Medical Image Analysis 82 (2022) 102618

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Beyond fine-tuning: Classifying high resolution mammograms using function-preserving transformations

Tao Wei [a], Angelica I. Aviles-Rivero [b,*], Shuo Wang [c,d], Yuan Huang [e], Fiona J. Gilbert [e], Carola-Bibiane Schönlieb [b], Chang Wen Chen [a]

[a] The Department of Computer Science, State University of New York at Buffalo, NY, USA
[b] The Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
[c] The Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai, China
[d] Shanghai Key Laboratory of MICCAI, Shanghai, China
[e] The Department of Radiology, University of Cambridge, UK

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Deep learning
Transfer learning
Mammogram classification
High resolution
Network morphism
Function-preserving transformations

ABSTRACT

The task of classifying mammograms is very challenging because the lesion is usually small in the high resolution image. The current state-of-the-art approaches for medical image classification rely on using the de-facto method for convolutional neural networks-fine-tuning. However, there are fundamental differences between natural images and medical images, which based on existing evidence from the literature, limits the overall performance gain when designed with algorithmic approaches. In this paper, we propose to go beyond fine-tuning by introducing a novel framework called MorphHR, in which we highlight a new transfer learning scheme. The idea behind the proposed framework is to integrate function-preserving transformations, for any continuous non-linear activation neurons, to internally regularise the network for improving mammograms classification. The proposed solution offers two major advantages over the existing techniques. Firstly and unlike fine-tuning, the proposed approach allows for modifying not only the last few layers but also several of the first ones on a deep ConvNet. By doing this, we can design the network front to be suitable for learning domain specific features. Secondly, the proposed scheme is scalable to hardware. Therefore, one can fit high resolution images on standard GPU memory. We show that by using high resolution images, one prevents losing relevant information. We demonstrate, through numerical and visual experiments, that the proposed approach yields to a significant improvement in the classification performance over state-of-the-art techniques, and is indeed on a par with radiology experts. Moreover and for generalisation purposes, we show the effectiveness of the proposed learning scheme on another large dataset, the ChestX-ray14, surpassing current state-of-the-art techniques.

Screening mammography is the primary imaging test for early detection of breast cancer as it offers a reliable and reproducible test for diagnosis. However, a major challenge is the interpretation of the mammogram, which requires specialised training and substantial experience by the reader (Lehman et al., 2016; Becker et al., 2017). Mammograms can be difficult to read due to high variability in the patterns and subtle appearances of small cancers. This can result in variation in performance between experts (Gilbert et al., 2006; Elmore et al., 2009; Lehman et al., 2015). As a result, it is often necessary to advocate to double reading of mammograms at the expense of increasing the cost and expert workload.

The aforementioned drawbacks have motivated the rapid development of automatic and robust algorithmic approaches to support the experts' outcome. In particular, computer-aided systems, that aim to influence the expert interpretation, have shown limited performance (Lehman et al., 2015). This limitation is mainly because traditional systems are designed using hand-crafted features (Warren and Duffy, 1995). Most recently, with the astonishing success of deep learning (DL), it has been shown that these systems can substantially improve their performance by learning features as data representatives (Hamidinekoo et al., 2018).

At the algorithmic level, breast diagnosis can be casted as a classification task, in which several developments using deep ConvNets have been reported e.g. Huynh et al. (2016), Lévy and Jain (2016), Geras et al. (2017) and Shen et al. (2019). However despite the rapid development of DL based techniques, mammography classification remains

* Corresponding author.
E-mail address: ai323@cam.ac.uk (A.I. Aviles-Rivero).

https://doi.org/10.1016/j.media.2022.102618
Received 28 July 2021; Received in revised form 3 August 2022; Accepted 2 September 2022
Available online 6 September 2022

an unsolved problem. Therefore, the question of how to improve the classification performance is of great interest from the technical and clinical points of view and is the problem that we address in this work.

There have been different attempts to design end-to-end solutions, in which the majority are based on pre-trained and fine-tuned (Krizhevsky et al., 2012; He et al., 2016) using for example ResNet (He et al., 2016) type architectures. However, one of the major challenges in mammography classification is that *mammograms are high resolved in nature* and *the abnormality is significantly smaller than the whole mammogram image*, for example, 100×100 vs. $3k \times 5k$. The majority of existing algorithmic approaches are usually not scalable to hardware (doubling the image resolution would require 4x the computational and memory costs). They cannot directly handle the high resolution nature of medical images and are designed to work around the limits of the GPU memory by either: (i) decreasing significantly the image resolution or (ii) using image patches for the classifier. Although these two alternatives have shown promising results, they are not ideal as they compromise the clinical outcome. When the resolution is decreased, the negative effects are reflected at the level of losing relevant small clinical regions; for example, the structures might appear blurred. Whilst when a patch-based solution is designed, it might suffer from significant boundary artefacts (Innamorati et al., 2018).

In this work, we propose a novel learning scheme, that we called MorphHR, that addresses the current drawbacks in the existing literature. We show that the proposed approach significantly boosts the classification accuracy whilst demanding low GPU usage. Our main contributions are as follows:

- We reveal that resolution is critical for mammogram classification. It is intuitive that mammograms are high resolved (~4K) in nature. However, due to the limit of GPUs memory, current mainstream approaches either decrease the resolution or extract patches. We argue that the importance of resolution for medical image classification should be emphasised. In particular, this work is proposed to resolve this resolution issue.
- We propose a new learning scheme, which uses function-preserving transformation to regularise the network. In particular, we build the proposed framework using Network Morphism (Wei et al., 2016) principles as proxy task for transferring knowledge from different domains, specially, from low resolution natural images to high resolution mammograms. This leads to the next advantages.
  - By introducing this proxy task in the training, we can go beyond fine-tuning by modifying not only the last layers but also the first few ones of the deep Net. This will cater to its own domain feature learning.
  - The proposed training scheme is *scalable* to hardware and allows for using high resolution mammograms that fit standard GPU memory. This will avoid losing relevant clinical regions — as the abnormalities are very small.
- We extensively evaluate the proposed approach on the CBIS-DDSM (Lee et al., 2017) dataset using an extensive numerical and visual experiments. Moreover and for generalisation purposes, we show how the transfer knowledge can be effective on other large datasets such as the ChestX-ray14 (Wang et al., 2017).
- We show that the proposed learning scheme mitigates the current limitations of the body of literature, by outperforming current state-of-the-art techniques in mammography, which is on a par with radiology experts, and X-ray classification.

1. Related work

The problem of classifying mammography data has been widely investigated in the community, in which the solutions are based on using hand-crafted or automatic selected features. In this section, we review the body of literature in turn. We then remark the current drawbacks and motivate the proposed novel learning scheme.

There have been different attempts to deal with the mammography classification task. Early developments were limited by their own construction, as they were based on the use of hand-crafted features such as texture analysis or intensity-based reasoning e.g. Brzakovic et al. (1990), Petrosian et al. (1994) and Vyborny and Giger (1994). These algorithmic approaches were based on strict modelling hypothesis such as conditioning the intensity histogram e.g. Brzakovic et al. (1990) - therefore, parameters adjustments were necessary for every single image. Therefore, these methodologies were not robust and generalisable to sightly changes across them.

