Computer Methods and Programs in Biomedicine 236 (2023) 107558

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Lightweight multi-scale classification of chest radiographs via size-specific batch normalization

Sofia C. Pereira [a,b,*], Joana Rocha [a,b], Aurélio Campilho [a,b], Pedro Sousa [c], Ana Maria Mendonça [a,b]

[a] Faculty of Engineering of the University of Porto, Portugal
[b] Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC), Portugal
[c] Hospital Center of Vila Nova de Gaia / Espinho, Portugal

ARTICLE INFO

Article history:
Received 15 November 2022
Revised 17 April 2023
Accepted 17 April 2023

Keywords:
Chest X-ray
Multi-label
Multi-scale
Deep learning
Ensemble

ABSTRACT

Background and Objective: Convolutional neural networks are widely used to detect radiological findings in chest radiographs. Standard architectures are optimized for images of relatively small size (for example, 224 × 224 pixels), which suffices for most application domains. However, in medical imaging, larger inputs are often necessary to analyze disease patterns. A single scan can display multiple types of radiological findings varying greatly in size, and most models do not explicitly account for this. For a given network, whose layers have fixed-size receptive fields, smaller input images result in coarser features, which better characterize larger objects in an image. In contrast, larger inputs result in finer grained features, beneficial for the analysis of smaller objects. By compromising to a single resolution, existing frameworks fail to acknowledge that the ideal input size will not necessarily be the same for classifying every pathology of a scan. The goal of our work is to address this shortcoming by proposing a lightweight framework for multi-scale classification of chest radiographs, where finer and coarser features are combined in a parameter-efficient fashion.

Methods: We experiment on CheXpert, a large chest X-ray database. A lightweight multi-resolution (224 × 224, 448 × 448 and 896 × 896 pixels) network is developed based on a Densenet-121 model where batch normalization layers are replaced with the proposed size-specific batch normalization. Each input size undergoes batch normalization with dedicated scale and shift parameters, while the remaining parameters are shared across sizes. Additional external validation of the proposed approach is performed on the VinDr-CXR data set.

Results: The proposed approach (AUC 83.27 ± 0.17, 7.1M parameters) outperforms standard single-scale models (AUC 81.76 ± 0.18, 82.62 ± 0.11 and 82.39 ± 0.13 for input sizes 224 × 224, 448 × 448 and 896 × 896, respectively, 6.9M parameters). It also achieves a performance similar to an ensemble of one individual model per scale (AUC 83.27 ± 0.11, 20.9M parameters), while relying on significantly fewer parameters. The model leverages features of different granularities, resulting in a more accurate classification of all findings, regardless of their size, highlighting the advantages of this approach.

Conclusions: Different chest X-ray findings are better classified at different scales. Our study shows that multi-scale features can be obtained with nearly no additional parameters, boosting performance.

© 2023 The Author(s). Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

1. Introduction

Computer-Aided Diagnosis (CAD) systems provide a second opinion and can assist doctors in their decision-making process, especially when the available human resources are scarce or inexperienced. Nowadays, many CAD systems use Deep Learning (DL) techniques applied to medical images [33].

* Corresponding author.
E-mail addresses: sofia.c.pereira@inesctec.pt (S. C. Pereira), joana.m.rocha@inesctec.pt (J. Rocha), campilho@fe.up.pt (A. Campilho), pedro.teixeira.sousa@chvng.min-saude.pt (P. Sousa), amendon@fe.up.pt (A.M. Mendonça).

https://doi.org/10.1016/j.cmpb.2023.107558
0169-2607/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

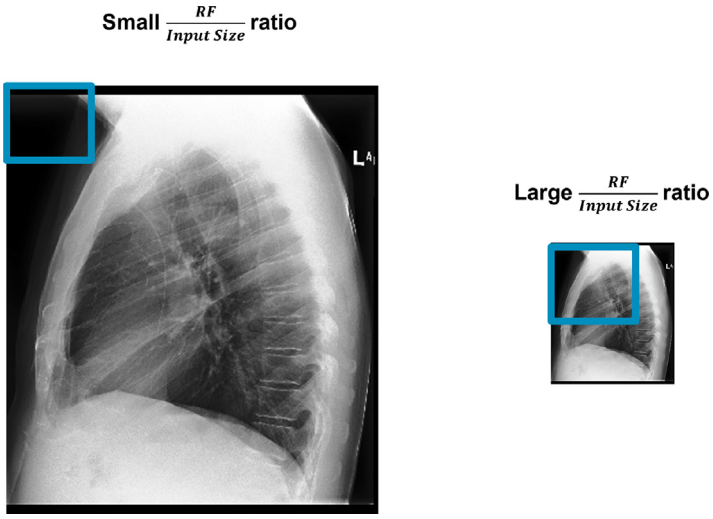


Fig. 1. The ratio between the size of the receptive field (RF) of a given CNN (illustrated by the blue square) and the input image size changes when the size of the input is adjusted. The receptive field has a fixed number of pixels. The granularity of the features extracted by the convolutions will increase as the size of the input images increases.

Convolutional Neural Networks (CNNs) are the most popular type of DL model used for processing and classifying image data. In most domain fields, CNNs take as input images of relatively small size. Some generic data sets frequently used in computer vision contain very small images (32 × 32 in the case of the CIFAR data set [12]). Others, such as ImageNet [4] and COCO [16], contain larger images. For example, the images of ImageNet have an average size of 469 × 387 pixels. Nevertheless, most pretrained off-the-shelf models are trained on images resized to 224 × 224 pixels. This size is considered to result in a good trade-off between computational cost and model performance.

Medical imaging has the particularity of employing large scale images. Histopathology images can be larger than 100,000 × 100,000 pixels [11], and X-ray images typically have a height/width of a few thousand pixels [30]. Simply re-scaling such large images to much smaller sizes may result in a significant loss of information. For this reason, even though studies from the medical domain often rely on models pretrained on images of smaller sizes (224 × 224) for weight initialization, it is common to fine-tune them to medical data using larger image sizes [26,29]. Off-the-shelf CNN architectures typically perform better when the input is provided within a certain size range. This happens due to the size of the Receptive Field (RF) in their convolutional layers [25]. The RF can be defined as the patch of the input image that fires a single cell of a feature map produced by a layer of a CNN [17]. Its size is fixed, and it depends on the size and number of filters in the CNN. For a given CNN with a pre-defined number of filters and filter sizes, smaller inputs will result in a RF that occupies a larger percentage of the image, while the opposite happens for larger images. This concept is illustrated in Fig. 1. Larger inputs result in finer features that are generated based on a smaller portion of the image, while smaller inputs result in coarser features that are based on a larger region of the input.

