Medical Image Analysis 83 (2023) 102690

Contents lists available at ScienceDirect

Medical Image Analysis

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

Symmetry-based regularization in deep breast cancer screening

Eduardo Castro [a,b,*], Jose Costa Pereira [a,c], Jaime S. Cardoso [a,b]

[a] INESC TEC, Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[b] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[c] Huawei Technologies R&D, Noah's Ark Lab, Gridiron building, 1 Pancras Square, 5th floor, London N1C 4AG, United Kingdom

ARTICLE INFO

Keywords:
Breast cancer
Deep neural network
Regularization
Equivariance
Computer-aided diagnosis

ABSTRACT

Breast cancer is the most common and lethal form of cancer in women. Recent efforts have focused on developing accurate neural network-based computer-aided diagnosis systems for screening to help anticipate this disease. The ultimate goal is to reduce mortality and improve quality of life after treatment. Due to the difficulty in collecting and annotating data in this domain, data scarcity is – and will continue to be – a limiting factor. In this work, we present a unified view of different regularization methods that incorporate domain-known symmetries in the model. Three general strategies were followed: (i) data augmentation, (ii) invariance promotion in the loss function, and (iii) the use of equivariant architectures. Each of these strategies encodes different priors on the functions learned by the model and can be readily introduced in most settings. Empirically we show that the proposed symmetry-based regularization procedures improve generalization to unseen examples. This advantage is verified in different scenarios, datasets and model architectures. We hope that both the principle of symmetry-based regularization and the concrete methods presented can guide development towards more data-efficient methods for breast cancer screening as well as other medical imaging domains.

1. Introduction

Breast cancer is the most common and lethal form of cancer in women, accounting for one fourth of the total number of new cancer cases within this population (Sung et al., 2021). Early diagnosis reduces the risk of dying and often allows for additional treatment options, such as breast-conserving surgery (American Cancer Society, 2021), reducing the negative impact of the disease on the patient's life after treatment. Several countries have implemented screening programs ensuring all asymptomatic women over a certain age have access to periodic checkups (Fryback et al., 2006; Altobelli et al., 2017).

Typically, breast cancer screening is based on mammography. One or two trained human readers look at two X-ray views of each breast for possible signs that support a positive diagnosis (e.g., masses, calcifications, distortions). While a second reader increases sensitivity, i.e., the proportion of positive cases that are detected, it also results in more women being recalled for further examination, many of whom do not have the disease. This trade-off is preferable since the potential harm of a false negative diagnosis is considered higher than the cost of further examination of some healthy women. Nevertheless, the cost of false positives is considerable in the form of additional radiation, stress, and financial cost.

Computer-Aided Detection (CADe) and Diagnosis (CADx) systems can aid the decision-making process in clinical practice. While CADe systems detect regions of interest in images, CADx systems predict the pathology or probability of malignancy for an exam, image, or image region. Before the widespread of deep learning in medical imaging, some studies indicated there was no benefit when using CAD systems in single-reader settings (Lehman et al., 2015), while others suggested that it had a comparable effect to adding a second human reader (Gromet, 2008). Houssami et al. (2009) point out that even though CAD helps find otherwise missed cancers, it increases false-positive diagnoses. The authors also state that refining CAD algorithms may improve their potential in clinical practice. Recently, deep learning methods have fueled a new generation of CAD systems. Bahl (2019) highlighted that these approaches focus on learning specific outcomes rather than closely mimicking the process that guides specialists' assessment and thus can better distinguish between benign and malignant findings, effectively addressing the high false-positive rate of previous systems. Recent studies indicate that these new CAD systems can improve decision-making in screening. For instance, Schaffter et al. (2020) showed that combining the assessments of experts and algorithms can lead to better decisions, although their work does not focus on usability in clinical practice. In a retrospective study, Rodríguez-Ruiz et al. (2019) showed

* Corresponding author at: INESC TEC, Campus da Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal.
E-mail address: emcastro@inesctec.pt (E. Castro).

https://doi.org/10.1016/j.media.2022.102690
Received 15 December 2021; Received in revised form 28 October 2022; Accepted 9 November 2022
Available online 21 November 2022

that using an AI-based CAD to assist experts improved sensitivity and specificity in a single-reading setting. Similar results were found by Conant et al. (2019) in digital breast tomosynthesis (DBT), an imaging modality closely related to mammography also used for breast cancer screening. Given the relatively high incidence of the disease and its potential harms, improving the accuracy of AI-based algorithms in screening mammography could benefit women populations under screening programs (Gao et al., 2019).

Recent developments in deep learning (LeCun et al., 2015) have led to improved image recognition models and strongly impacted different fields, including CAD in breast cancer. Since, under these new methods, the visual features required for solving a task are learned rather than manually set, the need for application-specific solutions is reduced. More emphasis is placed on collecting and curating large and high-quality datasets, as the scale and quality of these largely influences the model's accuracy. Despite this, breast cancer screening still poses unusual research challenges. While mammography images are high-resolution, the information needed for a correct diagnosis is contained in just a tiny portion of the breast, and the remaining tissue can appear healthy. Additionally, signs of a positive diagnosis are often found in subtle features. Thus, an image-wide label is not as informative as local annotations. Most works in the field reflect this, at least partially (e.g., Kooi et al. (2017b), Arevalo et al. (2016), Ribli et al. (2018) and Wu et al. (2020)). Perhaps most notably, the top submissions on the Digital Mammography DREAM challenge (Schaffter et al., 2020), a one-year-long competition towards the development of better CAD systems for mammography, all resorted mostly to strongly annotated datasets, even though these were much smaller in size when compared to the ones provided by the organization with breast-level labels.

Unfortunately, it is not as easy to generate and annotate large high-quality datasets for mammography as it is for other computer vision problems. Manual annotation is a tedious process that requires expert knowledge. Further, most screened women have a negative diagnosis, which severely limits the amount of positive cases that can be collected. For example, the number of images on the well-known ImageNet (Deng et al., 2009) dataset corresponds to the estimated number of breast cancer cases in Europe for the next three years in women over 50, the age at which many screening programs start. Collecting and annotating such a large number of exams would be a daunting task.

Due to the inherent difficulties in collecting and annotating large datasets, it is critical to improve the data efficiency of current methods to increase the accuracy of CAD systems in data-scarce scenarios. One way to achieve this is to incorporate domain knowledge into the learning process of neural networks. For instance, the orientation of a lesion is not indicative of its malignancy, and thus the model response should be invariant to this feature. In this work, we propose different ways of brewing this and other known symmetries into deep learning models. We validate the proposed strategies in the task of mass classification and show that adopting the proposed techniques improves model accuracy. Further, we extend these results to malignancy prediction in images of the whole breast for different datasets. Although our experimental evaluation focuses on CADx scenarios, the proposed framework is general and can be applied in other systems.

This paper is organized as follows. In Section 2, we summarize previous work in regularization for breast cancer screening, and equivariance and invariance in other medical domains. The proposed symmetry-based regularization methods are presented in Section 3. In 4 and 5 we describe the experimental settings and discuss the results obtained, before concluding in 6.

2. Related work

The development of CAD algorithms in the context of breast cancer screening in mammography has been around for more than half a century (Winsberg et al., 1967). However, it has recently shifted towards the use of deep learning methodologies, as in other medical imaging applications (Litjens et al., 2017). Recent works use neural networks to process the mammography data directly and solve a specific task. These include: lesion detection (CADe) (Boot and Irshad, 2020; Mordang et al., 2016; Agarwal et al., 2020), lesion classification (CADx) (Arevalo et al., 2016; Kooi et al., 2017b,a), and end-to-end breast cancer diagnosis (CADx) (Ribli et al., 2018; Shen et al., 2019; Li et al., 2021; Geras et al., 2017; Wang et al., 2021; Cogan et al., 2019; Shu et al., 2020). Most works on these tasks use local expert annotations to optimize models with a few exceptions (Geras et al., 2017; Tardy and Mateus, 2021, 2022; Shu et al., 2020). These annotations are considered more informative than a single image-wide label. Interestingly, Geras et al. (2017) show that despite the use of a large dataset (800k images), performance has not saturated, and more data would increase model accuracy.

Due to the high capacity of modern neural networks, typically, these models fit the training data perfectly, but their accuracy on unseen data is limited. This overfitting effect is aggravated for small datasets, common in medical fields, which has led researchers to use transfer learning approaches, where the model is first pre-trained on large datasets of different domains and then fine-tuned to a specific task (Raghu et al., 2019; Mednikov et al., 2018). The extent to which the model replicates precision for unseen data is called generalization and has been an important topic of research in deep learning (Zhang et al., 2021). Generalization can be evaluated on new data sampled from the same distribution (in-dataset) or from a different but related distribution (cross-dataset). Wang et al. (2020) and Cardoso et al. (2017) have shown that deep learning models exhibit unsatisfactory generalization in cross-dataset scenarios. This is a substantial limitation for their effective use in clinical practice, which must be addressed.

Regularization methods improve the ability of neural networks to generalize to new data. The ubiquitous technique of data augmentation can be understood as a form of regularization and usually improves model accuracy. For instance, flips and random cropping improved end-to-end classification in the work of Li et al. (2021). Similar results were found by Kooi et al. (2017b) using translations, scaling, and rotations in lesion classification. Cogan et al. (2019) used flipping, rotations and scaling. Castro et al. (2018) proposed elastic deformations to mimic the breast's natural elasticity during image acquisition in mammography, leading to more accurate CNNs in lesion detection. A more recent line of research is using generative data as a form of increasing the available training data. Authors often resort to generative adversarial neural networks (GANs) for this task. Alyafi et al. (2019) identified that lesion patches are often the minority class in classification problems, and synthetic data can attenuate this disparity. Wu et al. (2018) proposed adding or removing lesions from image patches. Jendele et al. (2019) add malignant features to the whole image of the breast. Guan and Loew (2019) generate two types of synthetic patches, normal and abnormal, and shows improved accuracy when including these in the training dataset. These works show that GAN-generated synthetic samples increase model accuracy. Alternatively, De Sisternes et al. (2015) proposed a three-dimensional computational model for mass generation. Cha et al. (2019) showed that synthetic samples based on this model can reduce overfitting. Tardy and Mateus (2021) extended the generation procedure to account for distortions and clusters of microcalcifications. Their method can learn in a weakly supervised setting, hence reducing the burden of local expert annotation. Outside the domain of mammography, Zhang et al. (2020) applied a sequence of augmentation transformations to the data (BigAug) during optimization and showed that this strategy can significantly increase out-of-domain generalization in medical image segmentation tasks.

Innovations to the model's architecture or loss function have also been used to regularize training and improve accuracy. For instance, Wang et al. (2021) use a neural network with two binary classifiers and a modified loss function to learn to distinguish between malignant and benign examples. Examples classified inconsistently between the two classifiers are given more weight during training. The authors show

that their method improves accuracy when used on top of well-known architectures. A dual-path architecture was proposed by Li et al. (2021). One path captures image features using a standard CNN, while the other focuses on geometric features by first generating a segmentation mask and, only then, extracting features. A new loss function is presented by Li et al. (2019). The authors first find adversarial examples. Based on these examples and the original data, they build a signed graph by connecting points in the same neighborhood. A loss function is defined based on this graph which encourages neighboring points of the same class to be close and different classes to be separated by at least a fixed margin. Tardy and Mateus (2022) propose an image-level multitask objective based on image reconstruction and the classification of binary malignancy, cancer probability, breast density, and laterality.

Outside the breast cancer screening domain, several types of regularization have been proposed. We focus on those that use the concepts of equivariance and invariance as fundamental principles for improving model robustness. Test-time augmentation, which averages the model's output for multiple input transformations, leads to invariant methods and improves accuracy. For instance, Ciresan et al. (2013) followed this strategy, motivated by the knowledge that orientation is arbitrary in histology images acquisition. A general mathematical framework for equivariant convolutional models was laid out by Cohen and Welling (2016) and later adapted to account for different types of transformations (Esteves et al. (2018), Cohen et al. (2019) and Zhu et al. (2019)). Various authors validated these models in different applications, including medical imaging (Li et al., 2020; Graham et al., 2020; Chidester et al., 2019; Li et al., 2018; Lafarge et al., 2021). Dumont et al. (2018) showed that group equivariant architectures are more robust against adversarial attacks. Alternatively, some authors have introduced penalizing terms in the loss function to make the model learn invariant features. Cheng et al. (2016) introduced the rotation invariant layer by penalizing the norm of the difference between the transformed input and the average representation of the input. The model is used to detect objects in satellite images. Within the same domain, Qi et al. (2021) proposed to generate soft labels from the output of an image rotation instead and use them to train the classifier with the cross-entropy loss.

The use of neural networks characterizes recent breast cancer screening research and has led to significant advances in terms of precision. Some studies have proposed different regularization techniques to cope with the dependence of these models on large, strongly annotated datasets, which are difficult to obtain. We follow this line of work and study a comprehensive set of regularization techniques unified under the same design principle: using known symmetries of the breast cancer screening domain to restrict models to learning more appropriate functions. Our contributions are as follows:

- We investigate which symmetries, when incorporated into the learning process, lead to more accurate deep learning classifiers for breast cancer screening.
- We propose different methodologies for incorporating these symmetries, namely data augmentation, invariance regularization, and equivariant model architectures. We show how the concept of equivariance and invariance relates to these methods and help explain why they work in practice.
- We show that adopting these methods can improve model accuracy for different models, datasets, and tasks.

We hope that our work can be used to improve model accuracy in data-scarce scenarios, such as breast cancer screening and other medical imaging tasks domain. A better understanding of the concepts of equivariance and invariance can aid the better design of future CAD approaches. The code used for the experimental section is available online.[1]

3. Equivariance and invariance as regularization

The amount of available training data often limits the accuracy of deep-learning-based CAD systems. Even though current state-of-the-art models can fit most datasets perfectly, the features learned do not necessarily generalize to unseen examples. This phenomenon is called overfitting and happens due to models learning spurious correlations specific to the training samples. Regularization is thus essential to limit the model's capacity to overfit and improve overall accuracy.