The previous drawbacks were mitigated with the remarkable success of deep learning in computer vision; in tasks such as object recognition and detection (Krizhevsky et al., 2012; He et al., 2016). This has motivated the community to apply successfully DL techniques to medical data. However, this type of data has set new challenges as significantly different from natural images. For example, the low signal-to-noise ratios or small relevant anatomical structures need to be taken into account to avoid false positive or false negative outcomes.

In particular, for the task of mammography data classification, there have been different works reported. The mainstream approach for this problem has been the use of patch-based classifiers e.g. Xi et al. (2018), Rampun et al. (2018), Mercan et al. (2017), Agarwal et al. (2019), Chun-ming et al. (2019), Ragab et al. (2019) and Wu et al. (2019). The central idea, of this perspective, is to decompose the mammograms in image patches which reflect: (i) areas with abnormalities (positive patches) and (ii) normal regions (negative patches). Although these algorithmic approaches have reported promising results and their merits have been recognised in the literature, when a patch-based approach is applied to a whole image, one can observe significant boundary artefacts (Innamorati et al., 2018).

A key reference in this category is the work of Li Shen (Shen, 2017; Shen et al., 2019), which to the best of our knowledge holds the state-of-the-art results on the CBIS-DDSM dataset (Lee et al., 2017). The authors' central idea is to use an all convolutional design to convert a patch-based classifier to an image-based one. However, the approach is not scalable to the image input size. That is — when the image size doubles both the computational run-time and memory shall require 4x resources. The computational resources will then overflow and reach their capacity. Further, a pre-trained patch-classifier heavily rely on having region-of-interests (ROIs) annotations, which is expensive and laborious to create.

Another set of algorithmic approaches have addressed the mammography problem as the task of detecting (Akselrod-Ballin et al., 2016; Ribli et al., 2018; Agarwal et al., 2019) or segmenting (Ronneberger et al., 2015; Sun et al., 2018) the abnormal regions. However, the main drawback of these perspectives is that they require finer annotations including ROI boxes or contours (Shen et al., 2019). These annotations are expensive to collect and are generally unavailable in the datasets.

Existing High-Resolution Techniques & Comparison to our Work. Another set of works have explored the task of classification using high-resolved mammographs. For example, the work of Yala et al. (2019) used high resolution mammographs on a ResNet-18. Unlike that work that uses a plain ResNet, we introduce a new learning scheme for any network to handle the high-resolution nature of the data. Moreover, in our results we show that our scheme also offers better performance in comparison with a plain ResNet-18. In most recent works, the authors of Shen et al. (2021) proposed a globally-aware instance classifier along with a fusion module. That work also uses a ResNet-type architecture as a backbone. In contrast to that work, we do not add additional mechanisms to improve the performance. We instead integrate function-preserving transformations directly to any network front allowing it to be scalable to standard hardware. The work of Tardy and Mateus (2021) also uses high-resolution mammographs, but unlike our work, they uses a U-net as backbone with a ResNet-22

and for the setting of self-and weakly supervised learning. We underline that direct performance against some of those techniques is not feasible due to either the learning paradigm that they use or/and the additional mechanism that they use. A commonality of existing works is the use of a ResNet-type network as backbone along with additional mechanisms. Therefore, our work can be seen as a complement to those techniques that can adopt our strategy to improve their backbone.

Goal of This Work. Although, the aforementioned approaches have reported promising results, they are limited by their own construction. The limitation comes from a commonality: the use of fine-tuning. That is, a DL model is first trained on a large dataset, such as ImageNet (Deng et al., 2009), and then it is fine-tuned for another task. However, the limitation of fine-tuning is that one can only modify the last several layers in a deep net. It is then not possible to alter the front layers. This limitation has motivated our current work. We propose a learning scheme that allows modifying the first several layers of any deep net, this, with the goal of *transferring knowledge from natural images to high resolution medical images without requiring high GPU memory*.

The proposed framework is inspired by the ideas of using function-preserving transformations, i.e., Network Morphism (Wei et al., 2016), in which one seeks to transfer the deep nets knowledge effectively. However, we further clarify the difference between that work and ours. Firstly, we use the principles from Wei et al. (2016) as proxy task as part of the proposed framework. Secondly, unlike (Wei et al., 2016), our goal is to improve the mammograms classification performance. Thirdly, we carefully design the morphing operations and strides to handle the resolution problem for mammogram classification. While in Wei et al. (2016), the authors carry out the morphing operations on the same level without resolution changes.

2. Proposed approach

This section contains two main parts: (i) the differences between fine-tuning and our philosophy and (ii) the proposed framework for classifying mammography data. In what follows, we first start formalising our problem.

Problem Definition Giving a training set in the form of pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, with input-target attributes x_i and y_i correspondingly. We seek to find a optimal classifier $f : X \rightarrow Y$ that maps the input X to the output Y space, with minimum generalisation error, such that f works well with unseen instances.

In this work, we follow the standard workflow of 'Callback' or 'Cancer-free' e.g. Yala et al. (2019). That is, cancer or not cancer. This has motivated, in extensive number of works, to consider the benign and malign cases. Whilst there are indeed works that explicitly define the three class classification e.g. Kim et al. (2020) instead of two or combining benign and healthy cases, a vast amount of technical and clinical (e.g. triages studies) works also follow such standard clinical workflow (Shen, 2017; McKinney et al., 2020; Hickman et al., 2022).

2.1. Beyond fine-tuning: Morphing deep nets for mammography data

In this section, we underline the need for improving upon fine-tuning. We highlight the major drawbacks of it when using medical images and give initial insights into the proposed approach.

Deep neural networks usually have millions of parameters and require a significant amount of data samples to train, whereas mammogram images are usually expensive to collect and the publicly available datasets are smaller than is needed for training. To address the limited-data problem, a technique to-go is fine-tuning. The key idea of it is to train a deep Net on a reasonably large dataset such as the ImageNet (Deng et al., 2009) dataset; and then the last layer (or last few layers) of the pre-trained neural network are dropped and replaced with a new layer (or several layers) to be fine-tuned for classification on a new task.

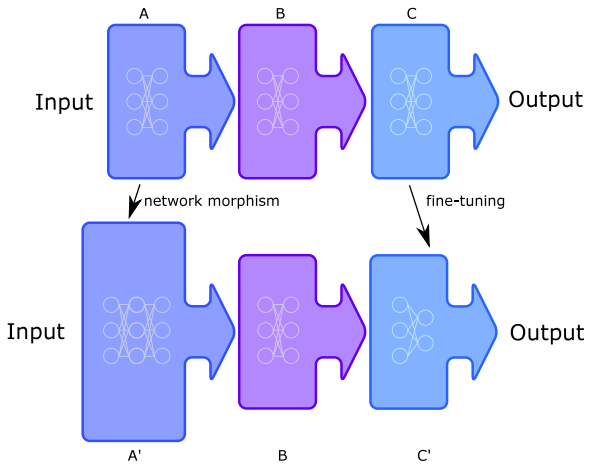


Fig. 1. Network morphism vs fine-tuning for knowledge transferring. Fine-tuning is the de-facto method for deep learning applications to medical imaging. However, it only allows to alter the last few layers ($C \rightarrow C'$) of a neural network. In this research, we propose to use network morphism to alter the first few layers ($A \rightarrow A'$) of a neural network to help to transfer knowledge from the natural images domain to the medical images domain.