In Chest X-rays (CXRs), radiological findings vary significantly in size and shape. As a consequence, for a given CNN, the ideal input size will not necessarily be the same for classifying every pathology. For example, a cardiomegaly (enlargement of the heart) might occupy a very large portion of the input image, while a lung nodule can be significantly smaller (Fig. 2a and b). Multiple types of findings of different sizes commonly coexist in a single scan (Fig. 2c). Moreover, different instances of the same type of finding might have a different appearance among scans or even in the same scan. For example, a lung lesion can either be a small nodule or a large mass, and a single scan can contain multiple lesions (Fig. 2d).

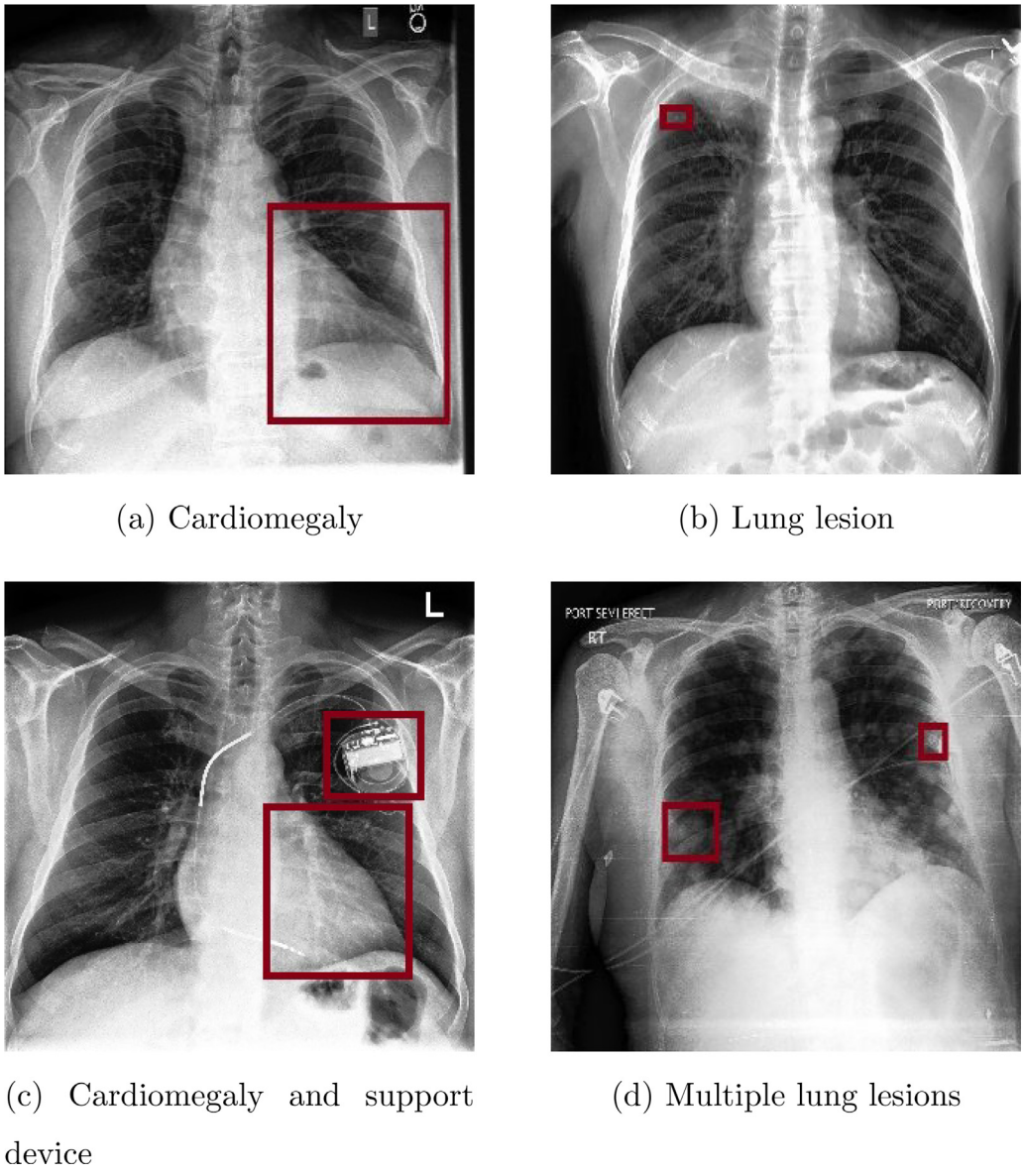
We developed a multi-scale ensemble model that takes advantage of several input image sizes to better characterize each data point, achieving better performance than any of the implemented single-scale baselines. We propose a new Size-Specific Batch Normalization (SSBN) layer, inspired by the domain-specific batch normalization layer used in multi-task learning [3]. When compared to ensembles using one individual CNN for each input size (3× the number of single-scale baseline parameters), ensembles based on SSBN rely on a significantly smaller number of parameters (1.02× the number of single-scale baseline parameters, in the case of the Densenet-121 architecture), while maintaining similar performance. We also explored other parameter-sharing settings, namely, sharing a variable number of initial layers of a CNN across input sizes, resulting in models known as TreeNets [13].

1.1. Related work

1.1.1. Multi-scale approaches

The fact that image size affects the performance of deep-learning based computer vision systems is well studied [25]. For this reason, many try to investigate ways to incorporate multi-scale information into their models. This is transversal to several domain applications, such as crowd counting [31], remote sensing [28] and medical imaging [8].

