its performance using Finnish full-field digital mammography data. Our aim was to acquire understanding on what kind of size and richness is required from a dataset for a successful finetuning, and whether differing demographics pose a significant hindrance for the process. Among the identified challenges were issues arising from the medical practice and issues related to modelling. One of the medical practise related issues arise from the broad definition of the benign class in the Finnish screening system, resulting a large within-class variation, thus making it challenging to retain the performance reported for the evaluated model in the literature. Moreover, a difficult subcategory to model were the malignant suspects with nonspecific findings. Therefore, we propose that the future research should concentrate on examining the model performance to identify those breast cancer subgroups adversely affecting to the results. This is one key requirement for increasing the models readiness level for a clinical setting after first performing a finetuning with a rich enough dataset, locally or via collaborative effort. Having a strong pre-trained model, which performs well also in an out-of-distribution setting, can be seen crucial when operating with single center datasets. Thus far, training such a well generalizable model has required a sizable dataset.

Conflict of interest statement

Authors declare no conflicts of interest.

Data statement

The datasets generated and analyzed during the current study are not publicly available due to personal data content.

The INbreast data that support the findings of this study are available from Breast Research Group, INESC Porto, Portugal, upon reasonable request.

The tool used in facilitating the reference annotations for this study is made available at https://github.com/MIPT-Oulu/MammogramAnnotationTool_public/.

Funding statement

The funding sources had no involvement in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Acknowledgement

Tiina Ihme is acknowledged for her contributions in research project management. Helinä Heino is acknowledged for her contribution in the data collection. Jungkyu Park, Nan Wu, and Krzysztof J. Geras are acknowledged for their help in deploying the New York University pre-trained models. Inês Domingues is acknowledged for providing the INbreast database.

CRediT authorship contribution statement

A. Isosalo: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Visualization. **S.I. Inkinen:** Conceptualization, Methodology, Software, Investigation, Resources, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition. **T. Turunen:** Investigation, Ground Truth Annotations, Writing - Review & Editing. **P.S. Ipatti:** Conceptualization, Writing - Review & Editing, Resources, Supervision. **J. Reponen:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision. **M.T. Nieminen:** Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition.

A. Supplementary evaluation metrics

This section provides Youden index, Sensitivity, Specificity, F1 score, Precision and Recall values for malignant (Table A.1) and benign (Table A.2) classification tasks for different datasets.

References

[1] Agarwal, R., Diaz, O., Lladó, X., Yap, M.H., Martí, R., 2019. Automatic mass detection in mammograms using deep convolutional neural networks. Journal of Medical Imaging 6, 1. doi:10.1117/1.JMI.6.3.031409.

[2] Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., Barkan, E., Herzel, E., Naor, S., Karavani, E., Koren, G., Goldschmidt, Y., Shalev, V., Rosen-Zvi, M., Guindy, M., 2019. Predicting breast cancer by applying deep learning to linked health records and mammograms. Radiology 292, 331–342. doi:https://doi.org/10.1148/radiol.2019182622.

[3] Bassett, L.W., 1994. Quality determinants of mammography. 95, United States Government Printing.

[4] Berg, W.A., Sechtin, A.G., Marques, H., Zhang, Z., 2010. Cystic Breast Masses and the ACRIN 6666 Experience. Radiologic Clinics of North America 48, 931–987. doi:10.1016/j.rcl.2010.06.007.

[5] Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis 71, 102062. doi:https://doi.org/10.1016/j.media.2021.102062.

[6] Caruana, R., 1997. Multitask Learning. Machine Learning doi:10.1023/A:1007379606734.

[7] Chen, Y., Wang, H., Wang, C., Tian, Y., Liu, F., Liu, Y., Elliott, M., McCarthy, D.J., Frazer, H., Carneiro, G., 2022. Multi-view local co-occurrence and global consistency learning improve mammogram classification generalisation, in: Medical Image Computing and Computer Assisted Intervention: MICCAI 2022. Lecture Notes in Computer Science., Springer. pp. 3–13. doi:https://doi.org/10.1007/978-3-031-16437-8_1.

[8] DeVries, T., Taylor, G.W., 2017. Improved Regularization of Convolutional Neural Networks with Cutout. arXiv:1708.04552 URL: http://arxiv.org/abs/1708.04552, arXiv:1708.04552.