Fine-tuning is the de-facto method for DL applications to medical imaging. However, there are fundamental differences in data sizes, features, and resolutions between the natural image domain and the medical image domain. Deep ConvNets are usually designed for natural image tasks. Therefore, the question of — how much of the ImageNet (or other natural images datasets) feature reuse is helpful for medical images. This question has been addressed in several works. For example, the authors of that Raghu et al. (2019) showed that fine-tuning transfer could offer little benefit to medical classification performance. In Kornblith et al. (2019), it was illustrated that pre-trained features may be less general than previously thought. Besides, fine-tuning is not scalable to the resolution of the input image. As illustrated in Fig. 1, suppose that the input image resolution is doubled, all convolutional layers of A, B, C shall require 4x GPU FLOPs and memory. Hence, current mainstream approaches based on fine-tuning, using either reduced resolutions or patches, are limited in the exploration of full high resolved mammograms.

Therefore, there is a need to mitigate the fine-tuning limitations in the medical domain, allowing to be scalable to high resolved inputs. It is important to modify not only the last layers but also the first few ones of a deep Net. This to cater for the own domain feature learning. To this aim, we propose a novel learning scheme that allows morphing deep nets without altering the function of the backbone network, it builds upon the transformation principles from Wei et al. (2016). The major differences between fine-tuning and the adopted learning scheme is illustrated in Fig. 1. In that figure, one can alter the front of a neural network from A to A'. *The first major benefit of this operation is that we can design the block A' to be suitable for learning domain specific features.*

Whilst *the second major benefit of the proposed scheme is that it allows for using high resolution mammograms that fit in a standard GPU memory.* Why does it work with highly resolved images? As illustrated in Fig. 1, suppose that the input image resolution is doubled, we are able to carefully design the morphing front A' such that B' and C' maintain their original GPU FLOPs and memory consumption. Deep ConvNets are typically designed to fit images with input size of, for example, 224×224 or 227×227. This input image size is fixed until one modifies the network architecture by: (i) replacing fully connected layers with 1×1 convolutional layers or (ii) replacing the last pooling layer with a global/adaptive pooling layer. However, these two approaches have limitations: (a) they fail if there is a single layer that requires fixed input size and it cannot be modified to accept other input sizes; and

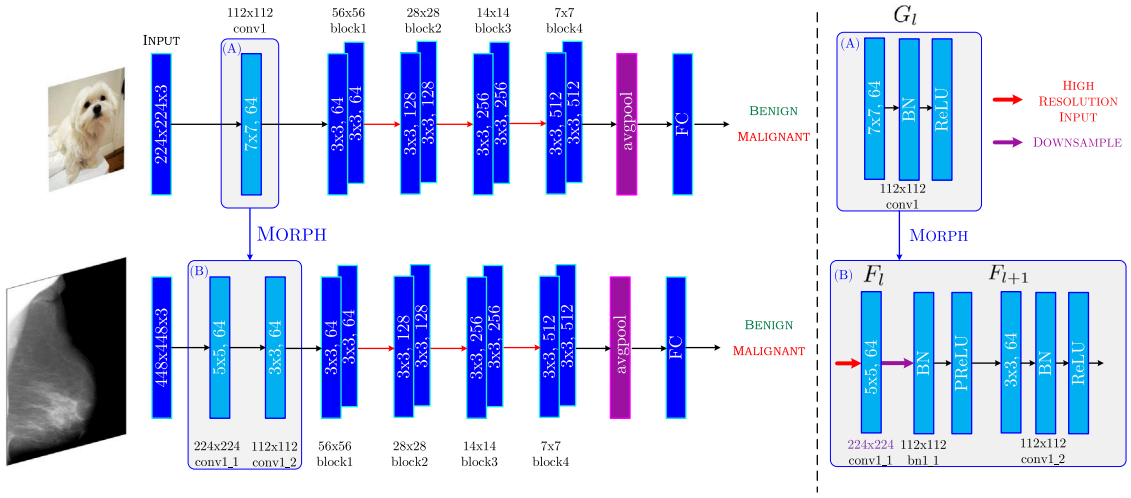


Fig. 2. Illustration of the proposed MorphHR scheme. The first convolutional layer with kernel size of 7×7 is morphed into two convolutional layers with kernel sizes of 5×5 and 3×3. The receptive field of the morphed network is kept and the network function is preserved. We double the input image size and set the stride of `conv1_1` to be 2 for downsampling. The layers in `block1` and after are exactly the same. The morphed network is further fine-tuned on the mammography data.

(b) they are not scalable, i.e., when the image size double both the computational run-time and memory shall require 4x resources.

We mitigate the existing aforementioned drawbacks by introducing a learning scheme that allows morphing deep Nets — it allows for using high resolution images whilst being scalable. More precisely, in our approach we morph the front of the network by adding more layers to allow sufficient learning of the natural-to-medical domain adaption instead of only modifying the last few layers. We hypothesise that we need more layers to learn the low-level features for a natural-to-medical domain knowledge transfer than a classical fine-tuning. Moreover, we morph the front of the network to allow high-resolution image input in a scalable way without significantly increase the GPU FLOPs and memory cost.

We therefore further underline the difference between fine-tuning and our proposed learning scheme. The proposed MorhHR does not lie in training epochs and stages. Instead, the difference is involved in the architecture change (see Fig. 1). In fact, fine-tuning only allows to alter the last few layers ($C \rightarrow C'$) of a neural network. In contrast to fine-tuning, our learning scheme allows altering the first few layers ($A \rightarrow A'$) of a neural network to help to transfer knowledge from the natural images domain to the medical images domain. We give precise details of the proposed algorithmic approach in the next section.

2.2. MorphHR: Classifying mammography data

We propose a novel framework for classifying mammography data, which is capable of transferring knowledge between natural and medical images. Moreover, it works on high resolved images which fit on standard GPU memory and is scalable. In what follows, we give details of the proposed approach, which involves function-preserving transformations.

In this work, we adopt ResNet18[1] as the backbone for our framework. However, the proposed scheme is generic and one can use other deep Nets architectures as long as the first layer is a convolutional layer. The first problem to deal with is the high resolution of the mammograms — typically 3k×5k pixels, which are significantly larger than the resolution of natural images. To mitigate this problem, we propose to morph the first layers of our backbone such that it allows for high resolved images whilst requiring standard GPU memory.

MorphHR Design. The proposed approach, MorphHR, is illustrated in Fig. 2, where we are using the ImageNet pre-trained weights to morph the network for mammograms. One can observe that the `conv1` is morphed into two convolutional layers `conv1_1` and `conv1_2`. *The added layer allows more precisely learning on the features of the mammographic images.* While primitive features, such as edges, corners, and textures, are universal for images, they can however be considerably different. By modifying the front layer, one can adapt these primitive features more efficiently. This modification on the backbone also allows the convolutional kernel size of `conv1_1` to be 5×5 and `conv1_2` to be 3×3. Hence, the effective receptive field size is $5 + 3 - 1 = 7$, which is equal to the original convolutional kernel size 7×7 and the receptive field size is kept.