To properly regularize training, it is crucial to understand which restrictions should be imposed on the function space (set of learnable functions). The properties of equivariance and invariance provide a natural way to design these constraints. Depending on the application domain, the model's output should vary predictably for transformed versions of the input. For instance, we often want image classifiers to be invariant to input translations, as these do not substantially change the scene's content. Enforcing or promoting this behavior constrains the function space, limiting the model's capacity to learn features specific to the training dataset.

More formally, we define equivariance as:

$$\Phi(T_g \circ x) = T'_g \circ \Phi(x) \tag{1}$$

where Φ is a function defined on set Ω and $x \in \Omega$. g is a group element of group G, and T_g, T'_g are group actions on the domain and codomain of Φ, respectively. The operator $\circ$ is used to denote both the group operation and group action.[2] If Eq. (1) holds, we say that Φ is equivariant under T. Invariance is obtained when T'_g is the identity transformation for all g. In other words, the output remains unchanged for a set of transformations on the input.

In the case of breast cancer screening, different arbitrary factors can alter the appearance of the mammogram, many of which are independent of whether the patient has breast cancer or not. Examples of these factors include the type of sensor used (Yaffe and Mainprize, 2004), the amount of radiation employed, the positioning and pressure applied to the patient's breast, and the post-processing done by the imaging system. A robust model should be invariant to changes in appearance induced by these factors while remaining sensitive to patterns indicative of breast cancer. Throughout the rest of this section, we will study different ways of regularizing neural networks by enforcing or promoting this behavior. We focus on three main approaches: (i) data augmentation, (ii) invariance regularization loss and (iii) equivariant architectures. Although diverse, all these techniques seek to induce symmetries deemed appropriate in the final model (Fig. 1).

3.1. Data augmentation

Data augmentation is a ubiquitous technique when it comes to training neural networks. Its use is often motivated by prior knowledge that the function being approximated is invariant under a specific set of transformations (i.e., the transformations used are label-preserving). In other words, knowledge about the problem and the data acquisition process determines the operations used for augmentation. This technique is simple and frequently used to improve generalization (Shorten and Khoshgoftaar, 2019). In this subsection, we review the relationship between data augmentation and invariance and discuss some transformations useful in the context of breast cancer screening (see Fig. 2).

For a classification task, we formulate data augmentation as:

$$\mathcal{L}_{batch} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(f(T_{g_i} \circ I_i), y_i\right), \quad g_i \sim G \tag{2}$$

[1] https://github.com/edux300/symmetry-based-regularization-in-deep-breast-cancer-screening.

[2] We compiled the definition of groups and group actions as supplementary material (A.).

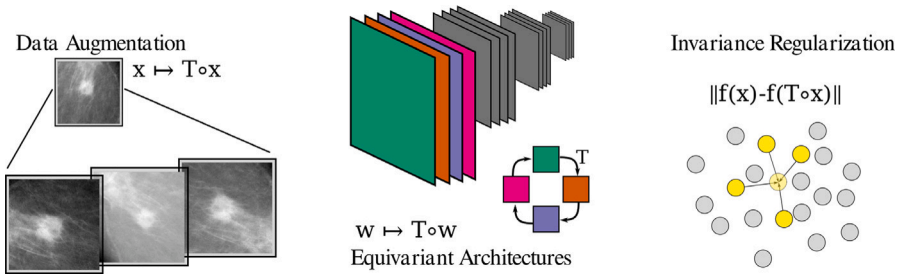


Fig. 1. This work focuses on different ways of baking symmetry into deep learning models. We promote or impose this property throughout the learning pipeline through changes in the input data (Section 3.1), model architecture 3.3 or loss function 3.2.

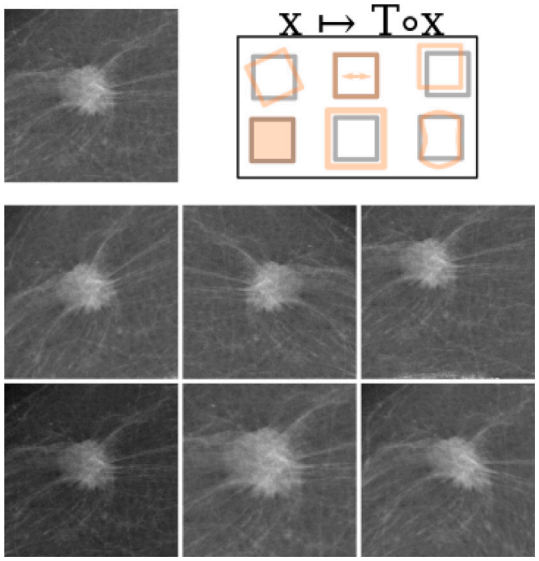


Fig. 2. Transformations studied in this work for data augmentation. The original mass image is in the top left corner. From left to right, in the first row rotation, reflection and translation are depicted. The second row shows contrast and brightness, scale and elastic transformations.

$\mathcal{L}$ is the loss function, typically cross-entropy, and $\mathcal{L}_{batch}$ is the loss for an entire batch. N denotes the batch size, f the output of the network, and $\{(I_1, y_1), \ldots, (I_N, y_N)\}$ the set of examples in the batch. Transformation T_{g_i} is sampled for each example (i.e., online data augmentation).

Loss minimization ensures that for the training data, the network output is approximately equal to the label, $f(I_i) \approx y_i$. When augmentation is used, the previous equality takes the form of $f(T_g \circ I_i) \approx y_i$, for any $g \in G$. In other words, data augmentation promotes invariance, under the definition of Eq. (1), to the set of transformations used, $\{T_g | g \in G\}$. It is thus essential to determine which operations should be used for augmentation and which should be avoided.

Different transformations are typically considered in mammography, depending on the task addressed. Rotations, reflections, and translations are frequently used for patch classification problems. Under the formulation of Eq. (2), these transformations are defined as:

- **Rotation:**

$$T_\theta \circ I(u_1, u_2) = I(c_\theta \cdot u_1 + s_\theta \cdot u_2, -s_\theta \cdot u_1 + c_\theta \cdot u_2) \tag{3}$$

where position u is separated in its two components u_1 and u_2, and c_θ, s_θ indicate the cosine and sine functions of angle θ.

- **Reflection:**

$$T_m \circ I(u_1, u_2) = I((-1)^m \cdot u_1, u_2) \tag{4}$$

where position u is separated in its two components u_1 and u_2, and $m \in \{0, 1\}$. Notice that horizontal reflections ($I(u_1, -1^m \cdot u_2)$) are not explicitly included since they correspond to a composition of a vertical reflection and a 180° rotation.

- **Translation:**

$$T_{\Delta u} \circ I(u) = (u + \Delta u) \tag{5}$$

where $\Delta u \in \mathbb{R}^2$ is the amount of translation.

The reader can easily verify that each of the operations described constitutes a group action. Their use as data augmentation is justified since their application to the input should not change the network's output. For instance, most local breast structures, including lesions indicative of breast cancer, do not have a particular orientation, which motivates rotation and reflection operations. Regarding translations, they do not remove the object of interest (e.g., mass) from the patch, provided they are small. This requirement for small translations, which differs from the definition in Eq. (5), is addressed later in this subsection. The same rationale is not valid for more global structures, such as the whole breast or the pectoral muscle, limiting the range of operations considered in whole image problems.

Some operations change the image appearance in ways that may correlate both with extraneous factors and breast cancer. Therefore, it is not clear whether invariance to these factors is a desirable property. We discuss and empirically evaluate three of such transformations.

- **Contrast and brightness** have been used in other image domains and often simulate different image acquisition conditions (i.e., exposure and light intensity). In mammography, factors such as radiation dose and breast density modify the contrast and brightness of the image. The presence of lesions also correlates with these quantities since these are usually bright, high contrast regions. The transformation is given by:

$$T_{(c,b)} \circ I(u) = c \cdot I(u) + b \tag{6}$$

where $I(u)$ is the image intensity at position u, and c and b are contrast and brightness values, uniformly sampled.

- **Scale** transformations increase or decrease the objects' size in the image. In mammography, they are motivated by the fact that the size of lesions may vary. Despite this, size is not independent of malignancy. Scaling is defined as:

$$T_s \circ I(u) = I(s \cdot u) \tag{7}$$

where the scale factor s is uniformly sampled. Interpolation is used to obtain the pixels' value for a fixed grid.

- **Elastic** transformations can also be used to expand the training data in the context of mammography. During the exam, the breast is compressed under two plates, stretching the tissues and allowing the radiologist to find abnormalities more easily. This process is not deterministic, as the original position of the breast and the amount of compression force applied in different regions can vary due to different factors (Mercer et al., 2013), which are external

to whether the patient has breast cancer or not. Elastic transforms have been previously used to model these changes (Castro et al., 2018) and are defined as:

$$T_{\Delta\mu} \circ I(u) = I(u + \Delta\mu(u)) \tag{8}$$

where $\Delta\mu$ is the displacement at each point u. This displacement is: (i) sampled at each point from a 2-dimensional uniform distribution, and then (ii) smoothed using a Gaussian filter with standard deviation σ. The displacement at the patch's central point is subtracted to all positions to ensure the final result is centered. Interpolation is used to obtain the pixels' value for a fixed grid. Elastic transformations have the potential to increase the dataset's variability. However, they need to be considered with caution since high displacement values can make images look unrealistic (i.e., out of domain). Also, although deformation can originate from compression forces, some deformation patterns can be indicative of malignancy (e.g. architectural distortions).

The parameters for the three transformations described above are sampled from bounded distributions, similarly to the translation operation. The combination of two transformations of the same type may lie outside the defined bounds. For instance, the sum of two translations inside the defined interval may result in a translation outside of it. This does not follow the closure requirement of the equivariance definition in Eq. (1) ($g, h \in G \Rightarrow g \circ h \in G$). In practice, this is an edge effect, and results that follow from the definition of these operations are valid within the defined bounds.

3.2. Invariance regularization loss

Data augmentation artificially expands the training dataset and promotes invariance to input transformations. The implicit assumption is that the relevant image content remains unchanged, and so should the output of the neural network. This subsection proposes a similar prior, imposed on the feature extraction process rather than just the output. The proposed regularization method takes the form of an additive term in the loss function.

For each training example, I_i, we define the average feature representation, $\overline{z_i}$, as:

$$\overline{z_i} = \mathbb{E}_{T_g}\left[f_{\text{inter}}(T_g \circ I_i)\right] \quad , \quad g \sim G \tag{9}$$

where f_{inter} is the normalized representation of the model at an intermediate layer. Notice that this quantity can be estimated by sampling K transformations from G and computing the average of the resulting feature vectors. We use $\hat{z}_i$ to refer to this estimate. Invariance is obtained when:

$$\|f_{\text{inter}}(T_g \circ I_i) - \overline{z_i}\| = 0, \forall g \in G \tag{10}$$

To encourage this property, we penalize the cosine distance between feature representations for different sampled transformations and $\hat{z}_i$:

$$\mathcal{R}(I_i, T_g) = \left(1 - \frac{\hat{z}_i^{\,T} f_{\text{inter}}(T_g \circ I_i)}{\|\hat{z}_i\| \cdot \|f_{\text{inter}}(T_g \circ I_i)\|}\right) \tag{11}$$

The use of the cosine similarity is a natural choice for comparing feature vectors. The choice of this metric is motivated by two additional factors: (i) loss functions based on the cosine similarity are common in the literature for similar tasks (Chen et al., 2020); and (ii) the use of unnormalized metrics (e.g., L2 distance) can be circumvented by neural networks by having small weights in the layer before the representation is taken and compensating with high weights in the layer immediately after. This mechanism allows for a small L2 distance for all examples in the dataset, independently of the features learned.

Since estimating $\hat{z}_i$ requires computing $f_{\text{inter}}(T_g \circ I_i)$ for K different input transformations, these inputs can also be used to compute the task-specific loss. For this, they should be passed by the remaining layers of the network, f_{remain} (such that $f_{\text{remain}}(f_{\text{inter}}(.)) = f$), and the loss function evaluated, as illustrated by Algorithm 1. By doing so, we can speed up the training process. With this in mind, the batch loss is changed to:

Data: x, y
Result: $\mathcal{L}_{batch}$
$\mathcal{L}_{batch} = 0$;
for x_i ***in*** x **do**
  $\hat{z}_i = 0$;
  for k ***in*** K **do**
    $g_k \sim G$;
    $z_{i,k} = f_{\text{inter}}(T_{g_k} \circ x_i)$;
    $\hat{z}_i$ += $(z_{i,k}/(K \times ||z_{i,k}||))$;
  end
  $\hat{z}_i = \text{no\_grad}(\hat{z}_i)$;
  for k ***in*** K **do**
    $\mathcal{L}_{batch}$ += λ.cos_dist$(z_{i,k}, \hat{z}_i)$;
    $\mathcal{L}_{batch}$ += $\mathcal{L}(f_{\text{remain}}(z_{i,k}), y_i)$;
  end
end
$\mathcal{L}_{batch}$ /= $(K \times N)$;

Algorithm 1: Example implementation for the invariance regularization. f_{remain} are the remaining layers of the network after f_{inter}. cos_dist denotes the cosine distance function.

$$\mathcal{L}'_{batch} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} \left[\mathcal{L}\left(f(T_{g_{i,k}} \circ I_i), y_i\right) + \lambda\mathcal{R}\left(I_i, T_{g_{i,k}}\right)\right] \tag{12}$$

The proposed formulation increases the batch size by a factor of K. In our experimental section, we adjusted N so that the effective batch size remains constant for a fair comparison of different methods. There is a conceptual difference between a non-regularized model and one trained with batches of repeated instances (Eq. (12) with $\lambda = 0.0$). Recent research (Hoffer et al., 2020) has focused on this and showed there are benefits to generalization when repeating examples with different augmentations within batches. The origin of these gains relates to the gradients of different samples being correlated within the same batch. In the experimental section, we appropriately quantify how both effects, (i) batch augmentation and (ii) invariance regularization, influence generalization.