Although accounting for the multi-scale nature of the discriminant image features is a standard practice for object detection tasks, where ground truth bounding boxes are available [24], it is not as common for image classification tasks, where only one label for the entire image is provided. One way of extracting multi-scale features is to use dilated convolutions [31] via spatial pyramid pooling modules. However, this operation may lead to poor spatial consistency, which ultimately results in gridding artifacts [7]. Another approach is to build one classification model per size and combine their outputs using an ensemble technique, such as boosting [28]. The approach taken by Lim and Yalim Keles [15] is close to ours, consisting of a triplet CNN to generate multi-scale features by feeding different input sizes to each component. Surprisingly, all the parameters are shared across components, meaning that in practice only one CNN is used.



(a) Cardiomegaly (b) Lung lesion

(c) Cardiomegaly and support device (d) Multiple lung lesions

Fig. 2. Chest radiological findings differ in size. Some findings are very large (a), while others are smaller (b). More than one finding can co-exist in a single scan (c) and, in some cases, there can be multiple different-looking instances of the same type of finding in one scan (d).

Specifically for multi-label chest radiograph analysis, the fact that the performance for each label is affected differently with changes to the input size has been observed in [27]. Given this observation, [8] take advantage of this behavior and build a deep model for each scale. After learning each deep expert, they learn how to best combine the separate models by learning weights to perform a weighted average of the predictions in a second training phase.

1.1.2. TreeNets

Model ensembling is usually done in a *post-hoc* manner. However, it is known that the initial layers of a CNN extract very simple and generic features (such as edges or corners). This means that early layers of a CNN can potentially be shared across models if they are trained in an end-to-end manner. TreeNets, introduced by Lee et al. [13], explore this concept by showing that between a single CNN and an ensemble of several fully independent models, there is a spectrum of possible architectures that leverage parameter sharing.

The extent to which parameters are shared usually results in a trade-off between the number of parameters/computational efficiency and model performance. TreeNets allow a reduction in the number of parameters of the model, frequently without a decrease in performance. In some cases, performance might actually increase due to a reduction in overfitting. Multiple studies have adopted TreeNet structures in their models [14,19,35]. We extrapolate this structure, commonly used for multi-task learning or traditional ensembling, to the proposed multi-scale setting.

1.1.3. Batch normalization

Before feeding data into a neural network, a normalization step is carried out to ensure that the features are calibrated, resulting in faster and often better convergence. However, when networks are very deep, the distribution of the data changes during training as successive operations are performed, a problem known as internal covariate shift [9]. This problem is addressed in most modern CNN architectures by re-normalizing the input of each layer of a neural network using mini-batch statistics, in a process known as batch normalization. During training, the mean μ_B and variance σ_B^2 of a

mini-batch $B = \{x_1, x_2, \ldots, x_m\}$, where x_m corresponds to mth image of the batch, are calculated in each batch normalization layer (usually placed after each convolutional layer) and used to normalize the data, according to Eq. (1), where ϵ is a value added to the denominator for numerical stability. Additionally, the normalized data $\hat{x}$ is scaled and shifted using two learnable parameters β and γ, resulting in y (Eq. (2)). During training, the moving average of the mean and variance are registered (as non-learnable parameters) and the β and γ parameters are learned, being later used for batch normalization during the inference stage.

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{1}$$

$$y_i = \gamma \hat{x}_i + \beta \tag{2}$$

Batch normalization has proved to be extremely important in deep learning by making deeper models much faster to train and less prone to divergence. Numerous studies have focused on exploring the importance of this layer. It is possible to achieve high performance in standard benchmark computer vision tasks performing exclusively affine transforms (namely, shifting and rescaling) on random features, by training only the batch normalization layers in a network and leaving the remaining (randomly initialized) layers untrained [5]. The importance of batch normalization can also be highlighted in multi-task learning scenarios, where models that have individual batch normalization layers for each task but share all other parameters of the feature extractor can achieve good performance [3]. We adapt this last concept to multi-scale data by creating a model that shares all convolutional layers across tasks, but contains distinct batch normalization layers per input size.

1.2. Contributions

In this article, we propose alternatives for combining multi-scale information for multi-label classification of chest radiographs. The contributions of this work can be summarized as follows:

- We made a detailed per-label analysis of how the performance is affected when the input size is changed. Specifically, we experimented with the input sizes 224 × 224, 448 × 448 and 896 × 896.
- We developed a lightweight end-to-end ensemble model that leverages multiple input sizes, boosting performance. Additionally, we also explored the use of TreeNets to reduce computational cost and examine the trade-off between performance and the number of parameters.
- We proposed a new scale-specific batch normalization layer that applies batch normalization to the data separately for each input size, resulting in per-size personalized trainable (scale and shift) and non-trainable (mean and standard deviation) parameters. The multi-scale model built using this layer achieves a performance similar to that of a baseline ensemble that uses one deep expert per size, while relying on a much smaller number of parameters.
- We showed that the proposed approach generalizes well in an external data set, maintaining its superiority in terms of predictive performance and robustness, when compared to single-scale baselines.

2. Methodology

In this section, we start by describing the data used in the experiments and how it was preprocessed (Section 2.1). Then, we detail the experiments and models that were implemented (Section 2.2) and describe their training procedure (Section 2.3).

Table 1
Frequency of the labels in the selected CheXpert subset.

Label	Frequency
No Finding	22,271
Enlarged cardiomediastinum	7106
Cardiomegaly	17,851
Lung opacity	53,190
Lung lesion	5621
Edema	35,756
Consolidation	8716
Pneumonia	3914
Atelectasis	23,749
Pneumothorax	15,069
Pleural Effusion	55,837
Pleural Other	1982
Fracture	6037
Support Devices	74,012

2.1. Data set and preprocessing

The CheXpert [10] data set, a large publicly available CXR collection, is used in the experiments. It is a multi-label data set of 223,414 frontal and lateral chest X-ray scans from 64,540 patients labeled for the presence of twelve radiological findings/pathologies, plus “support devices” and an additional “no finding” label. Originally, the data set contains a training set labeled using Natural Language Processing (NLP), where each pathology is labeled as *present*, *absent* or *uncertain*. Additionally, the data set also contains a hand-labeled validation set with 234 images (200 patients) and a hidden hand-labeled test set (the images are not made publicly available) with 668 images (500 patients). The validation and test sets are very small compared to the training set and their label distribution is not representative for every pathology [10], both in terms of absolute and relative frequencies. Specifically, eight out of the fourteen labels contain less than 50 positive instances. Considering that the performance of the NLP labeler is very good (specifically, 0.969, 0.952 and 0.848 micro-F1 scores for entity mention, entity negation and entity uncertainty, respectively [10]) and that our goal is to access the effect of image size across all pathologies, we decided to rely solely on the training set and perform five-fold cross validation using 80-20 stratified splits without patient overlap. To further ensure the integrity of the data, the instances containing at least one pathology labeled as *uncertain* were removed. Ultimately, the filtered data set contained 138,358 instances. Table 1 shows the frequency of each label in the filtered data.