Besides the added parameters at the front of the neural network to facilitate learning, we also change the size the image input in order to promote the resolution. We double the image input size and also insert a maxpool layer of kernel size 2×2 after `conv1_1`. It is easy to verify that the feature output dimensions of layer `conv1` and `conv1_2` are both of shape (64,112,112). Hence, the computation of layers `conv1/conv1_2`, and the ones after, are preserved. At the implementation level, the network function shall be perturbed when we double the image input size and half the feature output. However, ConvNets are robust enough to handle such small perturbations. In the proposed experiments, the network recovers to the original accuracy right after several iterations. In summary, the above operation allows us to promote the image input size on a pre-trained neural network without significantly altering the overall network architecture. This operation can be applied for a second round to promote the image input size to be 896.

How to Morph Deep Nets? Function-Preserving Transformations. Mathematically, to morph the deep net displayed in Fig. 2, to a new one with the network function completely preserved, one needs to compute the convolution operation as follows:

$$O_j(c_j) = \sum_{c_i} O_i(c_i) * G_l(c_j, c_i), \tag{1}$$

where the output blobs O_* are 3D tensors of shape (C_*, H_*, W_*) and the convolutional filter G_l is a 4D tensor of shape (C_j, C_i, K_l, K_l). We define C_*, H_*, and W_* as the number of channels, height and width of O_* correspondingly; and K_l is the convolutional kernel size.

The convolutional filter G_l is morphed into two convolutional filters F_l and F_{l+1} (Fig. 2 right), where F_l and F_{l+1} are 4D tensors of shapes

[1] In the proposed experiments, ResNet50 did not boost the performance significantly over ResNet18 on the mammography data. In Raghu et al. (2019), the authors showed similar findings, that is — smaller architectures can perform comparably to standard ImageNet models. Hence, we selected ResNet18.

(C_l, C_i, K_1, K_1) and (C_j, C_l, K_2, K_2). In order to preserve the network function, one needs to morph the network, which expression reads:

$$\tilde{G}_l(c_j, c_i) = \sum_{c_l} F_l(c_l, c_i) * F_{l+1}(c_j, c_l), \tag{2}$$

where $\tilde{G}_l$ is a zero-padded version of G_l whose effective kernel size is $\tilde{K}_l = K_1 + K_2 - 1 \geq K_l$. We remark that we random initialise F_l and F_{l+1}, F_l and F_{l+1} are then iteratively updated through the minimisation process from Wei et al. (2016). The sufficient condition to morph a network was shown in Wei et al. (2016), we therefore seek to fulfil the next condition:

$$\max(C_l C_i K_1^2, C_j C_l K_2^2) \geq C_j C_l (K_1 + K_2 - 1)^2, \tag{3}$$

in the proposed case illustrated in Fig. 2, we have $C_i = 3, C_j = 64, C_l = 64, K_1 = 5, K_2 = 3$. From (3) one can see that $\max(4800, 36864) \geq 9408$ holds. We underline that for the morphing operation, the mapping function from input to output is preserved.

Besides convolutional layers, the BatchNorm layers and ReLU layers needs to be carefully addressed. We adopt the approach of that Wei et al. (2019) for BatchNorm layers by setting: `bn1_2` as `bn1`; and `bn1_1` by using $\gamma = 1$ and $\beta = 0$. We refer to γ and β as the parameters of a BatchNorm layer (Ioffe and Szegedy, 2015). For the non-linear activation layer ReLU, we adopt the approach in Wei et al. (2016) by using a proxy version of PReLU with the slope set to 1.

Morphing Deep Nets for Classification. For the mammogram classification task, we use the cross entropy loss function to train our neural network. For mammography data, there are only two classes, either a benign tumour or a malignant tumour. However, the number of malignant tumour cases are usually significantly less comparing against the benign cases. This imposes a significant class-imbalance problem. Hence, we use the following weighted version of the cross entropy loss, which reads:

$$l(o_i, y_i) = -w[y_i] \left(\log \left(\frac{\exp(o_i[y_i])}{\sum_j \exp(o_i[j])} \right) \right) \tag{4}$$

$$= w[y_i] \left(-o_i[y_i] + \log \left(\sum_j \exp(o_i[j]) \right) \right), \tag{5}$$

where o_i is the predicted output for input image x_i whilst y_i the class label of x_i. Moreover, w refers to the weight vector for all the classes. We set $w[y_i]$ to be inverse proportional to the number of cases in class y_i, and the mean equals to 1. That is:

$$w[y_i] = \frac{C}{\#cases[y_i]} \Big/ \left(\sum_j \frac{1}{\#cases[j]} \right), \tag{6}$$

where C is the total number of classes, $\#cases[y_i]$ is the number of cases in the training dataset for class y_i. With this setup, in one training epoch, all the classes have equal weight of contribution to the loss function. Experimental results show that this weighted cross entropy loss is critical to the performance.

3. Experimental results

In this section, we detail the experiments carried out to evaluate the proposed MorpHR scheme.

3.1. Datasets description & evaluation protocol

We use the benchmarking CBIS-DDSM (Lee et al., 2017) to evaluate the proposed approach. It is composed of 3103 mammography images from 1566 women (see official dataset full description in TCIA: The Cancer Imaging Archive Public Access (2021)). For each breast, both craniocaudal (CC) and mediolateral oblique (MLO) views are included for most of the exams. We treated each view as a separate image in our experiments.

Although the dataset has an official train-test split, the body of literature does not follow the suggested split. For a fair comparison, we randomly split the training data 85:15 at the *patient level* to create independent training and validation datasets. Overall, there were 2097, 361, 645 mammograms of 1061, 187, 349 women in the training, validation, test data respectively. We emphasise that our partition (original test split) is comparable with that of Shen et al. (2019) and other works.

The CBIS-DDSM dataset contains Screen-Film Mammographs (SFM) and their benefits with respect to Full-Field Digital Mammographs (FFDM) has been extensively discussed in the literature e.g., Vinnicombe et al. (2009) and Farber et al. (2021), where the value of SFM is recognised. For example, since early works (Vinnicombe et al., 2009) the authors found that FFDM yields to detection rates at least as high as those for SFM. In most recent works, the authors of that Farber et al. (2021) suggested that the transition from SFM to FFDM did not result in health benefits for screened women. Moreover, SFM can be used along with FFDM in algorithmic development as showed in Shen et al. (2019). These findings merit the value of the CBIS-DDSM dataset.