3.3. Equivariant architectures

In neural networks, a set of operations (layers) is sequentially applied to the input to generate an output. Depending on their parameters, neural networks can encode different functions and thus can solve various problems. The model's architecture (sequence of layers) defines a prior on the set of functions that it can learn. Thus, some architectures may be better suited than others depending on the task at hand.

Convolutional Neural Networks (CNNs) are very effective for computer vision problems. The main difference of this large class of models is the use of the convolutional layers to map data, which are translation equivariant. In other words, CNNs recognize similar local patterns independently of their position in the image. This prior, along with other properties such as pooling, make these models better equipped to deal with natural signals (LeCun et al., 2015). Previous work has generalized this equivariance property to other transformations (Cohen and Welling, 2016). We use this methodology to generate architectures equivariant to rotation and evaluate them in breast cancer screening.

The G-convolution, $*_G$, is defined as:

$$[f *_G w](g) = \sum_{u \in \Omega} f(u) \cdot w(g^{-1} \circ u) \tag{13}$$

Here, f and w are functions defined on the set Ω. G is a group that acts on set Ω. Notice that the result of this operation is also a function, defined on G.

Eq. (13) is the standard 2D convolution if we consider G as all pairs of integers equipped with the sum operation, $G = (\mathbb{Z}^2, +)$. By considering other groups, we can define new convolutional layers, which are equivariant under different sets of transformations. To show this, consider the group action $T = \{T_b | b \in G\}$, which acts on f (and w) in the following way:

$$T_b \circ f(u) = f(b^{-1} \circ u) \tag{14}$$

In the supplementary material (B.) we show that $T = \{T_b | b \in G\}$ is indeed a group action. The G-convolution is equivariant under T:

$$\begin{aligned} T_b \circ [f *_G w](g) &= [f *_G w](b^{-1} \circ g) \\ &= \sum_{u \in \Omega} f(u) \cdot w((b^{-1} \circ g)^{-1} \circ u) \\ &= \sum_{u \in \Omega} f(u) \cdot w(g^{-1} \circ b \circ u) \\ &= \sum_{u \in \Omega} f(b^{-1} \circ u) \cdot w(g^{-1} \circ u) \\ &= \sum_{u \in \Omega} [T_b \circ f](u) \cdot w(g^{-1} \circ u) \\ &= \left[[T_b \circ f] *_G w \right](g) \end{aligned} \tag{15}$$

Notice that an equivalent proof could be used if f and w were functions defined on G (or any other set in which G acted on). In this work, two types of convolutions are considered:

$\mathbb{Z}^2$-convolution — This is the standard 2D convolution and serves as a baseline. As previously described, CNNs are equivariant to translations due to the use of this operation. Both the input and the resulting feature maps are functions of $\mathbb{Z}^2$.

$p4$-convolution — The group $p4$ includes translations and rotations of 90 degrees around the origin of the plane. Regarding rotation this is a cyclic group of size 4 ($C_4 = R_0, R_{90}, R_{180}, R_{270}$). Thus, $p4 = \mathbb{Z}^2 \times C_4$. We will consider inputs defined on two domains, $\Omega = \mathbb{Z}^2$, (e.g., input image), $\Omega = p4$, (e.g., feature maps resultant from applying this layer), and show how G acts on Ω in each case.

- $\Omega = \mathbb{Z}^2$
  Let $g = (v_1, v_2, \theta)$ and $u = (u_1, u_2)$. Then the group action is defined as:
  $$\begin{aligned} g \circ u = (&v_1 + c_\theta \cdot u_1 - s_\theta \cdot u_2, \\ &v_2 + s_\theta \cdot u_1 + c_\theta \cdot u_2) \end{aligned} \tag{16}$$
- $\Omega = p4$
  Let $g = (v_1, v_2, \theta)$ and $u = (u_1, u_2, \phi)$: Then the group action is defined as:
  $$\begin{aligned} g \circ u = (&\theta + \phi, \\ &v_1 + c_\theta \cdot u_1 - s_\theta \cdot u_2, \\ &v_2 + s_\theta \cdot u_1 + c_\theta \cdot u_2) \end{aligned} \tag{17}$$

In the supplementary material (B.) we show that the group actions defined above (G on $\mathbb{Z}^2$ and on $p4$) are indeed group actions.

A very common interpretation of the traditional convolution operation, is that the filter "slides" across all the positions in the image to generate the output. For the $p4$-convolution, the filter not only slides but also rotates over the image. As such, each $p4$-feature map is a function of position and orientation. Similarly, the neurons of standard CNNs share weights across spatial dimensions. For $p4$-CNNs, weight sharing happens across spatial dimensions and across an additional *orientation dimension* of size 4, where filters are rotated accordingly.

Regarding implementation, the $p4$-convolution on a $\mathbb{Z}^2$ input is equivalent to performing four convolutions, each one with a rotated version of the original filter. In that case, if the input were to be rotated, the output feature maps would be rotated and shifted in the *orientation dimension*. This equivalence between transformations done on the input and the output obeys the definition of equivariance. The $p4$-convolution can be directly implemented in current deep learning frameworks by stacking the rotated filters and using the standard 2D convolution as illustrated in Fig. 3. For inputs defined on $p4$ (e.g., after the first layer), the filters need not only be rotated but also shifted in the *orientation dimension* to account for the output shift mentioned above. Note that the described method is not equivalent to feeding multiple rotated copies of the input to the model and concatenating the output. Each $p4$-convolution has as input all the features of the previous layer in all orientations, in the same way, that a standard convolution has access to all the feature maps in the previous layer. When concatenating the output for multiple rotated copies, this is not true.

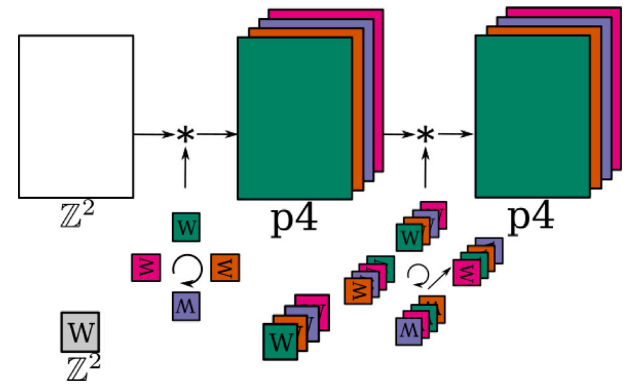


Fig. 3. Illustration of the filter transformations required to implement the $p4$-convolution using the standard convolution.

The composition of equivariant maps is still equivariant. Consequentially, as long as every operation in a network exhibits this property, the deep architecture as a whole is equivariant. We have seen how equivariance relates to the convolution layer. For other layers:

- Point-wise operations, such as ReLU, depend only on the value at each point. These functions are equivariant since applying a transformation to the input produces the same change in the output.
- Batch normalization maintains equivariance as long as weights and batch statistics are the same across orientations. In that case, it equals a scaling and offset operation for each $p4$-feature map. Thus, for inputs defined on $p4$, the *orientation dimension* should be treated as a spatial dimension.
- Pooling operations partially break equivariance due to the use of strides larger than 1, even for standard CNNs. Typically, a stride and kernel size of 2×2 is used. The resulting map retains equivariance to translations multiple of 2 (in each dimension), a subgroup of the original translation group. For feature maps defined on $p4$ the same loss of structure will happen. Rotation symmetry is unaffected. Interestingly, pooling across the *orientation dimension* leads to rotation invariance, a particular case of equivariance, which may be an appealing building block in some domains and is enabled by the use of the $p4$-convolution.

New architectures can be generated by substituting the standard 2D convolution with its $p4$ counterpart. However, some details are worth taking into consideration. Previous research has suggested that this prior is more relevant in the early layers of the network (Castro et al., 2020). As noted by Yosinski et al. (2014), many early filters resemble rotated copies of each other independently of the task they were optimized on. The introduction of the $p4$-equivariance prior is a way to brew knowledge into the network architecture rather than learn it. On the contrary, for features with no orientation (i.e., rotation

Table 1
Number of collected patches for the CBIS-DDSM and INbreast datasets.

Dataset	Set	Background	Benign	Malignant
CBIS-DDSM	Train	1950	583	544
	Valid.	343	122	101
	Test	565	207	139
		Background	Benign	Abnormal
INbreast	Test	401	28	88

invariant), the $p4$-convolution is time-inefficient since it will compute the same value four times.

Recent work (Dehghani et al., 2021) has highlighted the importance of adequately comparing deep learning models. We argue that, in CAD systems, where the model is trained once and runs potentially millions of times, time efficiency is more meaningful than the number of model parameters. Thus, although $p4$-convolutions lead to fewer parameters, we compare architectures based on their time complexity in the experimental setting.

3.4. Relationship between the proposed methods

The three methods discussed work by imposing or promoting equivariance in deep neural networks. The first two, data augmentation and invariance regularization, promote invariance under the desired set of input transformations. They differ in the stage of the network at which this property is sought. Differently, equivariant architectures enforce equivariance at all layers, but only transformations applicable to the network's filters can be used. Contrary to the previous methods, equivariant architectures are theoretically guaranteed to maintain equivariance for unseen data. The proposed methods induce different priors, can be used together, and are easily applicable within most frameworks in breast cancer screening.

4. Experimental settings

In the experimental part of this work, we evaluated the impact of the proposed methods on the model's ability to generalize to new data. We set up a patch classification problem considering three classes: background, benign masses, and malignant/abnormal masses. Two publicly available mammography datasets were used. The train/test splitting, preprocessing, and patch extraction are detailed below. In our experimental results we also consider a whole-image setting 5.6 where a third publicly available dataset was used.

CBIS-DDSM (Lee et al., 2017) is the largest publicly available dataset for developing breast cancer screening algorithms. This collection is an updated and standardized version of DDSM (Heath et al., 2000) divided into two subsets, *masses* and *calcifications*. For each, standardized splits for train and testing are provided. Each finding in the dataset is associated with a segmentation mask and its pathology (malignant or benign). Images were obtained from scanned film mammography. In total, the dataset contains 1566 patients and 3032 mammography images. The image height (px), width (px) and pixel size (μm) vary in the following ranges $[3721 - 7111]$, $[1546 - 5386]$, and $[42 - 50]$, respectively.

We consider the union of the two standard test sets (*masses* and *calcifications*) as our test set, resulting in 318 patients. The remaining patients were divided into train (85%) and validation (15%) using a stratified multi-label splitting algorithm (Szymański and Kajdanowicz, 2017), ensuring a more similar distribution between the two. The labels considered for this were (i) presence of masses, (ii) presence of calcifications, and (iii) malignancy. Notice that while both *masses* and *calcifications* subsets are used in this study, the latter is only used for the extraction of background patches, in regions with no annotated lesion.

We downscaled the images so that their height equals 1152 while maintaining the aspect ratio. This step ensures that a standard patch size of 224×224 is large enough to cover most of the mass annotations in the dataset. For larger masses, no adjustment in patch size was made. The pixel intensity was rescaled to the interval $[0, 1]$ at the image level. The breast was segmented and artifacts removed by keeping the largest object after using a binary threshold. Artifacts simultaneously close to the breast region and the image border remained after this operation. In order to remove them, the breast contour was smoothed and prolonged until the image border and pixels outside of it were set to zero.

At the model's input, a patch size of 224×224 was adopted, which is standard in the computer vision community. However, when sampling, a larger region was considered so that transformations, such as rotations and translations, did not require padding. A patch centered in each mass was taken. A background patch was also taken for each image by sampling a random point within the breast while ensuring no overlap with any lesion. For some images, the space occupied by lesions did not allow the extraction of the background patch. The total number of examples in each set is shown in Table 1.

The second dataset used was **INbreast** (Moreira et al., 2012). It contains 410 full-field digital mammography images along with precise lesion annotation. Image quality is superior to the CBIS-DDSM dataset, but the size is smaller. The height and width in pixels varies within the ranges $[3328 - 4084]$ and $[2560 - 3328]$, and the pixel size is 70μm. The whole dataset was used for testing. The procedure described for the previous dataset was followed with some exceptions. The images in INbreast do not have artifacts, so segmentation was done with binary thresholding only. One patch was taken from each mass and one background patch for each image (when possible). The annotations for malignancy in the INbreast dataset follow the standard BI-RADS. Masses were considered abnormal if the total assessment for that exam was a BI-RADS > 2. Notice that some benign lesions are still within this range, but this is the threshold at which screening patients undergo further examination. The number of examples for each class is also shown in Table 1.

Additionally, we also considered the **Chinese Mammography Database (CMMD)** (Cui et al., 2021; Cai et al., 2019). In this dataset, a total of 5202 images are available, out of which 3744 have either benign or malignant image-level annotation. The remaining images correspond to the collateral breast of some exams with no findings, and thus considered normal. The pixel size and image size are the same for all images, namely 94.1 μm and 2294×1914. We used the segmentation methodology for the INbreast dataset described in the previous paragraph to preprocess this data. Since no lesion-level annotations are available for the CMMD dataset, this data was only considered for the whole-image experiment in Section 5.6.

In all experiments, we evaluated four metrics: accuracy, balanced accuracy (average of recall obtained on each class), rocAUC, and F1-score. The definition and analysis of these metrics in imbalanced data are provided in the supplementary material (D.). Importantly, to adapt the rocAUC metric to multi-label classification, we followed a one vs. one approach between all pairs of classes. This formulation is more robust to imbalanced datasets. For the f1-score, one vs. rest was used since it does not suffer from the same issue. Each experiment was repeated five times to account for the randomness in neural network training, and the average and standard deviation were reported.