For further analysis of the proposed model, we performed an external validation using the publicly available version of the VinDr-CXR data set [21], made available on the Kaggle platform.[1] This data set contains 15,000 CXRs from two hospitals in Vietnam and was fully annotated by radiologists. Since the fourteen annotated labels made available do not fully overlap with those from CheXpert, we limit our analysis to the set of overlapping labels between data sets: No Finding, Cardiomegaly, Lung Opacity, Lung Lesion (equivalent to Nodule/Mass), Atelectasis, Consolidation, Pleural Effusion and Pneumothorax. Since each scan is annotated by three radiologists, we use majority voting for generating the final label set, following the approach taken in the original paper describing the data [21].

All scans were first resized to a height of 1,024 pixels (preserving aspect ratio) and then center cropped to 896 × 896, which is the largest input size used in our experiments. For each image, two lower-sized copies were generated (448 × 448 and 224 × 224 pixels). This way, the starting point for the analysis was an image

[1] https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection

size of 224 × 224 pixels, which is widely used in computer vision applications, including for pretraining models on the ImageNet [4] data set. All the images were standardized using ImageNet's mean and standard deviation. Random horizontal flips (with a probability of 1%) and random rotations (in the ±20° range) were performed to augment the data set. To make our CXR data set compatible with ImageNet pretrained weights, we copy the grayscale channel two times to make 3-channel images.

2.2. Experiments

The explored architectures are all based on a Densenet-121 model pretrained on the ImageNet data set, which is the model used by the authors of CheXpert [10] and others [1,22], due to its good performance on this data and task. We start by creating individual baselines (standard Densenet-121 models) for each input size (Section 3.1). The baseline models are then compared with the multi-scale ensemble models. These include the baseline approach of using one individual CNN per scale where only the final fully-connected layer is shared) (Section 3.1), several TreeNet alternatives (Section 3.2) and the proposed SSBN model (Section 3.3). Furthermore, we compared the proposed model to each single-scan baseline in an additional external validation set (Section 3.3). In all cases, the outputs obtained for each size are averaged to obtain the final probabilistic prediction.

Figure 3 illustrates the concept of parameter sharing in TreeNets and shows all the sharing scenarios that were considered in this study, ranging from a fully joint model to a fully separate model. Note that the fully separate model from the figure corresponds to the baseline ensemble approach that uses one individual expert per scale. The linear layer that performs classification after the feature extraction process is always shared across the individual models of all multi-scale ensembles. Even though this might not work well under multi-task settings where the tasks of each individual model are very different in our setting, the performed classification task is the same across ensemble members (only the scale varies). Considering that sharing this layer results in less parameters, we adopted this approach.

The SSBN model is a Densenet-121 model where all the regular batch normalization layers are replaced with SSBN layers, illustrated in Fig. 4. This way, each input size will have its own batch normalization layers and thus undergo a set of affine transforms optimized for that size alone, while sharing the remaining parameters of the network. For every batch of data, the three resolution versions of the data were run through the model, using their own set of batch normalization layers. We take the network's output for each resolution and average them. During backpropagation, the shared parameters are updated by all three versions, while each BN layer inside the SSBN layers is only updated by its corresponding version of the data. This results in a parameter-efficient, multi-scale ensemble. The number of parameters of all the models are shown in Table 2.

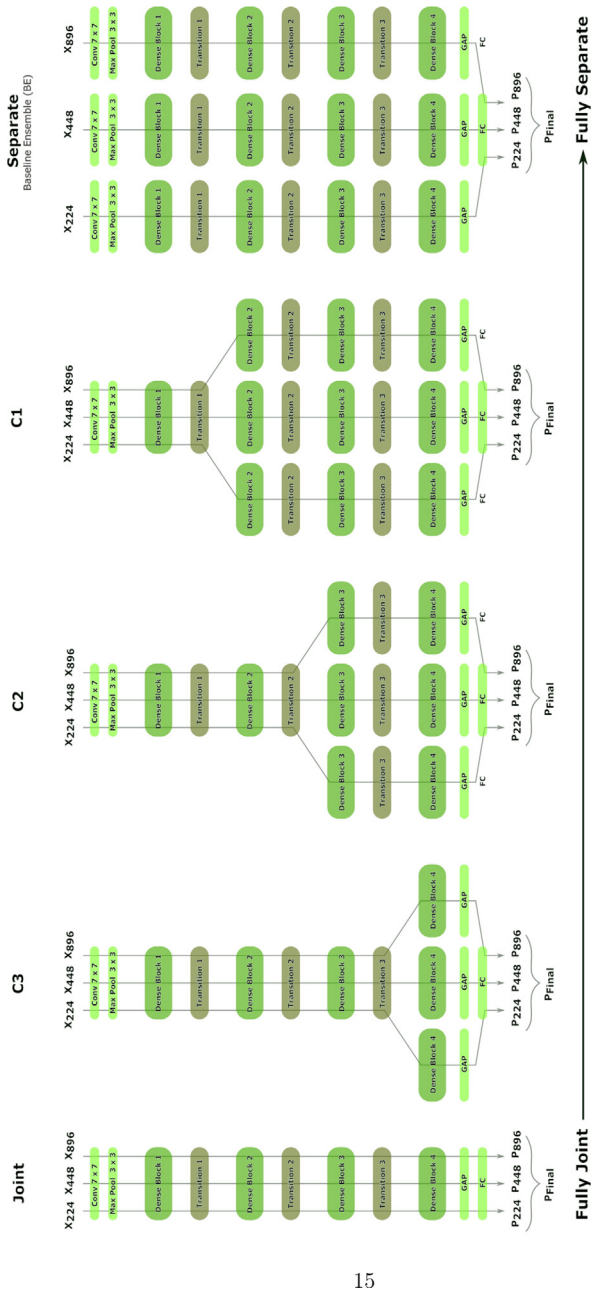


Fig. 3. TreeNet structures used in this study. Between a fully joint (Joint) model and fully separate (BE) model, there are three TreeNet structures that differ in their branching point, which can be after the third (C3), second (C2) or first (C1) block of a DenseNet-121 structure. The gray arrows indicate the layer trajectory of each input size.

