Video Processing and Understanding Lab

Universidad Autónoma de Madrid

SI1/PJI/2019-00414 AISEEME (2020-2022)

Aiding diagnosis by self-supervised deep learning from unlabeled medical imaging

D4

Multi-task SSL framework for applications in medical imaging

Video Processing and Understanding Lab

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Supported by CONSEJERÍA DE EDUCACIÓN E INVESTIGACIÓN Comunidad de Madrid

AUTHORS LIST

Pablo Carballeira López	pablo.carballeira@uam.es
Marcos Escudero Viñolo	marcos.escudero@uam.es
Juan Carlos San Miguel Avedillo	juancarlos.sanmiguel@uam.es
Kirill Sirotkin	kirill.sirotkin@uam.es

HISTORY

Version	Date	Editor	Description
1.0	05/08/2022	Pablo Carballeira López	Added Section 1.2.1
1.1	09/12/2022	Juan Carlos San Miguel Avedillo	Added Section 1.2.2
1.2	09/12/2022	Kirill Sirotkin	Added Section 1.1
1.3	12/12/2022	Pablo Carballeira López	Revision and formatting

CONTENTS:

1. INTRODUCTION **1**

1.1. MULTI-TASK SSL APPROACHES FOR SKIN LESION ASSESSMENT 3

1.2. MULTI-TASK SSL APPROACHES FOR LUNG NODULE MALIGNANCY DETECTION 4

1.2.1. Curricular SSL training for COVID-19 pneumonia recognition in CXR images *5*

1.2.2. Generation of synthetic data to train deep learning models for CXR image classification *6*

REFERENCES **10**

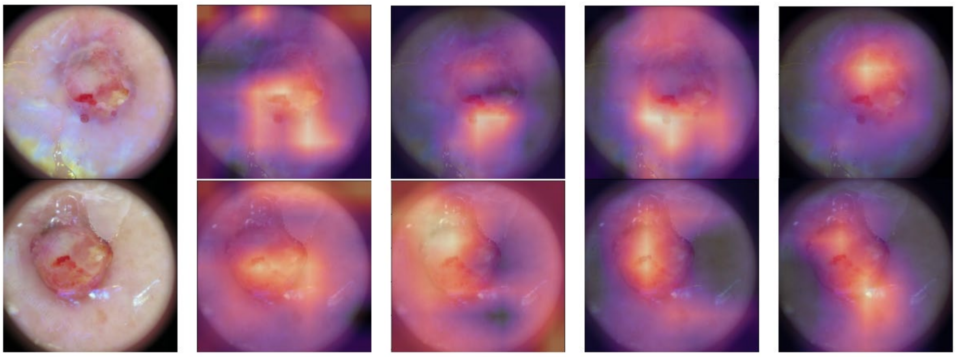
1. Introduction

This deliverable describes the work related to tasks T4.1: Multi-task SSL approaches for skin lesion assessment, and T4.2: Multi-task SSL approaches for lung nodule malignancy detection.

1.1. Multi-task SSL approaches for skin lesion assessment

In this task, we have studied the applicability of the multi-task Self-Supervised Learning (SSL) approach, described in D3, for the recognition of skin lesion images. Specifically, we used the International Skin Image Collaboration (ISIC) 2019 dataset that presents 8 types of skin lesions. The dataset used for this study consists of three skin lesion datasets collected independently, namely, HAM 10000 [12], BCN_20000 [13], and MSK [14]. Combined, the three datasets contain 25331 annotated training images which we split into the training and validation set of 20264 and 5067 images, respectively.

A pre-print manuscript of this work, with the complete description of experiments and results, can be found in [11]. In it, we explore the benefits of combining multiple-SSL in the training of a deep training model for the classification of images depicting skin lesions and compare the obtained performances with a purely Supervised learning approach. In particular, we demonstrate that sequential curricular pre-training on multiple pretext tasks (Relative Location, MoCo-v2 and ODC) outperforms its fully-supervised counterpart, even when the latter is pre-trained on a large-scale dataset, such as ImageNet. We show that at least four combinations of three SSL tasks outperform ImageNet pretraining, with the best combination reaching 75.44% of balanced accuracy on the validation set (+2.94% compared to the ImageNet pretraining). Moreover, we present evidence that effective curriculum orderings of the SSL tasks correlate with increasing downstream accuracies obtained for the individual SSL task, therefore, reducing the search space when approaching a new task. A summary of the obtained results is presented in Table 1.