Unless otherwise stated, the ResNet50 (He et al., 2016) architecture was trained from scratch by minimizing categorical cross-entropy. He's initialization (He et al., 2015) was used. Class weights were used to address label imbalance, and for each class, set to $\frac{N}{|C|.N_c}$, where N is the total number of examples, N_c is the number of examples of class c, and $|C|$ the number of classes. The learning rate was set to 0.05, the weight decay to 5e−4, and momentum to 0.9. The batch size was set to 32 and the gradient accumulated over 4 steps, leading to an effective batch size of 128. The model was trained for 300 epochs. After this, the learning rate reduced 10-fold, and the model was trained for 60 additional epochs. At the end of each epoch, the best weights for each metric were kept. The inference was run separately for each metric using the best weights in the validation set.

Table 2
Metrics on CBIS-DDSM for models trained with different data augmentation schemes. The *conventional* scheme uses rotation, flips and translation. The *improv* uses transformations which, when used individually, improved all metrics. Namely: rotation, flips, scale and elastic. Results show the mean ± std over 5 runs.

Transform	Accuracy	Bal-Accuracy	rocAUC	F1score	Improves all
None	0.810 ± 0.007	0.692 ± 0.006	0.860 ± 0.007	0.693 ± 0.013	–
Rotation	0.840 ± 0.005	0.747 ± 0.013	0.896 ± 0.005	0.746 ± 0.009	✓
Flips	0.815 ± 0.006	0.710 ± 0.017	0.878 ± 0.005	0.708 ± 0.014	✓
Translation	0.791 ± 0.012	0.681 ± 0.013	0.855 ± 0.006	0.674 ± 0.016	✗
Intensity	0.796 ± 0.009	0.689 ± 0.007	0.866 ± 0.002	0.686 ± 0.013	✗
Scale	0.812 ± 0.008	0.696 ± 0.014	0.868 ± 0.009	0.702 ± 0.025	✓
Elastic	0.812 ± 0.011	0.711 ± 0.016	0.863 ± 0.007	0.702 ± 0.015	✓
Conventional	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	$\mathbf{0.775 \pm 0.011}$	–
Improv	$\mathbf{0.855 \pm 0.008}$	$\mathbf{0.781 \pm 0.006}$	$\mathbf{0.920 \pm 0.002}$	0.772 ± 0.006	–

Table 3
Parameters used for each transformation. $\mathcal{U}_{\text{min}}$ and $\mathcal{U}_{\text{max}}$ correspond to the bounds of the uniform distribution used to sample Δu in elastic deformations.

Transformation	Parameters
Rotation	$\theta \in [-180, 180]$
Flips	–
Translation	$\Delta x, \Delta y \in [-24, 24]$
Intensity	$c \in [0.5, 1.5], b \in [-0.5, 0.5]$
Scale	$s \in [0.75, 1.25]$
Elastic	$\mathcal{U}_{\text{min}} = -500, \mathcal{U}_{\text{max}} = 500, \sigma = 10$

Table 4
Metrics on CBIS-DDSM for models trained with the proposed invariance regularization loss using different values of λ. Results show the mean ± std over 5 runs.

$\mathcal{R}$	λ	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	0.775 ± 0.011
✓	0.0	$\mathbf{0.862 \pm 0.009}$	$\mathbf{0.790 \pm 0.009}$	0.914 ± 0.008	0.774 ± 0.004
✓	0.25	0.861 ± 0.004	0.789 ± 0.003	0.924 ± 0.003	0.782 ± 0.005
✓	1.0	0.861 ± 0.005	$\mathbf{0.790 \pm 0.003}$	$\mathbf{0.925 \pm 0.003}$	$\mathbf{0.786 \pm 0.007}$
✓	4.0	0.860 ± 0.012	0.787 ± 0.012	0.919 ± 0.007	0.779 ± 0.016

5. Results and discussion

The first part of this section follows the same structure as 3. We evaluate the effect of (i) **data augmentation**, (ii) **invariance regularization loss**, and (iii) **equivariant architectures** on generalization. In the second part, Section 5.4 evaluates the synergy between the different techniques; 5.5 performs a cross-dataset evaluation, and 5.6 extends the main conclusions of our work to a whole-image setting.

5.1. Data augmentation

We assessed how different transformations impact the model's correctness when used as data augmentation. For this, we considered the following settings: (i) no augmentation (*none*); (ii) only one transformation as data augmentation (*rotation, flips, translation, intensity, scale, elastic*); (iii) *conventional* augmentation, which includes rotations, flips and translations; and (iv) *improv*, which includes all the transformations that, when used individually, lead to a better model in all metrics. The parameters of each transformation are shown in Table 3. The results for this set of experiments are depicted in Table 2.

As expected, data augmentation can improve the model's performance across multiple metrics. When applied individually, this was verified for four of the six transformations, with *rotations* having a significantly higher impact than *flips, scale* and *elastic*. Translations and intensity changes were detrimental. This reflects the fact that the test set images are well controlled in terms of the position of the mass, contrast, and brightness. Altering these conditions increases the problem's difficulty on the training data, without real benefit for the testing data. Another possible contributing factor is the fact that modern-day CNNs are already well equipped to deal with these transformations. As discussed in Section 3.3, the standard convolution operation is translation equivariant (LeCun et al., 2015). Regarding intensity changes, batch normalization layers (Ioffe and Szegedy, 2015) normalize the distributions of the activations after each layer according to batch statistics. The adjustment after the first batch-normalization layer may cancel out the variation introduced by brightness and contrast changes. A mathematical argument for this is provided in the supplementary material (C.).

Combining multiple transformations further improves all metrics. The *improv* scheme slightly improves the model when compared to *conventional* augmentation (3 metrics out of four). The interaction between different types of transformations and a possible saturation effect may prevent this difference from being more significant.

5.2. Invariance regularization loss

The *conventional* data augmentation scheme from the previous section was used as a baseline. We then introduced the proposed invariance regularization method and assessed how it affects the evaluation metrics for different values of λ. We used the representation of the last layer before the model's output to compute the regularization loss term. We chose this layer as, in Resnet architectures, this is the first representation after the convolutional part of the model. $K = 4$ was used in all experiments. As seen in Section 3, the proposed method increases the number of iterations per epoch and, consequentially, reduces the total number of epochs required for model convergence. Therefore, optimization was reduced to 185 epochs for regularized models. Results are depicted in Table 4.

Globally, invariance regularization leads to more accurate models. Setting $\lambda = 0$ leads to a significant improvement in accuracy and balanced accuracy, which is in line with recent findings on the regularization effect of batch augmentation for general computer vision problems (Hoffer et al., 2020). Despite this initial improvement, further gains in rocAUC and f1-score can be obtained by increasing the value of λ to 0.25 and 1. At $\lambda = 4$, the model performance starts degrading, as the regularization loss term starts dominating the cross-entropy in optimization. Results show that the transformation invariance prior, promoted by invariance regularization, can further improve generalization after data augmentation. This suggests that the proposed method encodes a stronger prior than data augmentation alone. In Section 5.4 we investigate this further by measuring invariance at the networks' output.

5.3. Equivariant architectures

Depending on the computer vision problem, deep architectures perform differently due to their inductive biases. In this section, we evaluated how changing the architecture, using the $p4$-convolution, can impact model accuracy.

We used Resnet-50 as a base model. Two other architectures were generated: The first one, named $p4$, is obtained by substituting every

Table 5
Evaluation of different model architectures on the CBIS-DDSM dataset. The $\mathbb{Z}^2$ architecture (baseline) corresponds to the standard ResNet-50 model. The time column indicates the theoretical time taken for inference compared to the baseline. (mean ± std over 5 runs).

Architecture	No filt	Params	Time	Accuracy	Bal-Accuracy	rocAUC	F1score
$\mathbb{Z}^2$	32	2.6M	0.25×	0.846 ± 0.017	0.759 ± 0.019	0.910 ± 0.011	0.754 ± 0.018
	64	23.5M	1×	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	0.775 ± 0.011
	128	267.2M	4×	0.853 ± 0.008	0.768 ± 0.012	0.912 ± 0.006	0.763 ± 0.017
p4	32	0.6M	0.25×	0.858 ± 0.007	0.785 ± 0.021	0.915 ± 0.011	0.771 ± 0.029
	64	5.9M	1×	0.864 ± 0.010	0.788 ± 0.019	0.915 ± 0.008	0.785 ± 0.019
	128	66.8M	4×	0.858 ± 0.013	0.782 ± 0.020	0.921 ± 0.004	0.780 ± 0.022
hybrid	32	2.6M	0.25×	0.862 ± 0.003	0.794 ± 0.010	0.924 ± 0.003	0.791 ± 0.014
	64	23.3M	1×	0.862 ± 0.011	0.793 ± 0.017	**0.925 ± 0.007**	0.788 ± 0.018
	128	265.3M	4×	**0.870 ± 0.005**	**0.803 ± 0.006**	0.922 ± 0.004	**0.802 ± 0.008**

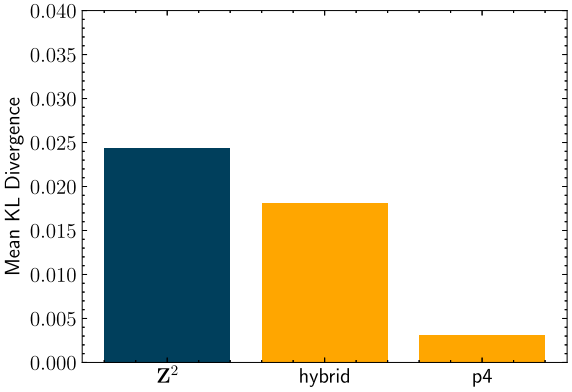


Fig. 4. Mean KL Divergence between outputs obtained for different transformations of the same input and their average. The test set of CBIS-DDSM was considered. Random $k \times 90°$ rotations were used as input transformations.

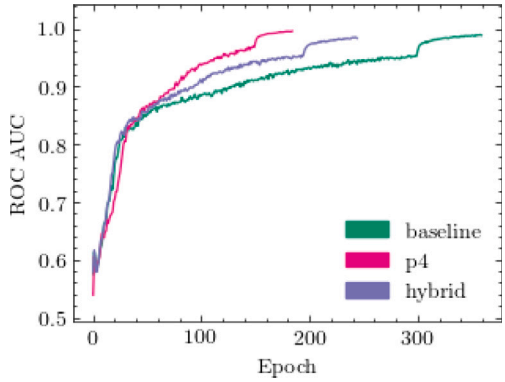


Fig. 5. Training rocAUC for different models (No filt = 64). As shown, equivariant models converge faster. The same plots for other metrics can be consulted on the supplementary material (E.).

convolution layer with the $p4$-convolution. Batch normalization was adapted as discussed in Section 3.3, and group pooling was used before the output layer. For the second architecture, named *hybrid*, we only changed the initial layer of the network and the convolutions in the first three residual blocks. The motivation behind this architecture is that the rotation equivariance prior may be more important in the initial layers of the network, as previously discussed. Batch normalization was adapted when it came after a $p4$-convolution. We also evaluated the impact of increasing (or decreasing) every layer's width by a factor of 2 for each architecture. Naturally, this leads to higher (or lower) inference and training times.

We confirmed that the introduction of $p4$-convolutions made models converge faster (as shown in Fig. 5). Consequently, we reduced the number of epochs to 245 for the *hybrid* and 185 for the $p4$ models. The same decrease in the baseline model led to worse results in all metrics. *Conventional* augmentation was used. The results for the different architectures are shown in Table 5.

Globally, model correctness is more determined by the architecture type than by the width of the convolutional layers. The increased capacity of wider models is not being efficiently employed, presumably due to a lack of data. Current CNN architectures have enough capacity to fit the data perfectly in settings with relatively small datasets, such as ours. Thus data efficiency plays a critical role.

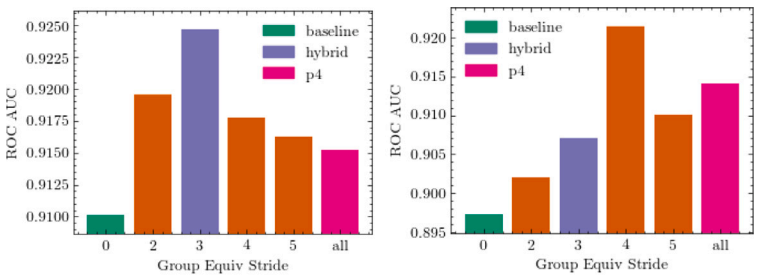
The $p4$ architecture compares favorably against the baseline, demonstrating that incorporating the rotation equivariance prior can benefit generalization. Together with the previous result on rotations for data augmentation, this provides evidence of the importance of rotational symmetry in mass classification. The *hybrid* model surpasses both the baseline and the $p4$ model, showing that the usefulness of the rotation equivariant layers is restricted to the early features of the network. These are often considered generic or task-independent, and small local patterns usually appear in different orientations (e.g., lines and corners). When moving to later layers in the architecture, features encode more abstract visual concepts. Here, the rotation equivariance prior appears to harm generalization. One possible explanation is that many of these more abstract features may not have a "preferred" orientation, and thus, using four channels to encode them is inefficient.

If we analyze the *hybrid* model, only a minority of the convolutional layers were changed. Despite this, the impact on the metrics is relatively high compared to the baseline. A key point in the design of this architecture is the reduction of the feature maps' resolution that happens for operations with stride higher than 1, namely convolution and pooling layers. In the lower parts of the Resnet-50 architecture, affected by the proposed architectural change, resolution decreases by a factor of 8. This is due to three out of the five operations that reduce the resolution in the architecture. Even though the *hybrid* model only uses a few $p4$-convolutions, they are the ones responsible for computing the low-level features.