2.3. Training details

All the models were trained for ten epochs, which is enough to reach convergence, using the standard binary cross-entropy loss, a batch size of 32 and a learning rate of 10^{-4}, which was empirically determined to optimize training. The learning rate is reduced by a factor of ten after five epochs. We select the weights from the epoch with the lowest validation loss. The models were created using the PyTorch framework and run on a cluster of two NVIDIA Titan RTX 24 GB Graphics Processing Units (GPUs). Automatic mixed precision [18] was used to accelerate training and reduce memory requirements.

3. Results

In this section, the results of the performed experiments are presented. First, the results of the single-scale baselines are compared to those of the baseline multi-scale ensemble model (Section 3.1). Then, the results of each TreeNet model are compared to those of the baseline ensemble model (Section 3.2), both in terms of performance and number of parameters. Finally, the results of the SSBN model are compared to those of the other models and additionally validated in the VinDr-CXR data set (Section 3.3).

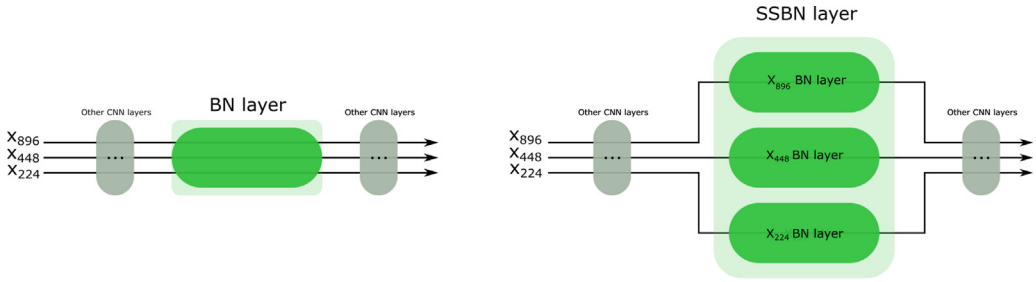


Fig. 4. Scale-Specific Batch Normalization (SSBN) layer. Instead of sharing a single batch normalization (BN) layer across scales, SSBN has one BN sub-module per scale. In the SSBN model, all the BN layers are replaced with SSBN layers, while all the remaining parameters are shared across input sizes.

Table 2
Number of parameters of the models discussed in this study. These include the the single-scale baselines (B224, B448 and B896), the TreeNet structures ranging from a fully joint network to the fully separate (except for the last fully connected layer) baseline ensemble (BE), and the SSBN model. The number of parameters of single-scale baseline models is similar to that of the Joint model.

Model	No. Parameters	% baseline parameters
Joint / B224 / B448 / B896	6,968,206	100%
C3	12,341,134	177%
C2	18,280,846	262%
C1	20,186,766	290%
BE	20,875,918	300%
SSBN	7,135,502	102%

3.1. Baseline multi-scale ensemble

We start by comparing the performance of each single-scale baseline model (224 × 224, 448 × 448, and 896 × 896 pixels) with the baseline multi-scale ensemble that takes as input all three scales, using one individual expert per scale (rightmost model in Fig. 3). Table 3 contains the per-label Area Under the Receiver Operating Characteristic Curve (AUC) of each model, as well as the macro-averaged and weighted-averaged AUCs. While the macro-averaged AUC is simply the average of the per-label AUCs, and therefore does not account for label imbalance in the data, the weighted averaged AUC weights the per-class AUCs with the number of instances of each label in the validation set, thus accounting for label imbalance. The results show that the multi-scale approach outperforms all individual scale baseline models.

3.2. Parameter sharing (TreeNets)

Aiming to design parameter-efficient models, we explore multiple TreeNet architectures as alternatives to the baseline multi-scale ensemble. The performance of each TreeNet can be found in Table 4. Model C1 is capable of achieving an AUC comparable to that of the baseline ensemble model, while using fewer parameters (it is 2.9 times larger than the single-scale baseline models instead of 3 times larger than the single-scale baseline models).

3.3. SSBN

The TreeNet models detailed in the previous subsection are more parameter efficient than the baseline ensemble model, and one of them (model C1) retains comparable performance. However, they are still much larger than the baseline models. We take a Densenet-121 model and replace its batch normalization layers with SSBN layers (Fig. 4), resulting in the SSBN model, which is used to process all input sizes as described in Section 2.2. The performance of this model is also shown in Table 4. This model yields an AUC that is almost identical to that of the baseline ensemble while resorting to a significantly smaller number of parameters (1.02× the size of the baseline models versus 3× the size of the baseline models).

Finally, we compare the proposed SSBN model to each individual single-scale model on the overlapping labels of the external VinDr-CXR data set, where SSBN still maintains superior performance (Table 5).

4. Discussion

4.1. Single-scale models

The results of the baseline models show that most labels have the best performance when the intermediate input size (448 × 448) is used. However, the difference between the best and second-best scales is not always notorious. While two labels (pneumothorax and support devices) are best classified using the largest scale (896 × 896), no label is best classified using the smallest scale (224 × 224). Some recent studies reported to have used input sizes equal or close to 448 × 448 pixels [10,34], while others still rely on smaller 224 × 224 input images [2,6,22,30]. Our results indicate that even though most computer vision applications default to the 224 × 224 image size, this input size seems to be suboptimal for chest radiograph classification.