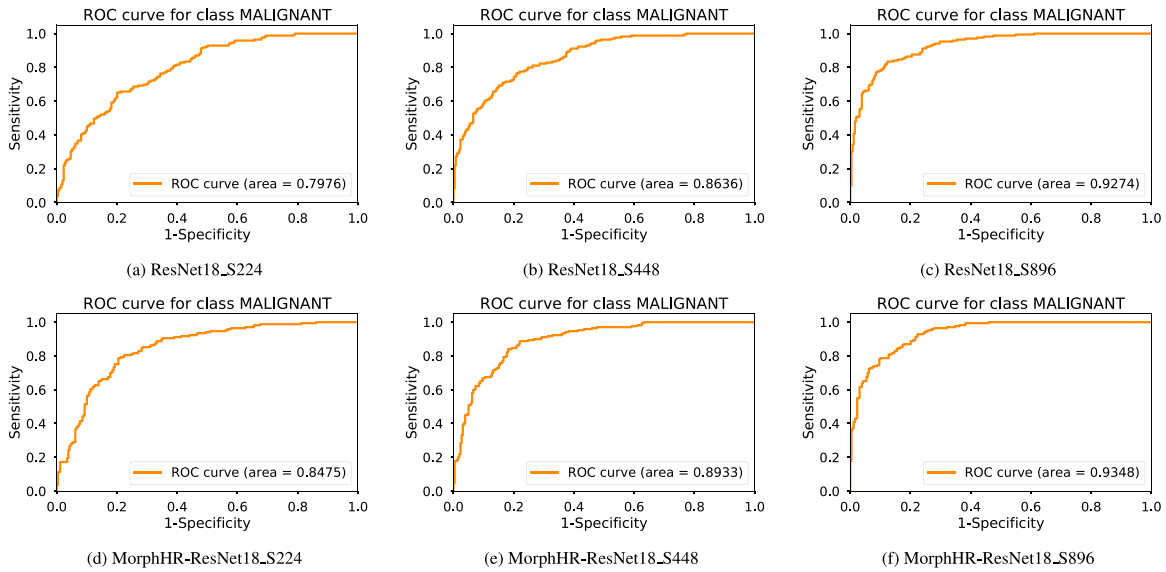
To show generalisation capabilities of the transfer learning of the transfer capabilities of the proposed approach, we use the ChestX-ray14 (Wang et al., 2017) dataset, which is a hospital-scale database consisting of 112,120 chest x-ray images with 14 abnormalities. We used the official dataset split — that is, 70% for training, 10% validation and 20% for testing (see Fig. 4).

We follow standard protocol in machine learning and clinical practice to evaluate our model. We performed a ROC analysis using the area under the curve (AUC). Our motivation to use such analysis is as follows. Firstly, from the machine learning perspective AUC-ROC analyses have been, since early developments e.g. Bradley (1997), a powerful tool for proving meaningful graphical visualisation and a great form to report the model performance with respect to the clinical screening gold-standard. Secondly, in clinical practice e.g. Hanley and McNeil (1982) they play a central role in evaluating diagnostic ability of tests to discriminate the true state of subjects, finding the optimal cutoff values, showing the trade-off between clinical sensitivity and specificity. Moreover and following standard protocol e.g. Shen et al. (2019), we also include 10-crop testing standard way of reporting results for the best performance of a single model. We clarify that 10-crop testing does not modify the model, it is a data augmentation strategy for testing, which includes center-crop plus four corner-crops, and another five crops (center + four corners) for the horizontally flipped image.

3.2. Implementation & training scheme

The mammograms in the CBIS-DDSM data are with a mean of 3138×5220 pixels. In the proposed experiments, we use random crop and random horizontal flip to augment data. The random crop ratio is set to be 0.875. For example, an input size of 1792×2304 means the mammogram is resized to 2048×2633 with an input image of 1792×2304 randomly cropped to feed into the neural network. The weights of the last fully-connected layer is initialised with the Xavier scheme (Glorot and Bengio, 2010). For the network training, we follow the hyper-parameters setup in Shen et al. (2019). The batch size was set to be 32, and Adam (Kingma and Ba, 2014) was used as the optimiser. Some visual samples for the CBIS-DDSM dataset, for malignant and benign cases, are displayed in Fig. 4. In our experiments, we use S224 and S448 to refer to the input size.

Moreover, we follow the 2-stage training strategy in Shen et al. (2019) with the following modifications: (i) we set weight decay to its default value 10^{-4} because there was not a noticeable benefit for changing its value; (ii) we did not freeze certain layers in stage-1 because it could slightly hurt the performance. The following strategy was adopted in the proposed experiments:



(a) ResNet18_S224 (b) ResNet18_S448 (c) ResNet18_S896

(d) MorphHR-ResNet18_S224 (e) MorphHR-ResNet18_S448 (f) MorphHR-ResNet18_S896

Fig. 3. ROC curves on the CBIS-DDSM dataset. Image resolution plays a critical role in the task of screening mammography classification.

Table 1
Experimental results of the proposed MorphHR scheme on the CBIS-DDSM dataset.

	Layer4 Computed on	Input Size	Val (1-crop)	Val (10-crops)	Test (1-crop)	Test (10-crops)
ResNet18_S224	7 × 9	224 × 288	78.15	79.76	72.57	75.33
MorphHR-ResNet18_S224	7 × 9	448 × 576	82.5	84.75	75.23	76.73
ResNet18_S448	14 × 18	448 × 576	83.82	86.36	78.82	79.95
MorphHR-ResNet18_S448	14 × 18	896 × 1152	87.48	89.33	78.36	80.13
ResNet18_S896	28 × 36	896 × 1152	91.26	92.74	79.58	80.68
MorphHR-ResNet18_S896	28 × 36	1792 × 2304	91.82	93.48	79.64	81.87
MorphHR-ResNet18_S896 (Ensemble)	28 × 36	1792 × 2304	93.60	**94.27**	82.16	**83.13**

Table 2
Numerical comparison of the proposed MorpHR scheme, the SOTA-model of that Shen et al. (2019) and human estimation (Lehman et al., 2016) on the CBIS-DDSM dataset.

	Sensitivity	Specificity	PPV	NPV	AUC (val)	AUC (test)
Single Model Single Crop (Shen et al., 2019)	–	–	–	–	85%	~75%
Single Model (Shen et al., 2019)	–	–	–	–	88%	–
Four Models Ensemble (Shen et al., 2019)	86.10%	80.10%	–	–	91%	–
Single Model Single Crop (MorphHR)	–	–	–	–	91.82%	79.64%
Single Model (MorphHR)	86.98%	82.57%	76.56%	90.68%	93.48%	81.87%
Four Models Ensemble (MorphHR)	**90.0%**	86.24%	80.85%	92.92%	**94.27%**	83.13%
Human (Estimation) (Lehman et al., 2016)	86.9%	**88.9%**	–	–	–	–

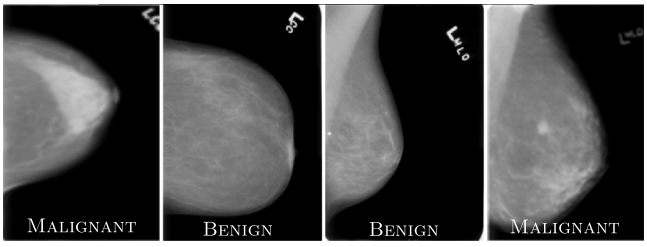


Fig. 4. Visual samples of the CBIS-DDSM dataset. These samples display malignant and benign cases.

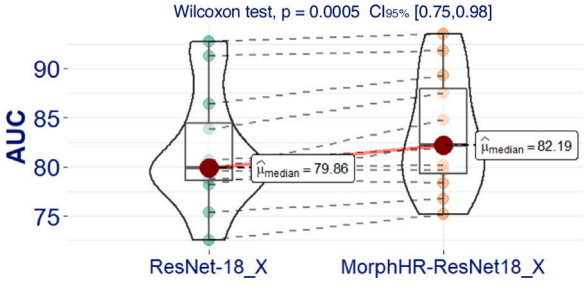


Fig. 5. Wilcoxon test of our proposed technique and ResNet. The statistical test uses the results from Table 1.