The number of parameters is neither a good surrogate for model accuracy nor for the time taken per image. Also, it is unlikely that space to store model weights is a concern in a CAD system in a real-world scenario. Accuracy is presumably the most critical attribute, followed by time complexity. Although the $p4$ model has much fewer parameters, it is unlikely to be preferred in any scenario over the *hybrid* architecture which performs better across the board. Notice that as new architectures are introduced in breast cancer screening, the same principles can be used to adapt them to use the $p4$-convolution.

To better understand the importance of rotational symmetry, we measured how invariant the different architectures were to rotation. To this end, we computed the average KL divergence between outputs obtained for different input rotations and their average. We considered $k \times 90°$ rotations, as this is the set of transformations that $p4$ addresses. This was repeated for the whole test set. Results are depicted in Fig. 4.

As expected, the $p4$ model is almost entirely invariant. Edge effects account for slight differences between the outputs. Interestingly, even though the *hybrid* model only ensures equivariance in the early layers, the learned function is more symmetric than a model trained with data



(a) default: (0.05, 5e-4, 0.9) (b) new: (0.01, 1e-4, 0.8)

Fig. 6. Test rocAUC for different number of equivariant layers in the Resnet-50 model (average over 5 runs). The same experiment was conducted using two different protocols in the format (learning rate, weight decay, momentum). Increasing regularization in the optimization process (high learning rates, high weight decay, high momentum) seems to favor models with less equivariant layers. The same plots for other metrics can be consulted on the supplementary material (E.).

Table 6
Evaluation of combining multiple regularization strategies for the CBIS-DDSM dataset. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same setting except for learning rate, weight decay, and momentum. Respectively, these hyper-parameters were (0.05, 5e−4, 0.9) for the Resnet-50 and (0.01, 1e−4, 0.8) for the DenseNet-121 (mean ± std over 5 runs).

ResNet-50						
improv Aug.	Invariance Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	–	0.850 ± 0.005	0.778 ± 0.013	0.910 ± 0.007	0.775 ± 0.011
–	–	✓	0.862 ± 0.011	0.793 ± 0.017	0.925 ± 0.007	0.788 ± 0.018
✓	–	✓	0.873 ± 0.007	0.800 ± 0.013	0.928 ± 0.007	0.803 ± 0.005
✓	✓	✓	**0.875 ± 0.008**	**0.805 ± 0.012**	**0.930 ± 0.004**	**0.804 ± 0.011**
DensetNet-121						
improv Aug.	Invariance Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	–	0.837 ± 0.008	0.750 ± 0.019	0.904 ± 0.009	0.743 ± 0.019
✓	–	–	0.864 ± 0.008	0.773 ± 0.023	0.915 ± 0.005	0.785 ± 0.014
–	✓	–	0.861 ± 0.011	0.779 ± 0.018	0.917 ± 0.004	0.790 ± 0.009
–	–	✓	0.850 ± 0.003	0.767 ± 0.007	0.908 ± 0.004	0.765 ± 0.005
✓	✓	✓	**0.874 ± 0.011**	**0.803 ± 0.015**	**0.931 ± 0.003**	**0.797 ± 0.016**

augmentation only. We conclude that the proposed prior is stronger than data augmentation alone.

Finally, we conducted an ablation experiment with different equivariant architectures to evaluate at which point in the network equivariance to rotation no longer helps generalization. Each architecture, L, was obtained by substituting all layers with a total stride smaller or equal to 2^L. Under this definition, the *hybrid* model corresponds to $L = 3$. Two different optimization settings (learning rate, weight decay, momentum) were considered, the first one equal to the previous experiments and the second one (0.01, 1e−4, 0.8) having reduced learning rate, decay, and momentum. Results are depicted in Fig. 6. Although the use of equivariant layers seems to have an overall positive impact on the model, the ideal number of equivariant layers depends on the optimization settings. Weight decay, as well as the implicit regularization of large learning rates (Smith et al., 2021) and large momentum (Wang et al., 2022), are alternative ways of reducing overfitting and thus lower the impact of the proposed regularization approach. Despite this, equivariant models perform better than the baseline in both settings.

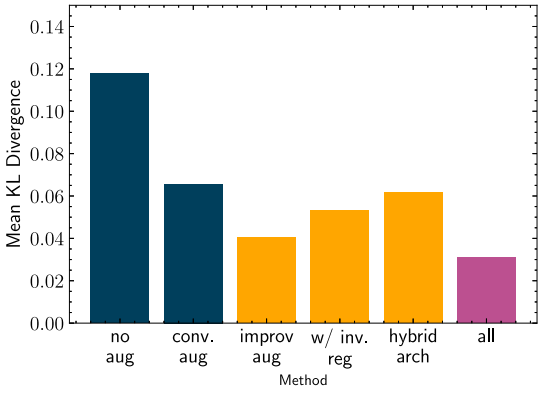


Fig. 7. Mean KL Divergence between outputs obtained for different transformations of the same input and their average. The test set of CBIS-DDSM was considered. Rotations, flips, scale and elastic transformations were used as input transformations.

5.4. Combining techniques

The various techniques considered in this work incorporate different priors into the network. This section evaluates the benefit of combining them in the same model. For this, we selected the top-performing settings in each set of experiments. The *hybrid* model was used, together with the *improv* data augmentation strategy, and invariance regularization ($\lambda = 1$). A first ablation study was conducted for the CBIS-DDSM dataset. The results are shown in Table 6. We also include the results for a different architecture, DenseNet-121 (Huang et al., 2017), which has been shown to perform well in mammography data by previous work (Wang et al., 2021). This model was trained with a learning rate of 0.01, weight decay of 0.0001, and a momentum of 0.8. Due to its similarity to Resnet-50, we used the $p4$-convolution in the same layers/blocks to obtain the *hybrid* architecture.

For the Resnet-50 model, the *improv* augmentation scheme improves the performance of the *hybrid* model in all metrics. Adding invariance regularization leads to further improvements. The best results are, therefore, obtained when combining the three techniques — *improv* augmentation scheme, invariance regularization, and a *hybrid* architecture. As more and more regularization methods are combined, the improvement becomes smaller for all metrics. The results extend to a new architecture, the DenseNet-121. In this case, the role of regularization is more significant. This can be attributed to the use of different optimization settings. In particular, high learning rates can be seen as

Table 7
Metrics for models optimized on CBIS-DDSM and evaluated on INbreast for three classes, {"Background", "Benign Mass", "Abnormal Mass"}. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same setting except for learning rate, weight decay, and momentum (mean ± std over 5 runs).

ResNet-50						
improv Aug.	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	–	0.773 ± 0.052	0.623 ± 0.022	0.839 ± 0.023	0.558 ± 0.022
✓	–	–	**0.888 ± 0.007**	0.702 ± 0.012	0.867 ± 0.019	0.688 ± 0.018
–	✓	–	0.819 ± 0.028	0.661 ± 0.025	0.832 ± 0.023	0.621 ± 0.034
–	–	✓	0.843 ± 0.020	0.700 ± 0.012	**0.873 ± 0.025**	0.667 ± 0.023
✓	✓	✓	0.882 ± 0.011	**0.705 ± 0.024**	**0.873 ± 0.028**	**0.694 ± 0.013**
DenseNet-121						
improv Aug.	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	–	0.827 ± 0.017	0.661 ± 0.027	0.846 ± 0.006	0.611 ± 0.026
✓	–	–	0.850 ± 0.013	0.674 ± 0.024	0.867 ± 0.014	0.635 ± 0.028
–	✓	–	0.856 ± 0.017	**0.719 ± 0.017**	0.870 ± 0.006	0.669 ± 0.016
–	–	✓	0.846 ± 0.008	0.699 ± 0.019	0.853 ± 0.010	0.647 ± 0.014
✓	✓	✓	**0.882 ± 0.009**	0.698 ± 0.029	**0.876 ± 0.008**	**0.681 ± 0.025**

Table 8
Metrics for models optimized on CBIS-DDSM (on the three class setting) and evaluated on INbreast on two classes, {"Background", "Mass"}. Models not trained with *improv* augmentation use the *conventional* strategy. Models were trained in the same setting except for learning rate, weight decay, and momentum (mean ± std over 5 runs).

ResNet-50						
improv Aug.	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	–	0.849 ± 0.041	0.859 ± 0.016	0.957 ± 0.007	0.718 ± 0.036
✓	–	–	**0.939 ± 0.004**	0.905 ± 0.008	0.958 ± 0.013	0.849 ± 0.020
–	✓	–	0.872 ± 0.029	0.884 ± 0.019	0.964 ± 0.005	0.766 ± 0.049
–	–	✓	0.891 ± 0.017	0.899 ± 0.006	0.964 ± 0.004	0.796 ± 0.016
✓	✓	✓	0.935 ± 0.005	**0.912 ± 0.006**	**0.966 ± 0.007**	**0.855 ± 0.008**
DenseNet-121						
improv Aug.	Inv. Reg.	*hybrid*	Accuracy	Bal-Accuracy	rocAUC	F1score
–	–	–	0.906 ± 0.008	0.889 ± 0.012	0.959 ± 0.005	0.866 ± 0.017
✓	–	–	0.927 ± 0.008	0.908 ± 0.007	**0.967 ± 0.005**	0.898 ± 0.011
–	✓	–	0.922 ± 0.017	0.904 ± 0.011	0.963 ± 0.002	0.893 ± 0.018
–	–	✓	0.915 ± 0.006	0.900 ± 0.007	0.958 ± 0.006	0.882 ± 0.009
✓	✓	✓	**0.947 ± 0.007**	**0.918 ± 0.007**	**0.967 ± 0.005**	**0.927 ± 0.006**

implicit regularization, which helps avoid local minima in the optimization landscape (Smith et al., 2021). The lower learning rate of the DenseNet-121 model optimization leads to a baseline model with less implicit regularization, and the impact of the proposed methodology is more considerable.

Fig. 7 depicts how sensible the model's output is to input transformations. The transformations shown to be useful in Section 5.1 were used (rotations, flips, scale, and elastic). We can see that when no data augmentation is used, the model's output is less robust to input transformation. *improv* augmentation and invariance regularization improve robustness against the *conventional* strategy. Notice that invariance regularization and *conventional* are trained using the same input transformations. Despite this, the proposed regularization method leads to more robust models, and so we conclude that this method is indeed a stronger prior, as the results of Section 5.2 suggested. The *hybrid* architecture does not significantly boost invariance under the considered set of transformations. Notice that: (i) architecture changes are only focused on $k \times 90°$ rotations, and (ii) the imposed prior is focused on the early layers and not on the model's output. Combining all strategies further boosts invariance, suggesting the effects of invariance regularization and *improv* accumulate.

5.5. Cross-dataset evaluation

We also evaluated the different methods on the INbreast dataset after being trained on CBIS-DDSM. Although images are from the same domain in this cross dataset evaluation, the acquisition conditions and quality significantly differ (Fig. 8). This setting is closer to the real-world scenario where a model is trained in one dataset and deployed to multiple clinics with different types of equipment. The only adjustment was to normalize the patches from the INbreast dataset so that their mean and standard deviation was the same as the training data.

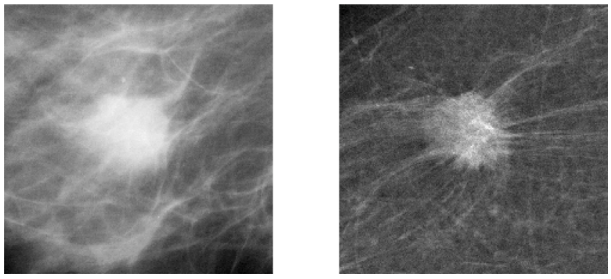


Fig. 8. Example of malignant masses on CBIS-DDSM (left) and INbreast (right). CBIS-DDSM images were acquired with scanned film mammography, while in INbreast full-field digital mammography was used. This is a more recent technique, which leads to images with better quality.

Two settings were considered:

1. Multiclass — with background, benign mass, and abnormal mass.
2. Binary — with background and mass. The model output was binarized by considering only the background class score and setting the probability of "mass" to be the opposite (i.e., 1 – P(background)).

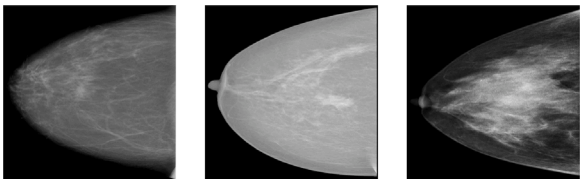


Fig. 9. Examples of images from different datasets in the whole-image experiment. From left to right: (i) CBIS-DDSM; (ii) INbreast; and (iii) CMMD.

Results are summarized in Tables 7 and 8. All the proposed strategies improve generalization on the INbreast dataset in both tasks and for both architectures. The relative influence of each strategy is different for the two architectures, but the *improv* augmentation appears to be the most impactful overall. When combined, the proposed methodology performs better in almost all metrics, demonstrating the benefit of combining all the techniques.

5.6. Evaluation in whole-image settings

An additional experiment was conducted to evaluate the proposed methodology in a scenario closer to a real-world application. For this, the whole breast is fed to the model, rather than just the region of interest (i.e., whole-image setting). We considered the case of weakly annotated data, where only the image-level labels are available for training (benign vs. malignant). For all datasets, images were resized to 800×800 after cropping the region containing the breast, as done in Shu et al. (2020), and the pixel intensity rescaled to the interval $[0, 1]$. In this experiment, INbreast examples are considered malignant if BIRADS is larger than 3 to keep consistency with the previous study. Examples from each dataset are provided in Fig. 9.