For labels representing larger findings (cardiomegaly, enlarged cardiomediastinum), using scales larger than 224 × 224 does not offer significant performance improvements and therefore does not compensate for the larger computational costs associated with using larger inputs. The labels that benefit from larger scales (pneumothorax and support devices) most likely exploit features generated from smaller input regions (small RF/input size ratio). These findings are in line with the RF/input size rationale described in Fig. 1, and also with the results from Sabottke and Spieler [27] and Haque et al. [8], mentioned in Section 1.1, which find analogous relationships among per-label performance and input size. The performance obtained in the external validation (Table 5) is higher than that obtained in the internal validation, most likely due to differences in the nature of the data. However, the performance is still lower than that obtained by others when both training and testing in the VinDr-CXR data set [20,23].

4.2. Multi-scale models

The baseline ensemble model achieves better performance than any of its single-scale counterparts, across all individual labels. The results for the TreeNet models (Section 3.2) show a clear trade-off between AUC and the number of parameters in the model. When the model is fully-shared across all scales (fewer parameters), the model performs poorly. On the other hand, when the model shares

Table 3
Mean and standard deviation 5-fold cross validation results of the three baseline models (B224, B448 and B896) and the baseline ensemble model (BE).

Category	AUC			
	B224	B448	B896	BE
No Finding	87.77 ± 0.5	87.92 ± 0.45	87.83 ± 0.45	88.37 ± 0.47
Enlarged Cardiomediastinum	68.94 ± 0.49	69.46 ± 0.42	68.92 ± 0.27	70.16 ± 0.28
Cardiomegaly	87.48 ± 0.3	87.64 ± 0.25	87.02 ± 0.21	88.24 ± 0.27
Lung Opacity	76.86 ± 0.25	77.26 ± 0.24	76.99 ± 0.25	77.74 ± 0.22
Lung Lesion	78.72 ± 0.41	80.58 ± 0.48	80.25 ± 0.73	81.49 ± 0.49
Edema	88.03 ± 0.24	88.53 ± 0.17	88.38 ± 0.19	88.98 ± 0.18
Consolidation	77.67 ± 0.66	78.01 ± 0.52	77.12 ± 0.45	78.77 ± 0.61
Pneumonia	80.12 ± 0.96	81.08 ± 0.75	80.25 ± 0.67	81.68 ± 0.82
Atelectasis	72.98 ± 0.4	73.83 ± 0.34	73.67 ± 0.33	74.47 ± 0.44
Pneumothorax	87.87 ± 0.23	90.42 ± 0.3	91.32 ± 0.28	91.62 ± 0.25
Pleural Effusion	90.3 ± 0.22	90.47 ± 0.27	90.21 ± 0.28	90.92 ± 0.26
Pleural Other	80.45 ± 1.45	81.57 ± 0.9	81.31 ± 1.13	81.95 ± 0.9
Fracture	78.3 ± 0.6	79.7 ± 0.39	79.74 ± 0.46	80.68 ± 0.53
Support Devices	89.15 ± 0.23	90.28 ± 0.23	90.45 ± 0.25	90.71 ± 0.22
MAUC	81.76 ± 0.18	82.62 ± 0.11	82.39 ± 0.13	83.27 ± 0.11
WAUC	84.58 ± 0.09	85.27 ± 0.06	85.14 ± 0.08	85.81 ± 0.06

Table 4
Mean and standard deviation 5-fold cross validation results of the TreeNet structures (ranging from the fully joint model to the fully separate model, which corresponds to the S) and the proposed SSBN model. MAUC = Macro AUC; WAUC = Weighted AUC; Enl. Card. = Enlarged Cardiomediastinum.

Category	AUC					
	Joint	C3	C2	C1	BE	SSBN
No Finding	87.58 ± 0.54	88.14 ± 0.5	88.22 ± 0.51	88.31 ± 0.48	88.37 ± 0.47	88.36 ± 0.46
Enl. Card.	68.67 ± 0.4	69.76 ± 0.4	69.84 ± 0.38	70.05 ± 0.3	70.16 ± 0.28	70.12 ± 0.44
Cardiomegaly	87.16 ± 0.14	87.91 ± 0.3	88.03 ± 0.31	88.08 ± 0.33	88.24 ± 0.27	88.04 ± 0.26
Lung Opacity	76.78 ± 0.26	77.41 ± 0.26	77.51 ± 0.24	77.65 ± 0.2	77.74 ± 0.22	77.75 ± 0.23
Lung Lesion	79.35 ± 0.51	80.86 ± 0.54	81.37 ± 0.54	81.53 ± 0.59	81.49 ± 0.49	81.53 ± 0.54
Edema	87.95 ± 0.19	88.64 ± 0.24	88.69 ± 0.15	88.85 ± 0.13	88.98 ± 0.18	88.91 ± 0.19
Consolidation	77.25 ± 0.68	78.46 ± 0.71	78.61 ± 0.67	78.6 ± 0.54	78.77 ± 0.61	78.77 ± 0.67
Pneumonia	79.88 ± 0.85	80.89 ± 0.86	81.56 ± 0.92	81.6 ± 0.86	81.68 ± 0.82	81.67 ± 0.77
Atelectasis	73.01 ± 0.36	73.98 ± 0.53	74.13 ± 0.34	74.33 ± 0.37	74.47 ± 0.44	74.51 ± 0.48
Pneumothorax	88.96 ± 0.23	90.5 ± 0.3	90.89 ± 0.17	91.44 ± 0.11	91.62 ± 0.25	91.50 ± 0.19
Pleural Effusion	90.27 ± 0.35	90.72 ± 0.27	90.76 ± 0.27	90.84 ± 0.24	90.92 ± 0.26	90.94 ± 0.27
Pleural Other	80.87 ± 0.8	81.76 ± 0.85	82.02 ± 0.85	82.09 ± 0.91	81.95 ± 0.9	82.27 ± 0.86
Fracture	78.69 ± 0.39	80.15 ± 0.54	80.53 ± 0.31	80.85 ± 0.43	80.68 ± 0.53	80.77 ± 0.54
Support Devices	89.01 ± 0.24	90.35 ± 0.16	90.57 ± 0.23	90.61 ± 0.2	90.71 ± 0.22	90.65 ± 0.28
MAUC	81.82 ± 0.12	82.82 ± 0.13	83.05 ± 0.11	83.2 ± 0.12	83.27 ± 0.11	83.27 ± 0.17
WAUC	84.54 ± 0.05	85.43 ± 0.07	85.59 ± 0.06	85.71 ± 0.05	85.81 ± 0.06	85.78 ± 0.07

Table 5
External validation mean and standard deviation 5-fold cross validation results of the single-scale baselines and proposed SSBN model. MAUC = Macro AUC; WAUC = Weighted AUC.