[9] Doi, K., Giger, M.L., Nishikawa, R.M., Hoffmann, K.R., Macmahon, H., Schmidt, R.A., Chua, K.G., 1993. Digital radiography: A useful clinical tool for computer-aided diagnosis by quantitative analysis of radiographic images. Acta Radiologica 35, 426–439. doi:10.3109/02841859309175379.

[10] Frazer, H.M., Tang, J.S., Elliott, M.S., Kunicki, K.M., Hill, B., Karthik, R., Kwok, C.F., Peña-Solorzano, C.A., Chen, Y., Wang, C., et al., 2022. Admani: Annotated digital mammograms and associated non-image datasets. Radiology: Artificial Intelligence 5, e220072. doi:https://doi.org/10.1148/ryai.220072.

[11] Gao, Y., Geras, K.J., Lewin, A.A., Moy, L., 2019. New Frontiers: An Update on Computer-Aided Diagnosis for Breast Imaging in the Age of Artificial Intelligence. American Journal of Roentgenology 212, 300–307. doi:10.2214/AJR.18.20392.

[12] Geras, K.J., Mann, R.M., Moy, L., 2019. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. Radiology 293, 246–259. doi:10.1148/radiol.2019182627.

[13] Geras, K.J., Wolfson, S., Shen, Y., Wu, N., Kim, S.G., Kim, E., Heacock, L., Parikh, U., Moy, L., Cho, K., 2017. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. arXiv:1703.07047 URL: http://arxiv.org/abs/1703.07047, arXiv:1703.07047.

[14] Gilbert, F., Tucker, L., Gillan, M., Willsher, P., Cooke, J., Duncan, K., Michell, M., Dobson, H., Lim, Y., Purushothaman, H., Strudley, C., Astley, S., Morrish, O., Young, K., Duffy, S., 2015. The TOMMY trial: a comparison of tomosynthesis with digital mammography in the UK NHS breast screening programme–a multicentre retrospective reading study comparing the diagnostic performance of digital breast tomosynthesis and digital mammography with digital mammography alone. Health Technology Assessment 19, 1–136. doi:https://doi.org/10.3310/hta19040.

[15] Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R., 2018. Deep learning in mammography and breast histology, an overview and future trends. Medical Image Analysis 47, 45–67. doi:10.1016/j.media.2018.03.006.

[16] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 770–778. doi:10.1109/CVPR.2016.90, arXiv:1512.03385.

[17] Holland, R., Hendriks, J., Mravunac, M., 1983. Mammographically occult breast cancer: a pathologic and radiologic study. Cancer 52, 1810–1819. doi:https://doi.org/10.1002/1097-0142(19831115)52:10<1810::AID-CNCR2820521009>3.0.CO;2-F.

[18] Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning, ICML 2015 37, 448–456. doi:10.5555/3045118.3045167.

[19] Iqbal, H., 2018. PlotNeuralNet: Latex code for making neural networks diagrams. URL: https://github.com/HarisIqbal88/PlotNeuralNet, doi:110.5281/zenodo.2526396.

[20] Isosalo, A., Heino, H., Inkinen, S.I., Nieminen, M.T., 2021. Mammogram annotation tool. GitHub, 14 Sep 2021. URL: https://github.com/MIPT-Oulu/MammogramAnnotationTool_public. (Accessed: 13 Sep 2022).

[21] Isosalo, A., Mustonen, H., Turunen, T., Ipatti, P.S., Reponen, J., Nieminen, M.T., Inkinen, S.I., 2022. Evaluation of different convolutional neural network encoder-decoder architectures for breast mass segmentation, in: Deserno, T.M., Park, B.J. (Eds.), Medical Imaging 2022: Imaging Informatics for Healthcare, Research, and Applications, International Society for Optics and Photonics. SPIE. pp. 207–214. doi:10.1117/12.2628190.

[22] Kervadec, H., Dolz, J., Tang, M., Granger, E., Boykov, Y., Ayed, I.B., 2019. Constrained-cnn losses for weakly supervised segmentation. Medical image analysis 54, 88–99.

[23] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. p. 9. doi:https://doi.org/10.1145/3065386.

[24] Kyono, T., Gilbert, F.J., van der Schaar, M., 2019. Multi-view multi-task learning for improving autonomous mammogram diagnosis, in: Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (Eds.), Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR. pp. 571–591. URL: https://proceedings.mlr.press/v106/kyono19a.html.