The same train/validation/test split described in Section 4 for the CBIS-DDSM dataset was used. For the INbreast and CMMD datasets, no standard test split is provided. Therefore we used 5-fold cross-validation to compare models. In each fold of CMMD, a validation split of 15% of the training data was used. This was not done for INbreast. Given the small size of this dataset, the validation split would be unrepresentative. All splits were done in a stratified fashion.

Similar to Shu et al. (2020), we use DenseNet169 as a backbone. An average pooling layer is used to aggregate the information from all input regions, followed by a linear layer for classification. The model was initialized with the pre-trained weights from ImageNet and fine-tuned to mammography data using the Adam optimizer with a learning rate of 2×10^{-5} and weight decay of 5×10^{-5}. The batch size was set to 16 and the gradient accumulated over 8 steps, leading to an effective batch size of 128. For the CBIS-DDSM and CMMD datasets, models were trained until rocAUC stopped improving in validation. For INbreast, the models were initialized with the final weights of the CBIS-DDSM experiment and optimized for a fixed number of epochs (300). This choice was based on the small size of this dataset. Since there are no pre-trained weights for the *hybrid* architecture, we trained this model in the ImageNet dataset using the same methodology as Huang et al. (2017).[3]

Data augmentation in the baseline model was done using rotations $\theta \in [-25, 25]$, small translations ($[-40, 40]$), and scaling ($s \in [0.8, 1.2]$). This model was then regularized by: (i) adding elastic transformations to the data augmentation pipeline; (ii) applying invariance regularization loss ($\lambda = 1$); and (iii) using an *hybrid* architecture. Architecture-wise, the first two blocks of the DensetNet architecture were converted to $p4$, along with the initial convolution and batch-norm layer. Accuracy and rocAUC are depicted in Table 9.

[3] After convergence, the model reached a top-5 error of 8.4% vs. 6.9% obtained with the original model. This difference is out of the scope of this paper, but we provide the values here for context.

Table 9
Accuracy and rocAUC for a whole-image models in three different datasets.

	Baseline		w/Regularization	
	Acc	AUC	Acc	AUC
CBIS	0.713	0.784	**0.750**	**0.812**
INbreast	0.844	0.828	**0.863**	**0.859**
CMMD	0.769	0.837	**0.779**	**0.850**

The baseline results are comparable to those obtained in Shu et al. (2020) for the CBIS dataset using the same methodology (DenseNet-169 with average pooling). Introducing symmetry-based regularization leads to higher accuracy and AUC for all datasets, demonstrating the potential of symmetry-based regularization in diverse settings. Notice that improved generalization was found even in a transfer-learning setting. Although the improvement was smaller for the CMMD dataset, it was still significant in a relatively large dataset of around 5k images.

6. Conclusion

The concept of symmetry is quintessential in the design of CAD systems. The ideal system should respond differently to input transformations depending on the application. This work shows how this general principle can be used to devise regularization strategies for breast cancer screening in mammography, where data efficiency is critical.

Three general approaches were followed: (i) data augmentation, (ii) invariance regularization loss, and (iii) equivariant architectures. These encode different priors on the functions learned by CNNs, based on a symmetry we want to induce in the network. Using the proposed methods for optimization is straightforward in most scenarios. The extensive evaluation showed that each approach improves generalization to unseen data. When combined, they further improve model robustness. These results were validated in different settings, including cross-dataset and whole-image scenarios, for different architectures and datasets.

Globally, including symmetry priors in the optimization of neural networks leads to better generalization and more robust models. We hope these principles can guide the development of more data-efficient methods for CAD in breast cancer screening and other medical imaging domains.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eduardo Castro reports financial support was provided by Foundation for Science and Technology.

Data availability

The data used is publicly available. The code used is available online with a link provided in the manuscript.

Acknowledgments

This work is financed by the ERDF — European Regional Development Fund, through the Norte Portugal Regional Operational Programme — NORTE 2020 under the Portugal 2020 Partnership Agreement, and by National Funds through the FCT — Portuguese Foundation for Science and Technology, I.P. on the scope of the CMU Portugal Program within project TAMI, with reference NORTE-01-0247-FEDER-045905, and within the PhD grant "*SFRH/BD/136274/2018*". The authors would also like to acknowledge NVIDIA for their generous donation of a TitanX gpu.

Appendix A. Notes on group and group actions

Groups

A group is a set equipped with a binary operation, in this work denoted as $\circ$, and which satisfies the four conditions below. Let G be a group. Then,

- Closure

  $g \circ b \in G, \qquad \forall g, b \in G$

- Associativity

  $(g \circ b) \circ a = g \circ (b \circ a), \qquad g, b, a \in G$

- Identity

  $\exists e \in G : \qquad e \circ g = g \circ e = g, \qquad \forall g \in G$

- Inverse

  $\forall g \in G \quad \exists g^{-1} \in G : \qquad g^{-1} \circ g = g \circ g^{-1} = e$

Group actions

A group G acts on a set X when there is a map $G \times X \to X$ such that the two conditions below are satisfied. The notation used in this work for this map is given by: $T_g \circ X$. Let G be a group that acts on X, then:

- Identity

  $T_e \circ x = x \qquad \forall x \in X$

- Compatibility

  $T_g \circ (T_b \circ x) = T_{g \circ b} \circ x \qquad \forall x \in X, \quad g, b \in G$

Appendix B. Group action proofs

If G is a group and X is a set, a group action is a function $G \times X \to X$, such that the following two axioms are satisfied:

1. Identity

   $e \circ x = x$

   e is the identity of G, $x \in X$
2. Compatibility

   $g \circ (b \circ x) = (g \circ b) \circ x$

   $g, b \in G, x \in X$

We now show that the group actions mentioned in this work satisfy these axioms.

General transformations

Consider the set of transformations $T = \{T_b | b \in G\}$, which acts on f in the following way:

$$[T_b \circ f](x) = f(b^{-1} \circ x)$$

1. Identity:

   $$[T_e \circ f](x) = f(e^{-1} \circ x) = f(e \circ x) = f(x)$$

2. Compatibility:

   $$\begin{aligned}[T_g \circ [T_b \circ f]](x) &= [T_b \circ f](g^{-1} \circ x) \\ &= f(b^{-1} \circ g^{-1} \circ x) \\ &= f((g \circ b)^{-1} \circ x) \\ &= [T_{g \circ b} \circ f](x)\end{aligned}$$

$p4$ on $\Omega = \mathbb{Z}^2$ and $\Omega = p4$

Consider the group $p4$. The group operation is defined as:

$$\begin{aligned}g \circ b = (&\theta_g + \theta_b, \\ &x_g + \cos(\theta_g).x_b - \sin(\theta_g).y_b, \\ &y_g + \sin(\theta_g).x_b + \cos(\theta_2).y_b)\end{aligned}$$

This is a group since:

- The identity element e exists and is equal to $(0, 0, 0)$.
- The above operation is associative. The first component is the sum of θ's, which is associative. For the second and third component consider the proof in Eq. (19).
- The inverse element takes the form of:

$$\begin{aligned}g^{-1} = (&-\theta_g, \\ &-\cos(\theta_g).x_g - \sin(\theta_g).y_g, \\ &+\sin(\theta_g).x_g - \cos(\theta_g).y_g)\end{aligned}$$

For $\Omega = \mathbb{Z}^2$, let $g = (x_1, y_1, \theta)$ and $u = (x_2, y_2) \in \Omega$, and the group action be defined as:

$$\begin{aligned}g \circ u = (&x_1 + \cos(\theta).x_2 - \sin(\theta).y_2, \\ &y_1 + \sin(\theta).x_2 + \cos(\theta).y_2)\end{aligned}$$

This group action can be defined in matrix form as:

$$\begin{pmatrix}\cos(\theta) & -\sin(\theta) & x_1 \\ \sin(\theta) & \cos(\theta) & y_1 \\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}x_2 \\ y_2 \\ 1\end{pmatrix} \tag{18}$$

1. Identity — with $e = (0, 0, 0)$, the identity element of group $p4$, the matrix in Eq. (18) becomes the identity matrix.
2. Compatibility:

$$\begin{aligned}g \circ (b \circ u) &= g \circ (x_b + \cos(\theta_b).x - \sin(\theta_b).y, \\ &\qquad y_b + \sin(\theta_b).x + \cos(\theta_b).y) \\ &= (x_g + \cos(\theta_g).[x_b + \cos(\theta_b).x - \sin(\theta_b).y] \\ &\qquad - \sin(\theta_g)[y_b + \sin(\theta_b).x + \cos(\theta_b).y], \\ &\qquad y_g + \sin(\theta_g).[x_b + \cos(\theta_b).x - \sin(\theta_b).y] \\ &\qquad + \cos(\theta_g)[y_b + \sin(\theta_b).x + \cos(\theta_b).y]) \\ &= ((x_g + \cos(\theta_g)x_b - \sin(\theta_g)y_b) \\ &\quad + \cos(\theta_g + \theta_b)x - \sin(\theta_g + \theta_b)y, \\ &\quad (y_g + \cos(\theta_g)y_b + \sin(\theta_g)x_b) \\ &\quad + \sin(\theta_g + \theta_b)x + \cos(\theta_g + \theta_b)y) \\ &= (g \circ b) \circ u\end{aligned} \tag{19}$$

For $\Omega = p4$, we have a group action on itself. It follows from the definition of a group that the group operation is a group action.

Appendix C. Relationship between contrast and brightness transformations and batch normalization

Considering the convolution parameterized by w, over an image, I, the result is given by:

$$[I * w](x) = \sum_{u \in \Omega} I(u) \cdot w(x - u)$$

When we apply contrast and brightness changes to the image, we have:

$$\begin{aligned}[(c.I + b) * w](x) &= \sum_{u \in \Omega} (c.I(u) + b) \cdot w(x - u) \\ &= c. \sum_{u \in \Omega} I(u) \cdot w(x - u) + \sum_{u \in \Omega} b.w(x - u)\end{aligned}$$

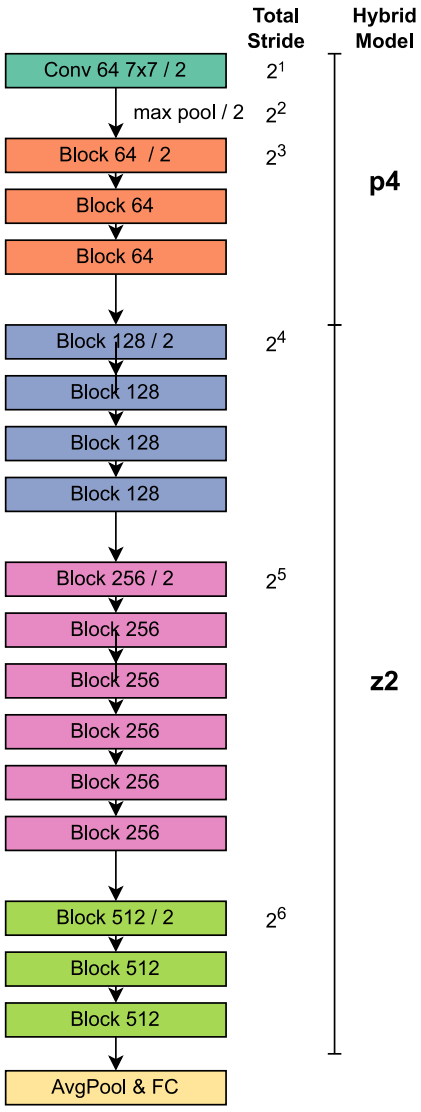


Fig. 10. Hybrid model structure.

$$= c.[I * w](x) + b.\sum_{u\in\Omega} w(x-u)$$

We can see that the changes in the output induced by the transformation do not depend on the input but only on c and the sum of the values of the convolutional weight multiplied by b, $b.\sum_{u\in\Omega} w(x-u)$. These changes in the activations' statistics are equal for all images in the batch. Thus, they are nullified by a batch normalization layer.