Category	AUC			
	224	448	896	SSBN
Cardiomegaly	86.03 ± 2.15	85.82 ± 1.5	82.69 ± 1.45	86.7 ± 0.77
Pneumothorax	98.24 ± 0.65	99.27 ± 0.25	99.31 ± 0.4	99.74 ± 0.18
Atelectasis	73.43 ± 3.24	70.83 ± 2.99	68.48 ± 4.33	70.43 ± 1.47
Pleural Effusion	95.92 ± 0.53	95.83 ± 0.31	94.95 ± 0.71	96.2 ± 0.76
Lung Lesion	83.63 ± 1.06	91.75 ± 1.08	93.13 ± 0.28	92.91 ± 0.35
Lung Opacity	86.49 ± 0.68	87.86 ± 1.15	87.18 ± 1.06	88.64 ± 0.33
Consolidation	95.76 ± 0.66	96.3 ± 0.26	95.29 ± 0.69	96.78 ± 0.19
No Finding	88.25 ± 0.68	90.1 ± 0.77	90.13 ± 0.64	91.29 ± 0.56
MAUC	88.47 ± 0.6	89.72 ± 0.51	88.89 ± 0.65	90.34 ± 0.21
WAUC	88.15 ± 0.51	89.78 ± 0.57	89.35 ± 0.54	90.86 ± 0.45

only a few parameters from the early layers and keeps the remaining parameters separate for each scale (model C1), the performance is nearly as good as the one from the baseline ensemble. Although model C1 performs almost as well as the baseline ensemble using a smaller number of parameters, it is still 2.9 times larger than the single-scale baseline models.

The proposed SSBN model is capable of achieving a performance comparable to that of a baseline ensemble, similar to that proposed in [8], while using a much smaller number of parameters. This not only highlights the extreme importance of affine transforms in convolutional neural networks and deep learning in general, but also enables the use of multi-scale inputs in less powerful machines. Moreover, Table 5 shows that SSBN generalizes well, as its performance it still overall superior to those of the single-scale baselines in an external data set. Note that for some labels, such as Atelectasis or Pneumothorax, the amount of

positive instances is quite small [21] (less than 100 scans), which may be insufficient to produce reliable results. Interestingly, in the external validation, the standard deviation across the five folds is notably smaller for SSBN, which shows that incorporating multiple scales into the predictions leads to more robust results.

4.3. Limitations

On the one hand, CheXpert's ground-truth labels are known to suffer from label inaccuracies due to the fact that these were automatically extracted using natural language processing. On the other, the agreement among the labelling of different radiologists for the same scan is low on the VinDR-CXR data set [32], which may also pose a problem. The external validation experiment may also be influenced by the different labelling schemes adopted in the data sets. While a direct label correspondence was adopted, labels with the same name may not mean exactly the same thing in both data sets. Finally, while building our model is extremely easy (simply replacing the BN layers by SSBN layers), it does require a personalized forward method to handle the two main modifications we make, which are (1) having three different-sized versions of the input and (2) a "switch" to alternate between the three sets of BN layers, as we use a different set for each resolution. Due to this, training time increases, and is comparable to that of the baseline ensemble. Nevertheless, memory requirements are smaller than those of the baseline ensemble (although naturally larger than those of single-scale models), due to the reduced number of parameters.

5. Conclusions

The variations in size and shape that radiological findings of chest radiographs exhibit pose a challenge to standard deep learning models, as models built based on a single input size do not account the multi-scale nature of the data. We developed a new Size-Specific Batch Normalization (SSBN) layer that can replace standard batch normalization layers to create a lightweight multi-scale ensemble, while adding a very small number of parameters to the model (+ 2%). This model achieves a performance similar to that of a baseline ensemble that uses one individual CNN per scale, with a much smaller number of parameters (approximately one third). Therefore, the proposed approach allows the multi-scale nature of the findings to be explicitly taken into account without the need for a model that is much larger than its single-scale counterparts.

Due to differences in the proportion between the receptive field of the convolutions and the input sizes, the proposed multi-scale approach leverages and combines features with different levels of granularity, boosting performance. In conclusion, this work highlights the per-label effect of varying the input scale of chest radiographs fed to deep learning models, offers a new lightweight alternative to combine and benefit from multiple scales and investigates the generalization capabilities of the proposed alternative. Future work can be focused on studying the effect of using a broader number of scales and different architectural backbones for the SSBN model.

Declaration of Competing Interest

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

Acknowledgments

This work was supported by the ERDF - European Regional Development Fund, through the Programa Operacional Regional do Norte (NORTE 2020) and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. within the scope of the CMU Portugal Program (NORTE- 01-0247-FEDER-045905) and LA/P/0063/2020. The work of S. C. Pereira was supported by the FCT grant contract 2020.10169.BD. The work of J. Rocha was supported by the FCT grant contract 2020.06595.BD. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M.U. Alam, J.R. Baldvinsson, Y. Wang, Exploring LRP and Grad-CAM visualization to interpret multi-label-multi-class pathology prediction using chest radiography, in: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), 2022, pp. 258–263, doi:10.1109/CBMS55023.2022.00052.

[2] M.M. Alshahrni, M.A. Ahmad, M. Abdullah, N. Omer, M. Aziz, An intelligent deep convolutional network based Covid-19 detection from chest X-rays, Alex. Eng. J. (2022), doi:10.1016/j.aej.2022.09.016.

[3] W.-G. Chang, T. You, S. Seo, S. Kwak, B. Han, Domain-specific batch normalization for unsupervised domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, Institute of Electrical and Electronics Engineers (IEEE), 2009, pp. 248–255, doi:10.1109/CVPR.2009.5206848.

[5] J. Frankle, D.J. Schwab, A.S. Morcos, Training BatchNorm and only BatchNorm: On the expressive power of random features in CNNs, in: International Conference on Learning Representations, 2021, doi:10.48550/ARXIV.2003.00152.