[25] Lehman, C.D., Arao, R.F., Sprague, B.L., Lee, J.M., Buist, D.S., Kerlikowske, K., Henderson, L.M., Onega, T., Tosteson, A.N., Rauscher, G.H., Miglioretti, D.L., 2017. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology 283, 49–58. doi:https://doi.org/10.1148/radiol.2016161174.

[26] Li, H., Mendel, K.R., Lan, L., Sheth, D., Giger, M.L., 2019. Digital mammography in breast cancer: additive value of radiomics of breast parenchyma. Radiology 291, 15. doi:https://doi.org/10.1148/radiol.2019181113.

[27] Liberman, L., Menell, J.H., 2002. Breast imaging reporting and data system (bi-rads). Radiologic Clinics of North America 40, 409–430. doi:https://doi.org/10.1016/S0033-8389(01)00017-3.

[28] Lipton, Z.C., Elkan, C., Naryanaswamy, B., 2014. Optimal thresholding of classifiers to maximize f1 measure, in: Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2014. Lecture Notes in Computer Science., Springer. Springer. pp. 225–239. doi:https://doi.org/10.1007/978-3-662-44851-9_15.

[29] Liu, K., Shen, Y., Wu, N., Chłędowski, J.P., Fernandez-Granda, C., Geras, K.J., 2021. Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis, in: Heinrich, M., Dou, Q., de Bruijne, M., Lellmann, J., Schläfer, A., Ernst, F. (Eds.), Proceedings of the 4th Conference on Medical Imaging with Deep Learning, PMLR. pp. 451–472. URL: https://proceedings.mlr.press/v143/liu21b.html.

[30] Ma, J., Yarats, D., 2019. Quasi-hyperbolic momentum and Adam for deep learning, in: International Conference on Learning Representations (ICLR 2019), pp. 1–13. URL: https://openreview.net/forum?id=S1fUpoR5FQ.

[31] Mason, D., et al., 2020. pydicom: An open source DICOM library. URL: https://github.com/pydicom/pydicom, doi:10.5281/zenodo.3614067.

[32] Meltzer, C., Skaane, P., 2022. Mammography screening, in: Breast Imaging. Springer, pp. 43–68. doi:10.1007/978-3-030-94918-1_3.

[33] Mongan, J., Moy, L., Kahn Jr, C.E., 2020. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. Radiology. Artificial Intelligence 2. doi:https://pubs.rsna.org/doi/10.1148/ryai.2020200029.

[34] Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. INbreast: Toward a full-field digital mammographic database. Academic Radiology 19, 236–248. doi:10.1016/j.acra.2011.09.014.

[35] Nodine, C.F., Kundel, H.L., Mello-Thoms, C., Weinstein, S.P., Orel, S.G., Sullivan, D.C., Conant, E.F., 1999. How experience and training influence mammography expertise. Academic Radiology 6, 575–585. doi:10.1016/S1076-6332(99)80252-9.

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. p. 12. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay, 2011. Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html, doi:10.5281/zenodo.3696718.

[38] Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J., 2021. Better Aggregation in Test-Time Augmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE. pp. 1214–1223. doi:https://doi.org/10.1109/ICCV48922.2021.00125.

[39] Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., Geras, K.J., 2021. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Medical Image Analysis 68, 101908. doi:10.1016/j.media.2020.101908.

[40] Sprague, B.L., Miglioretti, D.L., Lee, C.I., Perry, H., Tosteson, A.A., Kerlikowske, K., 2020. New mammography screening performance metrics based on the entire screening episode. Cancer 126, 3289–3296. doi:https://doi.org/10.1002/cncr.32939.

[41] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians 71, 209–249. doi:10.3322/caac.21660.

[42] Tardy, M., Mateus, D., 2021. Looking for abnormalities in mammograms with self-and weakly supervised reconstruction. IEEE Transactions on Medical Imaging , 1–1doi:10.1109/TMI.2021.3050040.

[43] Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P., 2020. Preparing Medical Imaging Data for Machine Learning. Radiology 295, 4–15. doi:10.1148/radiol.2020192224.

[44] Wu, N., Jastrzębski, S., Cho, K., Geras, K.J., 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks, in: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (Eds.), Proceedings of the 39th International Conference on Machine Learning, PMLR. pp. 24043–24055. URL: https://proceedings.mlr.press/v162/wu22d.html.