Appendix D. Experimental details — Metrics used

Four metrics were used to compare models in this work:

- **accuracy** is the most widely used metric in classification problems and corresponds to the proportion of correctly classified examples:

$$\sum_{i=1}^{N} \frac{\mathbb{I}[\hat{y}_i = y_i]}{N}$$

where $\mathbb{I}$ is an indicator function, $\hat{y}_i$ and y_i are the model prediction and label for ith sample, and N the number of samples.
- **balanced-accuracy** is the average proportion of correctly classified examples over the classes, contrary to the previous metric it weights every class equally, independently of the number of examples in each. This is often useful to assess the performance of models in imbalanced problems.

$$\frac{1}{|C|}\sum_{c\in C}\sum_{i=1}^{N_c} \frac{\mathbb{I}[\hat{y}_{c,i} = y_{c,i}]}{N_c}$$

where C is the set of classes in considered, N_c the number of examples for class c, and $\hat{y}_i$ and y_i are the model prediction and label for ith sample of class c.
- **rocAUC** score is the area and the curve of the ROC curve, often used for binary classifiers. In this work we used an extension for multi-class problems. It corresponds to the average AUC of each class (AUC_c). The average AUC of class c is found by averaging the AUC of the binary classifiers between class c and other classes (one vs. one). This is done to avoid large classes dominating the metric value. Concretely:

$$\text{AUC}_{c,k} = \frac{\sum_{i=1}^{N_c}\sum_{j=1}^{N_k}\mathbb{I}[s^c_{c,i} > s^k_{k,j}]}{N_c N_k}$$

$$\text{AUC}_c = \sum_{k\in C\setminus\{c\}} \frac{\text{AUC}_{c,k}}{|C|-1}$$

$$\text{AUC} = \sum_{c\in C} \frac{\text{AUC}_c}{|C|}$$

where $s^c_{k,i}$ is the classifier's confidence (score) that the ith sample of class k belongs to class c. We advocate for the use of this *one vs. one* formulation in our setting since, using a *one vs. rest* approach, AUC_c would lead to the biggest classes dominating in the formula. For instance, considering the classes "malignant", "benign" and "normal", where "normal" has an overwhelming support, the AUC of "malignant" vs. rest would be very close to the AUC of "malignant" vs. "normal". As such, the metric would benefit models that can distinguish "normal" from the rest in comparison to more equilibrate models. For mass classification, it is easier to classify "mass" (malignant or benign) vs. "normal" than to distinguish "malignant" and "benign" masses. As such, the numeric value of *one vs. one* was lower than *one vs. rest* in early experiments. In Appendix D.1 we provide an example showing how one class can dominate the computation of metrics that are sensible to class size.
- **F1-score** is the harmonic mean between the precision and recall of a classifier. We use a multi-class extension of the metric given by the average of f1-score for each class:

$$\text{Precision}_c = \frac{\sum_{i=1}^{N_c}\mathbb{I}[\hat{y}_{c,i} = y_{c,i}]}{\sum_{k\in C}\sum_{i=1}^{N_k}\mathbb{I}[\hat{y}_{k,i} = c]},$$

$$\text{Recall}_c = \frac{\sum_{i=1}^{N_c}\mathbb{I}[\hat{y}_{c,i} = y_{c,i}]}{N_c}$$

$$\text{F}_1 = \frac{1}{|C|}\sum_{c\in C} 2\frac{\text{Precision}_c.\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$$

D.1. Example — Metrics' sensibility to class size

Suppose we have a classification problem with three classes, A, B, C, where $N_A \gg (N_B + N_C)$ and $N_B = N_C$. A classifier is trained and its confusion matrix is given by:

		Actual		
		A	B	C
Predicted	A	1	0	0
	B	0	0.5	0.5
	C	0	0.5	0.5

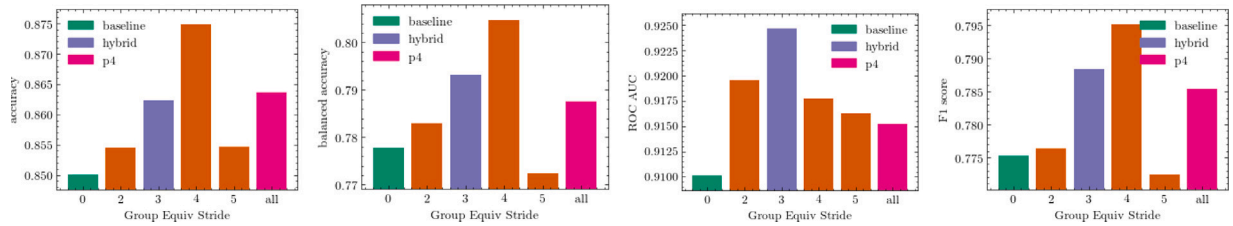
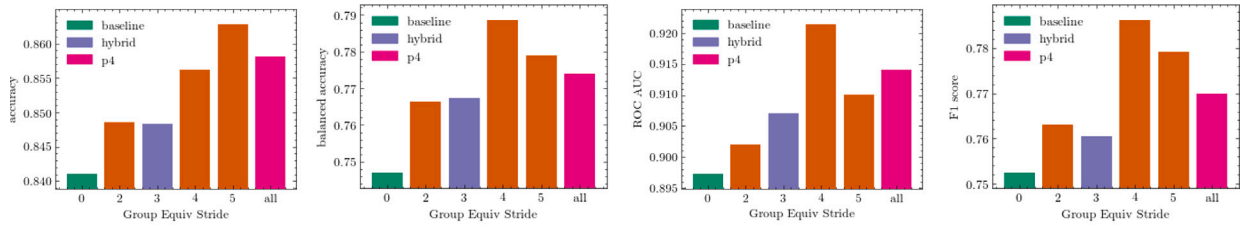
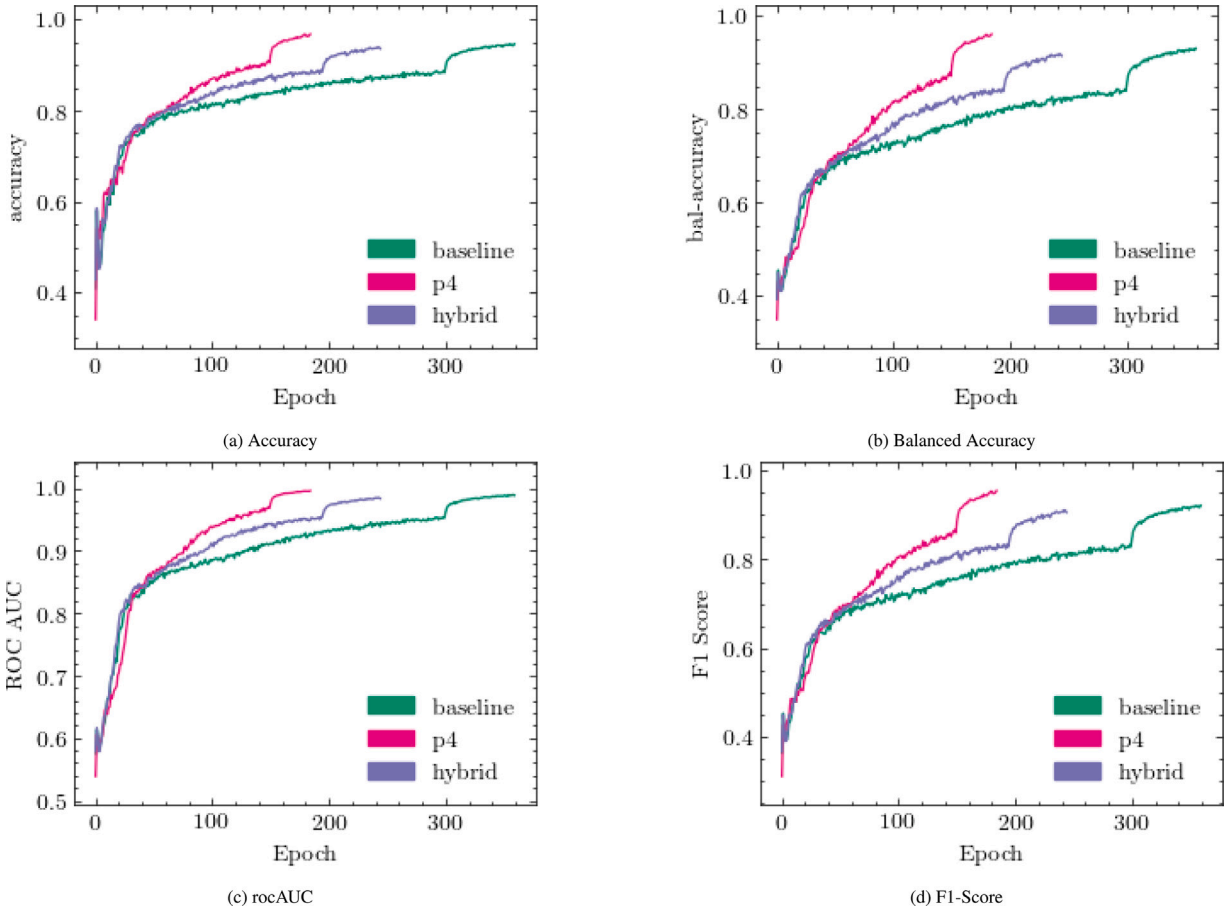


Fig. 11. Test metrics for different models architectures. Ablation experiments on the model architecture following the same experimental protocol as in the main paper.



(a) Accuracy (b) Balanced Accuracy (c) rocAUC (d) F1-Score

Fig. 12. Test metrics for different models architectures. Ablation experiments on the model architecture following a new experimental protocol (learning rate 0.01, weight decay 0.0001 and momentum 0.8).



(a) Accuracy (b) Balanced Accuracy

(c) rocAUC (d) F1-Score

Fig. 13. Training metrics for different model architectures (average over 5 runs).

This model perfectly distinguishes between "A" and "{B,C}", but does not distinguish "B" and "C". Because $N_A \gg (N_B + N_C)$, the model's accuracy is:

$$\sum_{i=1}^{N} \frac{\mathbb{I}[\hat{y}_i = y_i]}{N} \approx \frac{1N_A + 0.5N_B + 0.5N_C}{N_A} \approx 1$$

The balanced accuracy is:

$$\frac{1}{|C|} \sum_{k \in C} \sum_{i=1}^{N_k} \frac{\mathbb{I}[\hat{y}_{k,i} = y_{k,i}]}{N_k}$$
$$\approx \frac{1}{3} \left(\frac{1N_A}{N_A} + \frac{0.5N_B}{N_B} + \frac{0.5N_C}{N_C} \right) \approx \frac{2}{3}$$

As such, the balanced accuracy may be a more discriminative measure of model's performance if distinguishing B and C from each other is considered important. In summary, this metric gives more weight to small classes than standard accuracy. If we do the same analysis for the two rocAUC score settings, "one vs. one" and "one vs. rest", we get the following class scores:

Averaging class scores leaves us with AUCs of 0.8333 for the "one vs. one" case and 1 for the "one vs. rest". AUC with "one vs. one" may be a more discriminative measure for imbalanced classification problems. Regarding the F1-score, despite it being computed in a one vs. rest setting, it is still sensible to different degrees of "distinguishness" between B and C. The precision and recall for each class would be given by:

The average f1-score is $\frac{2}{3}$.

		$AUC_{c,k}$ k			AUC_c
		A	B	C	
c	A	–	1	1	1
	B	1	–	0.5	0.75
	C	1	0.5	–	0.75

(a) one vs. one

		AUC_c^{ovr}
c	A	$N_A(N_B+N_C)/N_A(N_B+N_C)=1$
	B	$N_B(N_A+0.5N_C)/N_B(N_A+N_C)\approx 1$
	C	$N_C(N_A+0.5N_B)/N_C(N_A+N_A)\approx 1$

(b) one vs. rest

	Precision	Recall
A	$1.N_A/(N_A+0N_B+0N_C)=1$	$N_A/N_A=1$
B	$.5N_B/(0N_A+.5N_B+.5N_C)=.5$	$.5N_B/N_B=.5$
C	$.5N_C/(0N_A+.5N_C+.5N_B)=.5$	$.5N_C/N_C=.5$

Appendix E. Additional results — Ablation for group equivariant architectures

As discussed, equivariant architectures can be generated from well-known models. For this, convolutional and other types of layers (e.g., batch normalization) need to be substituted by their equivariant counterparts. This can be done to the whole architecture (e.g., the p4 model in the main paper) or a set of early layers in the model (e.g., the *hybrid* architecture). The benefit of rotation equivariance is expected to be more significant at the beginning of the network. In this section, we provide additional experiments that help to characterize these architectures. Unless otherwise stated, they follow the experimental protocol described in the paper for mass classification in the CBIS-DDSM dataset.

We conducted an ablation experiment with different equivariant architectures to evaluate at which point in the network rotation equivariance no longer helps generalization. Each architecture, L, was obtained by substituting all layers with a total stride smaller or equal to 2^L. For instance, under this definition, the *hybrid* model presented in the paper corresponds to $L=3$, as shown in Fig. 10. The different metrics for each architecture are depicted in Fig. 11.

The architectures $L=2$, $L=3$, and $L=4$ are better than the baseline for all metrics. The same happens for the fully equivariant model p4. $L=4$ is better than the L = 3 (*hybrid*) architecture in 3 out of 4 metrics and worse in rocAUC. Unexpectedly, $L=5$ performs worse than the fully equivariant model in three out of four metrics. We suspect that different factors can contribute to this, including optimization difficulties. The same experiment was run again but with a different optimization protocol. Namely, we set the learning rate to 0.01, weight decay 0.0001, and used Nesterov momentum equal to 0.8. These changes were done to reduce the intrinsic regularization effect that the high learning rates, momentum, and weight decay have on learning. In this scenario (Fig. 12), all equivariant models beat the baseline. The $L=4$, $L=5$, and p4 architectures dominate across the four metrics. As there is less regularization intrinsic to the optimization protocol, models with more structure perform relatively better. In this experiment, the difference between the best and worse models was typically higher, but the models performed overall worse when compared to the previous optimization protocol.

Another attractive property of these models is that they require less time for optimization, as shown in Fig. 13. These architectures are simpler models in the sense that they have fewer parameters and more structure than their traditional counterparts. Rotation equivariance is brewed in the model's architecture rather than learned.

Appendix F. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2022.102690.

References

Agarwal, R., Díaz, O., Yap, M.H., Lladó, X., Martí, R., 2020. Deep learning for mass detection in full field digital mammograms. Comput. Biol. Med. 121, 103774.

Altobelli, E., Rapacchietta, L., Angeletti, P.M., Barbante, L., Profeta, F.V., Fagnano, R., 2017. Breast cancer screening programmes across the WHO European region: Differences among countries based on national income level. Int. J. Environ. Res. Public Health 14 (4), http://dx.doi.org/10.3390/ijerph14040452.

Alyafi, B., Diaz, O., Marti, R., 2019. DCGANs for realistic breast mass augmentation in X-ray mammography. URL: http://arxiv.org/abs/1909.02062. arXiv:1909.02062.

American Cancer Society, 2021. American Cancer Society. Cancer Facts & Figures 2021. American Cancer Society, Atlanta, pp. 1–72, 2021.

Arevalo, J., González, F.A., Ramos-Pollán, R., Oliveira, J.L., Lopez, M.A.G., 2016. Representation learning for mammography mass lesion classification with convolutional neural networks. Comput. Methods Programs Biomed. 127, 248–257.

Bahl, M., 2019. Detecting breast cancers with mammography: Will AI succeed where traditional CAD failed? Radiology 290 (2), 315–316. http://dx.doi.org/10.1148/radiol.2018182404.