[6] P. Ghose, M.A. Uddin, U.K. Acharjee, S. Sharmin, Deep viewing for the identification of Covid-19 infection status from chest X-Ray image using CNN based architecture, Intell. Syst. Appl. 16 (2022) 200130, doi:10.1016/j.iswa.2022.200130.

[7] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, S. Hikosaka, Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018-January, 2018, pp. 1442–1450, doi:10.1109/WACV.2018.00162. 1709.00179.

[8] M.I.U. Haque, A.K. Dubey, J.D. Hinkle, The effect of image resolution on automated classification of chest X-rays, 2021. doi:10.1101/2021.07.30.21261225.

[9] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37, JMLR.org, 2015, pp. 448–456, doi:10.5555/3045118.3045167.

[10] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A.Y. Ng, CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: 33rd AAAI Conference on Artificial Intelligence, 2019, pp. 590–597, doi:10.1609/aaai.v33i01.3301590. 1901.07031.

[11] D. Komura, S. Ishikawa, Machine learning methods for histopathological image analysis, Comput. Struct. Biotechnol. J. 16 (2018) 34–42, doi:10.1016/J.CSBJ.2018.01.001.

[12] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Technical Report, 2009.

[13] S. Lee, S. Purushwalkam, M. Cogswell, D. Crandall, D. Batra, Why M heads are better than one: training a diverse ensemble of deep networks, 2015. doi:10.48550/arxiv.1511.06314.

[14] H. Li, J.Y.-H. Ng, P. Natsev, EnsembleNet: end-to-end optimization of multi-headed models, 2019. doi:10.48550/arxiv.1905.09979.

[15] L.A. Lim, H. Yalim Keles, Foreground segmentation using convolutional neural networks for multiscale feature encoding, Pattern Recognit. Lett. 112 (2018) 256–262, doi:10.1016/J.PATREC.2018.08.002.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755, doi:10.48550/arXiv.1405.0312.

[17] W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 29, Curran Associates, Inc., 2016, doi:10.48550/arXiv.1701.04128.

[18] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, H. Wu, Mixed precision training, in: International Conference on Learning Representations, 2018, doi:10.48550/arXiv.1710.03740.

[19] A.R. Narayanan, A. Zela, T. Saikia, T. Brox, F. Hutter, Multi-headed neural ensemble search, 2021. doi:10.48550/arxiv.2107.04369.

[20] A.A. Nasser, M.A. Akhloufi, Classification of CXR chest diseases by ensembling deep learning models, in: Proceedings - 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science, IRI 2022, 2022, pp. 250–255, doi:10.1109/IRI54793.2022.00062.

[21] H.Q. Nguyen, K. Lam, L.T. Le, H.H. Pham, D.Q. Tran, D.B. Nguyen, D.D. Le, C.M. Pham, H.T. Tong, D.H. Dinh, C.D. Do, L.T. Doan, C.N. Nguyen, B.T. Nguyen, Q.V. Nguyen, A.D. Hoang, H.N. Phan, A.T. Nguyen, P.H. Ho, D.T. Ngo, N.T. Nguyen, N.T. Nguyen, M. Dao, V. Vu, VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations, Scientific Data 9 (2022) 1–7, doi:10.1038/s41597-022-01498-w.

[22] H.H. Pham, T.T. Le, D.Q. Tran, D.T. Ngo, H.Q. Nguyen, Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels, Neurocomputing 437 (2021) 186–194, doi:10.1016/j.neucom.2020.03.127.

[23] H.H. Pham, H.Q. Nguyen, H.T. Nguyen, L.T. Le, L. Khanh, An accurate and explainable deep learning system improves interobserver agreement in the interpretation of chest radiograph, IEEE Access 10 (2022) 104512–104531, doi:10.1109/ACCESS.2022.3210468.

[24] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2016, pp. 779–788, doi:10.1109/CVPR.2016.91. 1506.02640.

[25] M.L. Richter, W. Byttner, U. Krumnack, A. Wiedenroth, L. Schallner, J. Shenk, (Input) size matters for CNN classifiers, Lect. Notes Comput. Sci. 12892 LNCS (2021) 133–144, doi:10.1007/978-3-030-86340-1_11.

[26] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[27] C.F. Sabottke, B.M. Spieler, The effect of image resolution on deep learning in radiography, Radiol. Artif. Intell. 2 (1) (2020) e190015, doi:10.1148/ryai.2019190015. PMID: 33937810

[28] J.A. dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. da S. Torres, A.X. Falao, Multiscale classification of remote sensing images, IEEE Trans. Geosci. Remote Sens. 50 (10) (2012) 3764–3775, doi:10.1109/TGRS.2012.2186582.

[29] S. Sivakumar, C. Chandrasekar, Lung nodule detection using fuzzy clustering and support vector machines, Int. J. Eng. Technol. 5 (1) (2013) 179–185.

[30] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, 2017, pp. 3462–3471, doi:10.1109/CVPR.2017.369. 1705.02315.

[31] Y. Wang, S. Hu, G. Wang, C. Chen, Z. Pan, Multi-scale dilated convolution of convolutional neural network for crowd counting, Multimed. Tools Appl. 79 (1–2) (2020) 1057–1073, doi:10.1007/S11042-019-08208-6.

[32] Weronika, Detailed analysis of the competition database - examination of consistency in annotation and data quality, 2021, Accessed on 13.03.2023, https://www.kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection/discussion/251250.

[33] J. Yanase, E. Triantaphyllou, A systematic survey of computer-aided diagnosis in medicine: past and present developments, 2019, 10.1016/j.eswa.2019.112821

[34] H. Yasar, M. Ceylan, A new deep learning pipeline to detect Covid-19 on chest X-ray images using local binary pattern, dual tree complex wavelet transform and convolutional neural networks, Appl. Intell. 51 (2021) 2740–2763, doi:10.1007/S10489-020-02019-1.

[35] W. Zhang, J. Jiang, Y. Shao, B. Cui, Efficient diversity-driven ensemble for deep neural networks, in: International Conference on Data Engineering, Vol. 2020-April, IEEE Computer Society, 2020, pp. 73–84, doi:10.1109/ICDE48307.2020.00014.