[45] Wu, N., Jastrzębski, S., Park, J., Moy, L., Cho, K., Geras, K., 2020a. Improving the Ability of Deep Neural Networks to Use Information from Multiple Views in Breast Cancer Screening, in: Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C. (Eds.), Proceedings of the 3rd Conference on Medical Imaging with Deep Learning, PMLR. pp. 827–842. URL: https://proceedings.mlr.press/v121/wu20a.html.

[46] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L.L.Y., Ho, K., Weinstein, J.D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K.J., 2020b. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Transactions on Medical Imaging 39, 1184–1194. doi:10.1109/TMI.2019.2945514.

[47] Wu, N., Phang, J., Park, J., Shen, Y., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K., 2019. The NYU breast cancer screening dataset v1.0. URL: https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf.

[48] Yan, Y., Conze, P.H., Lamard, M., Zhang, H., Quellec, G., Cochener, B., Coatrieux, G., 2021. Deep active learning for dual-view mammogram analysis, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 180–189. doi:10.1007/978-3-030-87589-3_19.

[49] Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? Advances in Neural Information Processing Systems 27, 9. URL: https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

[50] Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35. doi:https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Table A.1
Malignant classification for the three datasets using pre-trained weights from NYU without finetuning, pre-trained weights from NYU with finetuning, and trained from random initialization

Dataset, model and projection	Youden index (threshold)	Sensitivity	Specificity	F1 score (threshold)	Precision	Recall
Annotated malignant suspect holdout subset (S1)						
Baseline w/o finetuning						
R-MLO	0.51 (0.03)	0.75	0.77	0.38 (0.19)	0.29	0.55
R-CC	0.57 (0.02)	0.79	0.78	0.51 (0.25)	0.49	0.53
L-MLO	0.45 (0.06)	0.59	0.86	0.46 (0.23)	0.59	0.37
L-CC	0.52 (0.02)	0.75	0.78	0.43 (0.22)	0.36	0.53
Finetuned						
R-MLO	0.49 (0.15)	0.73	0.77	0.40 (0.20)	0.42	0.37
R-CC	0.59 (0.11)	0.82	0.77	0.52 (0.26)	0.57	0.47
L-MLO	0.54 (0.11)	0.80	0.73	0.47 (0.23)	0.43	0.51
L-CC	0.56 (0.16)	0.73	0.83	0.49 (0.25)	0.52	0.47
Random initialization						
R-MLO	0.28 (0.11)	0.86	0.41	0.21 (0.10)	0.13	0.49
R-CC	0.30 (0.12)	0.63	0.67	0.26 (0.13)	0.17	0.56
L-MLO	0.21 (0.13)	0.69	0.52	0.20 (0.10)	0.12	0.69
L-CC	0.29 (0.14)	0.59	0.70	0.24 (0.12)	0.15	0.51
Entire holdout subset (S2)						
Baseline w/o finetuning						
R-MLO	0.49 (0.03)	0.69	0.79	0.20 (0.10)	0.18	0.22
R-CC	0.56 (0.02)	0.75	0.81	0.32 (0.16)	0.32	0.32
L-MLO	0.47 (0.06)	0.58	0.89	0.33 (0.17)	0.31	0.36
L-CC	0.53 (0.02)	0.78	0.75	0.28 (0.14)	0.25	0.33
Finetuned						
R-MLO	0.51 (0.13)	0.71	0.80	0.27 (0.13)	0.28	0.25
R-CC	0.58 (0.13)	0.77	0.81	0.38 (0.19)	0.46	0.32
L-MLO	0.58 (0.11)	0.82	0.77	0.30 (0.15)	0.41	0.24
L-CC	0.54 (0.16)	0.67	0.87	0.36 (0.18)	0.44	0.31
Random initialization						
R-MLO	0.39 (0.11)	0.85	0.54	0.09 (0.05)	0.06	0.31
R-CC	0.29 (0.12)	0.63	0.66	0.11 (0.05)	0.10	0.12
L-MLO	0.28 (0.13)	0.67	0.60	0.08 (0.04)	0.05	0.36
L-CC	0.26 (0.14)	0.56	0.69	0.09 (0.05)	0.30	0.05
Portuguese evaluation data (S3)						
Baseline w/o finetuning						
R-MLO	0.68 (0.03)	0.88	0.80	0.68 (0.34)	0.58	0.82
R-CC	0.46 (0.07)	0.69	0.77	0.53 (0.27)	0.57	0.50
L-MLO	0.75 (0.03)	0.90	0.85	0.67 (0.33)	0.53	0.90
L-CC	0.81 (0.05)	0.90	0.91	0.83 (0.41)	0.81	0.85
Finetuned						
R-MLO	0.62 (0.17)	0.76	0.86	0.65 (0.32)	0.57	0.76
R-CC	0.47 (0.20)	0.69	0.79	0.54 (0.27)	0.70	0.44
L-MLO	0.57 (0.05)	0.90	0.67	0.65 (0.32)	0.79	0.55
L-CC	0.73 (0.20)	0.90	0.83	0.77 (0.38)	0.79	0.75