- Stage-(1) Set learning rate to 10^{-4} and train all the layers for 30 epochs,
- Stage-(2) Set learning rate to 10^{-5} and train all the layers for an additional 20 epochs.

3.3. Results and discussion

In this section, we detail our findings and give deeper insights into the good performance of the proposed MorphHR approach.

Table 3
Experimental results of the proposed MorphHR scheme on the ChestX-ray14 (Mass) dataset.

	Layer4 computed on	Input Size	Test	Test (10-crops)
ResNet18_S224	7 × 7	224 × 224	79.74	81.55
MorphHR-ResNet18_S224	7 × 7	448 × 448	80.98	82
ResNet18_S448	14 × 14	448 × 448	80.67	82.44
MorphHR-ResNet18_S448	14 × 14	896 × 896	82.53	**83.77**

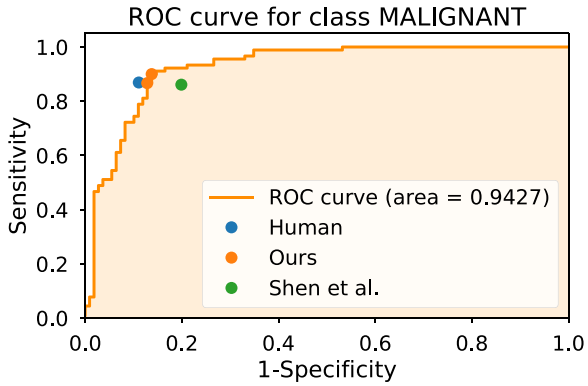


Fig. 6. Sensitivity and Specificity on ROC curve. The proposed system achieved performance on a par with mammography experts.

Fine-Tuning vs. MorphHR. We begin our evaluation by comparing the performance of the proposed approach against fine-tuning using a ROC-AUC analysis. We first compare the performances of ResNet and MorphHR-ResNet, on the CBIS-DDSM dataset, which results are displayed in Table 1. In this table, we show the AUC scores on the validation and test data, with both single-crop and 10-crops results. In a closer look at the results, one can see that the proposed MorphHR scheme has a consistent improvement over the baseline counterpart.

The performance improvement on the validation data is 4.35% (82.5% vs. 78.15%), 3.66% (87.48% vs. 83.82%), 0.56% (91.82% vs. 91.26%) for image size 224, 448, 896 respectively. To further support our results from Table 1, we ran a non-parametric Wilcoxon test. We report the results along with the 95% confidence interval in Fig. Fig. 5. From the results, we conclude that our approach is statistical significant different, $p = 0.005$, than the compared one at a significance level alpha = 0.05. There are also other well-known networks that perform lower than our technique and baseline. For example, in testing data, GoogleNet AUC = 69.65% (vs our 75.23%) and for AlexNet AUC = 71.11% (vs our 75.23%).

We remark that "Ensemble" denotes an average of the output scores run in four rounds, we include it in order to compare against the ensemble of fours models of that Shen et al. (2019). That is, for the ensemble, we average the four probabilities outputs as the final probability for decision. Because it is binary classification, if the final probability $p \geq 0.5$, we classify the mammograms as benign, otherwise, it is malignant. In terms of computational time, ResNet18_S224 and MorphHR-ResNet18_S224, the training time is 80 min and 100 min respectively on 4GPUs for 50 epochs. We highlight that it is meaningful that the latter is slightly slower *with more layers and doubling the image input resolution and significantly improved accuracy's and AUCs.*

Resolution is Critical for High Resolved Screening Mammogram Classification. We reveal that resolution is critical for mammogram classification. It is intuitive that mammograms are high resolved (~4K) in nature. However, due to the limit of GPUs memory, current mainstream approaches either decrease the resolution or to extract patches. We argue that the importance of resolution for medical image classification should be emphasised. In particular, this research is proposed to resolve this resolution issue.

To further support the results, we display the ROC curves of fine-tuning vs MorphHR using different input sizes. The results are reported in Fig. 3, where a set of ROC curves over several resolutions are displayed as a powerful graphical tool to show the trade-off between clinical sensitivity and specificity. In a close look at Fig. 3, one can observe that greater discriminant capacity on the malignant diagnostic is achieved by our MorphHR model than a standard network. As MorphHR displays curves closer to the top-left corner indicating a better performance. This results with an accompanying statistical significant, $p < 0.05$, improvement up to AUC 6% (p-value = 0.002). From a comparison at those plots, one can observe that the proposed approach performs better than fine-tuning at all thresholds and for all resolutions. Overall, one can observe that *the image resolution plays a very important role in the task of screening mammography classification.* This is meaningful because a higher resolution image gives more fine details, especially when considering the fact that the lesion regions only accounts for a small part of the whole image.

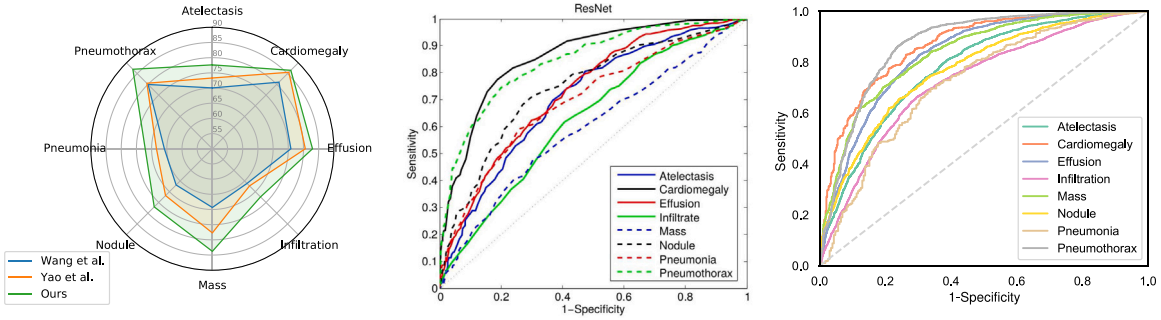
MorphHR vs. SOTA-Model. We now compare the proposed approach against the SOTA method of that Shen et al. (2019), which to the best of our knowledge holds the state-of-the-art results in the CBIS-DDSM dataset. The results are displayed in Table 2.

On the validation dataset, using four models ensemble, we improve current best results of 91% in Shen et al. (2019) to 94.27%. In summary, there is a 7% improvement for single model single crop, a 5.5% improvement for single model, and a 3% improvement for four models ensemble. The sensitivity and specificity are also improved from 86.10% and 80.10% to 90.00% and 86.24% respectively. The sensitivity and specificity of screening mammography of human radiologists are reported to be an average of 86.9% and 88.9% respectively (Lehman et al., 2016). It can be seen that the proposed CAD system is on par with mammography experts. This is further illustrated in Fig. 6, where the sensitivity and specificity pairs on the ROC curves are drawn.