Boot, T., Irshad, H., 2020. Diagnostic assessment of deep learning algorithms for detection and segmentation of lesion in mammographic images. In: MICCAI.

Cai, H., Huang, Q., Rong, W., Song, Y., Li, J., Wang, J., Chen, J., Li, L., 2019. Breast microcalcification diagnosis using deep convolutional neural network from digital mammograms. Comput. Math. Methods Med. 2019, 2717454.

Cardoso, J.S., Marques, N., Dhungel, N., Carneiro, G., Bradley, A., 2017. Mass segmentation in mammograms: a cross-sensor comparison of deep and tailored features. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). URL: http://publications/conferences/2017JaimeICIP.pdf.

Castro, E., Cardoso, J.S., Pereira, J.C., 2018. Elastic deformations for data augmentation in breast cancer mass detection. In: 2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI). pp. 230–234. http://dx.doi.org/10.1109/BHI.2018.8333411.

Castro, E., Pereira, J.C., Cardoso, J.S., 2020. Soft rotation equivariant convolutional neural networks. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN).

Cha, K.H., Petrick, N., Pezeshk, A., Graff, C.G., Sharma, D., Badal, A., Sahiner, B., 2019. Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. J. Med. Imaging 7 (01), 1. http://dx.doi.org/10.1117/1.jmi.7.1.012703.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (Eds.), Proceedings of the 37th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 119, PMLR, pp. 1597–1607, URL: https://proceedings.mlr.press/v119/chen20j.html.

Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Remote Sens. 54 (12), 7405–7415. http://dx.doi.org/10.1109/TGRS.2016.2601622.

Chidester, B., Zhou, T., Do, M.N., Ma, J., 2019. Rotation equivariant and invariant neural networks for microscopy image analysis. Bioinformatics 35 (14), i530–i537. http://dx.doi.org/10.1093/bioinformatics/btz353.

Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images using deep neural networks. In: Proc Medical Image Computing Computer Assisted Intervenction (MICCAI). pp. 411–418. http://dx.doi.org/10.1007/978-3-642-40763-5_51, arXiv:arXiv:1411.4389v3.

Cogan, T., Cogan, M., Tamil, L., 2019. RAMS: Remote and automatic mammogram screening. Comput. Biol. Med. 107, 18–29. http://dx.doi.org/10.1016/j.compbiomed.2019.01.024, URL: https://www.sciencedirect.com/science/article/pii/S0010482519300307.

Cohen, T., Weiler, M., Kicanaoglu, B., Welling, M., 2019. Gauge equivariant convolutional networks and the icosahedral CNN. In: ICML.

Cohen, T.S., Welling, M., 2016. Group equivariant convolutional networks. In: ICML 2016. pp. 2990–2999.

Conant, E.F., Toledano, A.Y., Periaswamy, S., Fotin, S.V., Go, J., Boatsman, J.E., Hoffmeister, J.W., 2019. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. Radiol.: Artif. Intell. 1 (4), e180096. http://dx.doi.org/10.1148/ryai.2019180096.

Cui, C., Li, L., Cai, H., Fan, Z., Zhang, L., Dan, T., Li, J., Wang, J., 2021. The Chinese mammography database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast.

De Sisternes, L., Brankov, J.G., Zysk, A.M., Schmidt, R.A., Nishikawa, R.M., Wernick, M.N., 2015. A computational model to generate simulated three-dimensional breast masses. Med. Phys. 42 (2), 1098–1118. http://dx.doi.org/10.1118/1.4905232.

Dehghani, M., Arnab, A., Beyer, L., Vaswani, A., Tay, Y., 2021. The efficiency misnomer.

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

Dumont, B., Maggio, S., Montalvo, P., 2018. Robustness of rotation-equivariant networks to adversarial perturbations. ArXiv abs/1802.06627.

Esteves, C., Allen-Blanchette, C., Makadia, A., Daniilidis, K., 2018. Learning so (3) equivariant representations with spherical cnns. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 52–68.

Fryback, D.G., Ph, D., Clarke, L., Zelen, M., Ph, D., Mandelblatt, J.S., Ph, D., Yakovlev, A.Y., Ph, D., Habbema, J.D.F., Ph, D., 2006. Effect of screening and adjuvant therapy on mortality from breast cancer: Commentary. Obstet. Gynecol. Surv. 61 (3), 179–180. http://dx.doi.org/10.1097/01.ogx.0000201966.23445.91.

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. AJR. Am. J. Roentgenol. 212 2, 300–307.

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 arXiv:1703.07047.

Graham, S., Epstein, D., Rajpoot, N., 2020. Dense steerable filter cnns for exploiting rotational symmetry in histology images. IEEE Trans. Med. Imaging 39 (12), 4124–4136.

Gromet, M., 2008. Comparison of computer-aided detection to double reading of screening mammograms: Review of 231,221 mammograms. Am. J. Roentgenol. 190 (4), 854–859. http://dx.doi.org/10.2214/AJR.07.2812.

Guan, S.Y., Loew, M., 2019. Breast cancer detection using synthetic mammograms from generative adversarial networks in convolutional neural networks. J. Med. Imaging 6 (3), http://dx.doi.org/10.1117/1.JMI.6.3.031411.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1026–1034. http://dx.doi.org/10.1109/ICCV.2015.123.

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 (CVPR).

Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P., 2000. The digital database for screening mammography. In: Proceedings of the Fourth International Workshop on Digital Mammography. http://dx.doi.org/10.1007/978-94-011-5318-8_75.

Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., Soudry, D., 2020. Augment your batch: Improving generalization through instance repetition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Houssami, N., Given-Wilson, R., Ciatto, S., 2009. Early detection of breast cancer: Overview of the evidence on computer-aided detection in mammography screening. J. Med. Imaging Radiat. Oncol. 53 (2), 171–176. http://dx.doi.org/10.1111/j.1754-9485.2009.02062.x.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269. http://dx.doi.org/10.1109/CVPR.2017.243.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pp. 1–11. http://dx.doi.org/10.1007/s13398-014-0173-7.2, Arxiv. URL: http://arxiv.org/abs/1502.03167. arXiv:1502.03167.

Jendele, L., Skopek, O., Becker, A.S., Konukoglu, E., 2019. Adversarial augmentation for enhancing classification of mammography images. pp. 1–14, URL: http://arxiv.org/abs/1902.07762. arXiv:1902.07762.

Kooi, T., van Ginneken, B., Karssemeijer, N., den Heeten, A., 2017a. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. Med. Phys. 44 (3), 1017–1027. http://dx.doi.org/10.1002/mp.12110.

Kooi, T., Litjens, G.J.S., van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R.M., den Heeten, A., Karssemeijer, N., 2017b. Large scale deep learning for computer aided detection of mammographic lesions. Med. Image Anal. 35, 303–312.

Lafarge, M.W., Bekkers, E.J., Pluim, J.P.W., Duits, R., Veta, M., 2021. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. Med. Image Anal. 68, 101849.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. http://dx.doi.org/10.1038/nature14539.

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

Lehman, C.D., Wellman, R.D., Buist, D.S.M., Kerlikowske, K., Tosteson, A.N.A., Miglioretti, D.L., for the Breast Cancer Surveillance Consortium, 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Med. 175 (11), 1828–1837. http://dx.doi.org/10.1001/jamainternmed.2015.5231.

Li, Y., Cao, G., Cao, W., 2020. A dynamic group equivariant convolutional networks for medical image analysis. In: Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020. pp. 1056–1062. http://dx.doi.org/10.1109/BIBM49941.2020.9313601.

Li, H.Y., Chen, D.D., Nailon, W.H., Davies, M.E., Laurenson, D.I., 2019. Signed Laplacian deep learning with adversarial augmentation for improved mammography diagnosis. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2019, PT VI, Vol. 11769. pp. 486–494. http://dx.doi.org/10.1007/978-3-030-32226-7_54.

Li, H., Chen, D., Nailon, W.H., Davies, M.E., Laurenson, D.I., 2021. Dual convolutional neural networks for breast mass segmentation and diagnosis in mammography. IEEE Trans. Med. Imaging PP (XX), 1. http://dx.doi.org/10.1109/TMI.2021.3102622, arXiv:2008.02957.

Li, X., Yu, L., Fu, C.-W., Heng, P.-A., 2018. Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis. Springer, pp. 235–243.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. http://dx.doi.org/10.1016/j.media.2017.07.005, URL: http://www.sciencedirect.com/science/article/pii/S1361841517301135.

Mednikov, Y.A., Nehemia, S., Zheng, B., Benzaquen, O., Lederman, D., 2018. Transfer representation learning using inception-V3 for the detection of masses in mammography. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 2587–2590.

Mercer, C., Hogg, P., Lawson, R., Diffey, J., Denton, E., 2013. Practitioner compression force variability in mammography: a preliminary study. Br. J. Radiol. 86 (1022), 20110596.

Mordang, J.-J., Janssen, T., Bria, A., Kooi, T., Gubern-Mérida, A., Karssemeijer, N., 2016. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In: Digital Mammography / IWDM.

Moreira, I., Amaral, I., Domingues, I., Cardoso, A.J.O., Cardoso, M.J., Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. Acad. Radiol. 19 2, 236–248.

Qi, K., Yang, C., Hu, C., Shen, Y., Shen, S., Wu, H., 2021. Rotation invariance regularization for remote sensing image scene classification with convolutional neural networks. Remote Sens. 13 (4), http://dx.doi.org/10.3390/rs13040569, URL: https://www.mdpi.com/2072-4292/13/4/569.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. Adv. Neural Inf. Process. Syst. 32 (NeurIPS), arXiv:1902.07208.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I., 2018. Detecting and classifying lesions in mammograms with deep learning. Sci. Rep. 8, http://dx.doi.org/10.1038/s41598-018-22437-z.

Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J., Schilling, K., Heywang-Köbrunner, S.H., Sechopoulos, I., Mann, R.M., 2019. Detection of breast cancer with mammography: Effect of an artificial intelligence support system. Radiology 290 (3), 305–314. http://dx.doi.org/10.1148/radiol.2018181371.

Schaffter, T., Buist, D.S.M., Lee, C.I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S., Feng, J., Feng, M., Kim, H.-E., Albiol, F., Albiol, A., Morrell, S., Wojna, Z., Ahsen, M.E., Asif, U., Jimeno Yepes, A., Yohanandan, S., Rabinovici-Cohen, S., Yi, D., Hoff, B., Yu, T., Chaibub Neto, E., Rubin, D.L., Lindholm, P., Margolies, L.R., McBride, R.B., Rothstein, J.H., Sieh, W., Ben-Ari, R., Harrer, S., Trister, A., Friend, S., Norman, T., Sahiner, B., Strand, F., Guinney, J., Stolovitzky, G., the DM DREAM Consortium, 2020. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw. Open 3 (3), e200265. http://dx.doi.org/10.1001/jamanetworkopen.2020.0265.

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

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.

Shu, X., Zhang, L., Wang, Z., Lv, Q., Yi, Z., 2020. Deep neural networks with region-based pooling structures for mammographic image classification. IEEE Trans. Med. Imaging 39 (6), 2246–2255. http://dx.doi.org/10.1109/TMI.2020.2968397.

Smith, S., Dherin, B., Barrett, D., De, S., 2021. On the origin of implicit regularization in stochastic gradient descent.

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: Cancer J. Clin. 71 (3), 209–249. http://dx.doi.org/10.3322/caac.21660, URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660. arXiv:https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660.

Szymański, P., Kajdanowicz, T., 2017. A network perspective on stratification of multi-label data. In: Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, Vol. 74.

Tardy, M., Mateus, D., 2021. Looking for abnormalities in mammograms with self- and weakly supervised reconstruction. IEEE Trans. Med. Imaging 40 (10), 2711–2722. http://dx.doi.org/10.1109/TMI.2021.3050040.

Tardy, M., Mateus, D., 2022. Leveraging multi-task learning to cope with poor and missing labels of mammograms. Front. Radiol. 1 (January), http://dx.doi.org/10.3389/fradi.2021.796078.

Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z., Jacobs, N., 2020. Inconsistent performance of deep learning models on mammogram classification. J. Am. Coll. Radiol. 17 (6), 796–803. http://dx.doi.org/10.1016/j.jacr.2020.01.006.

Wang, Y., Wang, Z., Feng, Y., Zhang, L., 2021. WDCCNet: Weighted double-classifier constraint neural network for mammographic image classification. IEEE Trans. Med. Imaging PP (XX), 1. http://dx.doi.org/10.1109/tmi.2021.3117272.

Wang, L., Zhou, Y., Fu, Z., 2022. The implicit regularization of momentum gradient descent with early stopping. ArXiv, arXiv:2201.05405.

Winsberg, F., Elkin, M., Macy, J., Bordaz, V., Weymouth, W., 1967. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. Radiology 89 (2), 211–215. http://dx.doi.org/10.1148/89.2.211, arXiv:https://doi.org/10.1148/89.2.211.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, 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., 2020. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging 39 (4), 1184–1194.

Wu, E., Wu, K., Cox, D., Lotter, W., 2018. Conditional infilling GANs for data augmentation in mammogram classification. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). In: LNCS, vol. 11040, pp. 98–106. http://dx.doi.org/10.1007/978-3-030-00946-5_11, arXiv:1807.08093.

Yaffe, M.J., Mainprize, J.G., 2004. Detectors for digital mammography. Technol. Cancer Res. Treat. 3 (4), 309–324.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS '14, MIT Press, Cambridge, MA, USA, pp. 3320–3328.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115. http://dx.doi.org/10.1145/3446776, arXiv:1611.03530.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., Xu, Z., 2020. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. Med. Imaging 39, 2531–2540. http://dx.doi.org/10.1109/TMI.2020.2973595.

Zhu, W., Qiu, Q., Calderbank, R., Sapiro, G., Cheng, X., 2019. Scaling-translation-equivariant networks with decomposed convolutional filters.