Table A.2
Benign classification for the three datasets using pre-trained weights from NYU without finetuning, pre-trained weights from NYU with finetuning, and trained from random initialization

Dataset, model and projection	Youden index (threshold)	Sensitivity	Specificity	F1 score (threshold)	Precision	Recall
Annotated malignant suspect holdout subset (S1)						
Baseline						
R-MLO	0.27 (0.21)	0.75	0.53	0.48 (0.24)	0.36	0.76
R-CC	0.24 (0.12)	0.89	0.35	0.49 (0.24)	0.34	0.86
L-MLO	0.34 (0.24)	0.76	0.58	0.52 (0.26)	0.40	0.73
L-CC	0.19 (0.29)	0.40	0.79	0.50 (0.25)	0.35	0.86
Finetuned						
R-MLO	0.29 (0.26)	0.71	0.58	0.49 (0.24)	0.38	0.69
R-CC	0.08 (0.25)	0.79	0.28	0.44 (0.22)	0.28	1.00
L-MLO	0.34 (0.30)	0.66	0.68	0.52 (0.26)	0.42	0.70
L-CC	0.11 (0.25)	0.79	0.32	0.48 (0.24)	0.31	0.98
Random initialization						
R-MLO	0.01 (0.23)	0.36	0.65	0.42 (0.21)	0.27	1.00
R-CC	0.02 (0.24)	0.07	0.95	0.43 (0.22)	0.28	1.00
L-MLO	0.00 (0.17)	1.00	0.00	0.43 (0.21)	0.27	1.00
L-CC	0.12 (0.25)	0.41	0.71	0.46 (0.23)	0.30	1.00
Entire holdout subset (S2)						
Baseline						
R-MLO	0.24 (0.26)	0.51	0.73	0.50 (0.25)	0.41	0.65
R-CC	0.01 (0.15)	0.55	0.46	0.50 (0.25)	0.34	1.00
L-MLO	0.20 (0.24)	0.60	0.61	0.51 (0.25)	0.36	0.88
L-CC	0.04 (0.35)	0.14	0.90	0.49 (0.25)	0.33	1.00
Finetuned						
R-MLO	0.27 (0.28)	0.56	0.71	0.51 (0.26)	0.42	0.64
R-CC	0.21 (0.27)	0.68	0.52	0.53 (0.27)	0.40	0.79
L-MLO	0.23 (0.30)	0.50	0.73	0.53 (0.26)	0.40	0.76
L-CC	0.18 (0.28)	0.56	0.62	0.51 (0.25)	0.35	0.92
Random initialization						
R-MLO	0.12 (0.23)	0.38	0.74	0.48 (0.24)	0.31	1.00
R-CC	0.07 (0.22)	0.76	0.31	0.50 (0.25)	0.34	0.98
L-MLO	0.12 (0.22)	0.70	0.42	0.50 (0.25)	0.34	0.89
L-CC	0.03 (0.25)	0.13	0.91	0.49 (0.25)	0.33	1.00
Portuguese evaluation data (S3)						
Baseline						
R-MLO	0.05 (0.01)	0.98	0.07	0.81 (0.40)	0.67	1.00
R-CC	0.07 (0.80)	0.07	1.00	0.81 (0.41)	0.69	1.00
L-MLO	0.10 (0.18)	0.66	0.44	0.74 (0.37)	0.58	1.00
L-CC	0.03 (0.01)	1.00	0.00	0.74 (0.37)	0.59	1.00
Finetuned						
R-MLO	0.08 (0.50)	0.19	0.89	0.81 (0.40)	0.69	0.98
R-CC	0.35 (0.18)	0.86	0.48	0.83 (0.42)	0.72	0.98
L-MLO	0.15 (0.20)	0.76	0.39	0.74 (0.37)	0.59	0.98
L-CC	0.30 (0.24)	0.73	0.57	0.76 (0.38)	0.66	0.90

Highlights

To reduce radiologists' overwhelming reading workload in routine mammography screening, one solution is to use computer assisted detection software for automated detection and classification of breast cancer.