In Table 2, we also include the experimental results on the testing dataset. In Shen et al. (2019), the authors used a custom split and did not report results on the official testing dataset. The 75% AUC score is adopted from Shen (2019), which is reported on the official testing data using the original authors' deployed models. It can be seen on the testing dataset, we are able to achieve a 4.64% AUC improvement (75% vs. 79.64%). One may notice that in Table 2, for both Shen et al. (2019) and the proposed MorphHR scheme, the performances on the testing data is significantly lower than those on the validation data. This is because the testing data is another holdout set acquired in a different time. It contains cases which are intrinsically more difficult and bear different distributions (Shen, 2019).

Generalisation Capabilities: The X-ray Case. To demonstrate the generalisation capability of the proposed approach for transferring knowledge, we further carry out experiments on the ChestX-ray14 (Wang et al., 2017) dataset. We follow the same experimental protocol as the previous section.

Fine-Tuning vs. MorphHR. We start by first supporting our claim regarding the limitation of fine-tuning for medical data. In Table 3, we report a performance comparison in terms of AUC, the results reflect the outputs on the "Mass" class, which is close related to the mammography case. From these results, one can observe that the proposed approach report the best performance — that is, the results are consistent with those findings on the CBIS-DDSM dataset. By using the proposed MorphHR scheme, we can achieve up to 2% (82.53% vs. 80.67%) AUC performance improvement than using only fine-tuning. To further support the performance improvement of our



(a) (b)

Fig. 7. Comparison of the proposed MorpHR scheme against existing approaches in (a) Radar plot (Wang et al., 2017; Yao et al., 2018). (b) Left: ROC curve from Wang et al. (2017); right: ROC curve of the proposed approach.

Table 4
Numerical comparison of the proposed MorpHR scheme and the SOTA-model on the ChestX-ray14 dataset. "In 8" means this category is in the original ChestX-ray8 dataset..

	In 8	Num. Images	Wang et al. (2017)	Yao et al. (2018)	Yao et al. (2018)	Aviles-Rivero et al. (2019)	MorphHR	MorphHR(10-crops)
Atelectasis	T	11 559	70.69	70.03	73.3	71.89	76.12	77.62
Cardiomegaly	T	2776	81.41	81	85.6	87.99	85.81	86.53
Consolidation		4667	–	70.32	71.1	73.36	77.38	77.60
Edema		2303	–	80.52	80.6	80.20	84.77	85.44
Effusion	T	13 317	73.62	75.85	80.6	79.20	81.50	82.97
Emphysema		2516	–	83.3	84.2	84.07	88.98	89.78
Fibrosis		1686	–	78.59	74.3	80.34	80.59	81.36
Hernia		227	–	87.17	77.5	87.22	85.42	88.22
Infiltration	T	19 894	61.28	66.14	67.3	72.05	71.13	72.42
Mass	T	5782	56.09	69.33	77.7	80.90	82.53	83.77
Nodule	T	6331	71.64	66.87	71.8	71.13	75.15	77.07
Pleural_Thickening		3385	–	68.35	72.4	75.70	78.44	80.00
Pneumonia	T	1431	63.33	65.8	68.4	76.64	69.70	72.11
Pneumothorax	T	5302	78.91	79.93	80.5	83.70	86.52	87.01
Average AUC		–	–	73.8	76.1	78.88	80.29	**81.56**

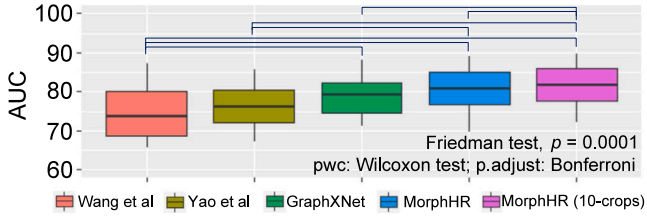


Fig. 8. Multiple pair-wise comparisons, using a paired Wilcoxon test and p-values adjusted using the Bonferroni method, between existing techniques and our proposed MorpHR on the ChestX-ray14.

technique, we ran a set of statistical tests. Firstly, we ran a non-parametric statistical test for supporting Table 4. Most precisely, we use the Friedman test for multiple comparison. We found that our AUC reported scores were statistically significantly different with respect to the compared approaches $\chi^2(4) = 45.44$, $p < 0.0001$. This test then was followed by a Wilcoxon test for multiple pair-wise comparison with adjusted p-values using the Bonferroni method. This test revealed that our method is statistically significant different in AUC for all methods (see Fig. 8). Hence, the effectiveness of the proposed MorphHR scheme is demonstrated.

MorphHR vs. SOTA-Model. To further support of our results, we also reported the average AUC scores on all pathologies, the results are reported in Table 4 against (Wang et al., 2017; Yao et al., 2018; Aviles-Rivero et al., 2019). In a detail inspection, one can observe that overall the proposed approach reports SOTA results for this dataset. We can also observe that there are only few pathologies with clear variability. It is because of the limitation of the dataset, for example "Hernia" is the one with the fewest samples in the dataset. The dataset, therefore, is not always representative for each class (despite the data augmentation).

We remark that other results reported on this dataset are, either not on the official test data (Guendel et al., 2018; Baltruschat et al., 2019), or using per-image split (Yao et al., 2017; Rajpurkar et al., 2017). Fig. 7(a) compares against (Wang et al., 2017; Yao et al., 2018) in radar plot and Fig. 7(b) compares against (Wang et al., 2017) in ROC curves. The advantage of the proposed MorphHR scheme is obvious.

Overall, to the best of our knowledge, we are reporting SOTA performances for both medical dataset cases — that is, the mammography and X-ray data classification.

4. Conclusions and future work

In this work, we proposed a new transfer learning framework for improving mammography data classification beyond fine-tuning, which learns domain specific features, accepts high resolution inputs and is scalable to hardware. Fine-tuning is the de-facto method for deep learning application to medical imaging. However, it only allows the modification of the last few layers of a neural network to adapt the high-level concept class information into the new dataset. In this research, we proposed to use network morphism to alter the front of a neural network to learn the considerable differences between natural images and medical images. In particular, we also proposed a concrete learning scheme to deal with the high resolution nature of mammographic images. Extensive experiments were carried out on the benchmark datasets CBIS-DDSM and ChestX-ray14 to achieve state-of-the-art results.

In this research, the proposed modification of the standard ImageNet models for high resolution mammogram classification is simple and effective. It is meaningful we can design more elegant patterns for block A' in Fig. 2 to further improve the performance, which will be the focus for future work. As future work we also further explore several type of architectures such as along with our proposed approach. Moreover, the behaviour of our approach when using tiny training sets will be another interesting area of future exploration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

AIAR gratefully acknowledges the financial support of the CMIH, UK (EP/T017961/1) and CCIMI University of Cambridge, UK. CBS acknowledges support from the Leverhulme Trust project on Breaking the non-convexity barrier, UK, the Philip Leverhulme Prize 2018, EPSRC Centre, UK EP/N014588/1 and EP/T017961/1, the RISE projects CHiPS and NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. YH acknowledges the financial support of the CMIH, UK (EP/T017961/1). FJG is an NIHR Senior Investigator.

References

Agarwal, R., Diaz, O., Lladó, X., Yap, M.H., Martí, R., 2019. Automatic mass detection in mammograms using deep convolutional neural networks. J. Med. Imaging 6 (3), 031409.