Development of reliable intelligent analytics (machine learning models) in computer assisted detection requires large amount of data with rich set of examples of anomalies and phenotypes to automatically learn the required representations for the task.

With transfer learning, breast cancer classification model can be finetuned to adapt to different demographics, without the immediate need to collect or share large sets of medical data.

Finally, having a strong pre-trained model, which performs well also in an out-of-distribution setting, can be seen crucial when operating with single center datasets.

Biography

Antti Isosalo received his MSc (Tech.) degree in information engineering (2010) from the University of Oulu, Oulu, Finland. In addition, he has conducted supplementary studies in mathematics and physics and the pedagogical studies for subject teachers. He is currently pursuing towards PhD degree in medical technology at the University of Oulu, Oulu, Finland. His most recent research interests lie in the area of deep learning and medical image analysis.

Satu I. Inkinen received her MSc (2013) and PhD (2017) degrees in applied physics from the University of Eastern Finland, Kuopio, Finland. After graduation, she was a postdoctoral researcher at the Research Unit of Medical Imaging, Physics and Technology at the Faculty of Medicine, University of Oulu, Oulu, Finland (2017-2021). Her research interests cover photon counting computed tomography and deep learning in medical imaging applications. Currently, she works as a medical physicist resident at the Helsinki University Hospital, Helsinki, Finland.

Topi Turunen received his licentiate degree in medicine (2015) and his specialist degree in radiology (2022) from the University of Oulu, Oulu, Finland. He has conducted studies also in Biochemistry. He is currently pursuing towards a subspecialization in emergency radiology where also his main professional interests lie.

Pieta S. Ipatti received her licentiate degree in medicine (2007), her specialist degree in radiology (2015), and subspeciality in breast radiology (2017) from the University of Oulu, Oulu, Finland. In 2016-2017 she served three months in Addenbrooke's Hospital in Cambridge, U.K., as a breast radiology fellow part of ESOR exchange programme. Currently, she serves as a deputy chief medical officer at the Oulu University Hospital, Oulu, Finland.

Jarmo Reponen received his licentiate degree in medicine (1985), medical specialist degree in radiology (1992) and MD PhD degree (2010) from the University of Oulu, Oulu, Finland. Currently, he is a professor of healthcare information systems at the Research Unit of Health Sciences and Technology at the Faculty of Medicine, University of Oulu, Oulu, Finland. Prof. Reponen has more than 30 years of experience in radiology information system and electronic patient record system development. He has carried out several telemedicine and digital radiology projects. Reponen holds a Finnish special competence in mammography screening and healthcare information technology. He is the recipient of the Carl Wegelius Award of the Radiological Society of Finland (2016) and the Lifetime Achievement Award of the International Society for Telemedicine and eHealth (2021). He has contributed to research in teleradiology, telemedicine, AI and machine learning in radiology, mobile communication, and patient record systems.

Miika T. Nieminen received his MSc (1999) and PhD (2002) degrees in medical physics from the University of Kuopio, Kuopio, Finland. After graduation, he was a postdoctoral research fellow at Harvard University. Currently, he is a professor of medical physics at the Research Unit of Health Sciences and Technology at the Faculty of Medicine, University of Oulu, Oulu, Finland and he serves as a chief physicist at the Oulu University Hospital, Oulu, Finland. Prof. Nieminen has actively published on different aspects of medical imaging and physics. His principal research interest is the development of novel imaging techniques for tissue characterization. He is internationally recognized as one of the key researchers in the field of quantitative magnetic resonance imaging of osteoarthritis. His other research interests include novel computed tomographic methods, dosimetry, radiation safety and quality assurance in radiology. He has received the Carl Wegelius Award of the Radiological Society of Finland (2012) and the Fellowship Award of the International Society for Magnetic Resonance in Medicine (2015).

Conflict of Interest Statement

Authors declare no conflicts of interest.