Akselrod-Ballin, A., Karlinsky, L., Alpert, S., Hasoul, S., Ben-Ari, R., Barkan, E., 2016. A region based convolutional network for tumor detection and classification in breast mammography. In: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 197–205.

Aviles-Rivero, A.I., Papadakis, N., Li, R., Sellars, P., Fan, Q., Tan, R.T., Schönlieb, C.-B., 2019. GraphX^{NET} - Chest X-Ray classification under extreme minimal supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 504–512.

Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A., 2019. Comparison of deep learning approaches for multi-label chest X-ray classification. Sci. Rep. 9 (1), 6381.

Becker, A.S., Marcon, M., Ghafoor, S., Wurnig, M.C., Frauenfelder, T., Boss, A., 2017. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Investig. Radiol. 52 (7), 434–440.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30 (7), 1145–1159.

Brzakovic, D., Luo, X.M., Brzakovic, P., 1990. An approach to automated detection of tumors in mammograms. IEEE Trans. Med. Imaging 9 (3), 233–241.

Chun-ming, T., Xiao-mei, C., Xiang, Y., Fan, Y., et al., 2019. Five classification of mammography images based on deep cooperation convolutional neural network. Am. Sci. Res. J. Eng. Technol. Sci. (ASRJETS) 57 (1), 10–21.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Elmore, J.G., Jackson, S.L., Abraham, L., Miglioretti, et al., 2009. Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 253 (3), 641–651.

Farber, R., Houssami, N., Wortley, S., Jacklyn, G., Marinovich, M.L., McGeechan, K., Barratt, A., Bell, K., 2021. Impact of full-field digital mammography versus film-screen mammography in population screening: a meta-analysis. JNCI: J. Natl. Cancer Inst. 113 (1), 16–26.

Geras, K.J., Wolfson, S., Shen, Y., Wu, N., Kim, S., Kim, E., Heacock, L., Parikh, U., Moy, L., Cho, K., 2017. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint: 1703.07047.

Gilbert, F.J., Astley, S.M., McGee, M.A., Gillan, M.G., Boggis, C.R., Griffiths, P.M., Duffy, S.W., 2006. Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. Radiology 241 (1), 47–53.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 249–256.

Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D., 2018. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In: Iberoamerican Congress on Pattern Recognition. Springer, pp. 757–765.

Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R., 2018. Deep learning in mammography and breast histology, an overview and future trends. Med. Image Anal. 47, 45–67.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hickman, S.E., Woitek, R., Le, E.P.V., Im, Y.R., Mouritsen Luxhøj, C., Aviles-Rivero, A.I., Baxter, G.C., MacKay, J.W., Gilbert, F.J., 2022. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. Radiology 302 (1), 88–104.

Huynh, B.Q., Li, H., Giger, M.L., 2016. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J. Med. Imaging.

Innamorati, C., Ritschel, T., Weyrich, T., Mitra, N.J., 2018. Learning on the edge: Explicit boundary handling in CNNs. arXiv:1805.03106.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint: 1502.03167.

Kim, H.-E., Kim, H.H., Han, B.-K., Kim, K.H., Han, K., Nam, H., Lee, E.H., Kim, E.-K., 2020. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. Lancet Digit. Health 2 (3), e138–e148.

Kingma, D., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint: 1412.6980.

Kornblith, S., Shlens, J., Le, Q.V., 2019. Do better imagenet models transfer better? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2661–2671.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 1097–1105.

Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci. Data 4, 170177.

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., et al., 2016. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology 283 (1), 49–58.

Lehman, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N., Miglioretti, D.L., 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Med. 175 (11), 1828–1837.

Lévy, D., Jain, A., 2016. Breast mass classification from mammograms using deep convolutional neural networks. arXiv preprint: 1612.00542.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al., 2020. International evaluation of an AI system for breast cancer screening. Nature 577 (7788), 89–94.

Mercan, C., Aksoy, S., Mercan, E., Shapiro, L.G., Weaver, D.L., Elmore, J.G., 2017. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. IEEE Trans. Med. Imaging 37 (1), 316–325.

Petrosian, A., Chan, H.-P., Helvie, M.A., Goodsitt, M.M., Adler, D.D., 1994. Computer-aided diagnosis in mammography: classification of mass and normal tissue by texture analysis. Phys. Med. Biol. 39 (12), 2273.

Ragab, D.A., Sharkas, M., Marshall, S., Ren, J., 2019. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 7, e6201.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning with applications to medical imaging. arXiv preprint: 1902.07208.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint: 1711.05225.

Rampun, A., Wang, H., Scotney, B., Morrow, P., Zwiggelaar, R., 2018. Classification of mammographic microcalcification clusters with machine learning confidence levels. In: 14th International Workshop on Breast Imaging (IWBI 2018), Vol. 10718. International Society for Optics and Photonics, p. 107181B.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. 8 (1), 4165.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Shen, L., 2017. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. arXiv preprint: 1711.05775.

Shen, L., 2019. https://github.com/lishen/end2end-all-conv/issues/5, accessed: 2019-11-11.

Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh, W., 2019. Deep learning to improve breast cancer detection on screening mammography. Sci. Rep. 9.

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., et al., 2021. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med. Image Anal. 68, 101908.

Sun, H., Li, C., Liu, B., Zheng, H., Feng, D.D., Wang, S., 2018. AUNet: Attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. arXiv preprint: 1810.10151.

Tardy, M., Mateus, D., 2021. Looking for abnormalities in mammograms with self-and weakly supervised reconstruction. IEEE Trans. Med. Imaging.

TCIA: The Cancer Imaging Archive Public Access, 2021. The CBIS-DDSM (Curated Breast Imaging Subset of DDSM). URL: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM.

Vinnicombe, S., Pinto Pereira, S.M., McCormack, V.A., Shiel, S., Perry, N., dos Santos Silva, I.M., 2009. Full-field digital versus screen-film mammography: comparison within the UK breast screening program and systematic review of published data. Radiology 251 (2), 347–358.

Vyborny, C.J., Giger, M.L., 1994. Computer vision and artificial intelligence in mammography. AJR Am. J. Roentgenol. 162 (3), 699–708.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106.

Warren, R., Duffy, W., 1995. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Br. J. Radiol. 68 (813), 958–962.

Wei, T., Wang, C., Chen, C.W., 2019. Stable network morphism. In: International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019. pp. 1–8. http://dx.doi.org/10.1109/IJCNN.2019.8851955.

Wei, T., Wang, C., Rui, Y., Chen, C.W., 2016. Network morphism. In: Proceedings of the 33rd International Conference on Machine Learning. pp. 564–572.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Févry, T., Katsnelson, J., Kim, E., et al., 2019. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging.

Xi, P., Shu, C., Goubran, R., 2018. Abnormality detection in mammography using deep convolutional neural networks. In: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA). IEEE, pp. 1–6.

Yala, A., Schuster, T., Miles, R., Barzilay, R., Lehman, C., 2019. A deep learning model to triage screening mammograms: a simulation study. Radiology 293 (1), 38–46.

Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., Lyman, K., 2017. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint: 1710.10501.

Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K., 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv preprint: 1803.07703.