(a) Original (b) No pretraining (c) Supervised pretraining (ImageNet) (d) MoCo v2 (best single) (e) ODC → MoCo v2 (best combination)

Figure 1 Class activation maps of classifiers with and without pretraining on pretext tasks. In the shown samples, classifiers that had no pretraining tend to focus on irrelevant parts of the images (black surrounding areas) and incorrectly classify skin lesions.

In [11], we demonstrate qualitative results, in the form of Class Activation Maps (CAMs), showing that curriculum SSL pretraining improves the final model, focusing more of its attention on the lesions. An example of such qualitative results can be found in Figure 1.

Table 1 Balanced accuracies for the evaluated single- and multisource transfer settings for the ISIC-19 skin lesion recognition task. The right-most column indicates whether the pretraining strategy led to a higher classification accuracy than supervised pretraining on ImageNet. The column "δ" indicates how the performance of a combination of pretext tasks differs from an individual pretext task. The left-most column shows whether a combination follows Curriculum (C), Anti-Curriculum (AC) or Mixed Curriculum (MC) ordering.

	1st task	2nd task	3rd task	Balanced accuracy (%)	δ (%)	Better than ImageNet
-	Rel. loc.			69.52	-	No
(AC)	Rel. loc.	ODC		70.68	1.16	No
(MC)	Rel. loc.	ODC	MoCo v2	75.00	5.49	**Yes**
(C)	Rel. loc.	MoCo v2		74.10	4.58	**Yes**
(MC)	Rel. loc.	MoCo v2	ODC	74.38	4.86	**Yes**
-	MoCo v2			72.74	-	No
(AC)	MoCo v2	ODC		72.72	-0.02	No
(MC)	MoCo v2	ODC	Rel. loc.	67.00	-5.74	No
(AC)	MoCo v2	Rel. loc.		66.72	-6.02	No
(AC)	MoCo v2	Rel. loc.	ODC	69.80	-2.95	No
-	ODC			63.52	-	No
(C)	ODC	Rel. loc.		68.23	4.71	No
(C)	ODC	Rel. loc.	MoCo v2	73.36	9.84	No
(C)	ODC	MoCo v2		75.44	11.92	**Yes**
(MC)	ODC	MoCo v2	Rel. loc.	65.73	2.21	No
ISIC-2019 challenge winner [3]				72.5 ± 1.7	-	-
Supervised ImageNet				73.76	-	-
No pretraining				49.27	-	-

1.2. Multi-task SSL approaches for lung nodule malignancy detection

Given the problems of lack of structure in the CT lung-scan datasets, which have been reported intermediate reports of the project, this task has been re-oriented to evaluation on medical imaging databases of a similar modality. Specifically, we have used Chest X-Ray (CXR) images which, contrary to skin lesion images, represent medical images not obtained by optical means. In this task, we have analyzed the benefits of two strategies for CXR image classification: i) a curricular SSL training scheme (Section 1.2.1), and the generation of synthetic CXR images to train deep learning models (Section 1.2.2).

1.2.1. Curricular SSL training for COVID-19 pneumonia recognition in CXR images

In this task, we evaluate the benefits of a curricular Self-Supervised Learning (SSL) pretraining scheme with respect to fully supervised training regimes for pneumonia recognition on CXR images of Covid-19 patients. The complete description of experiments and results can be found in [1]. We have used the curricular SSL training scheme proposed in D3, with learning-rate (LR) selection in each training step (both SSL steps and downstream classification) using the following policy:

- Each training step is repeated, using different LR values taken from a pre-defined range, for a limited number of epochs.
- The LR that leads to the highest performance (or lowest loss value) on the training task is used to train the model for the full number of epochs

For this evaluation, we have used the SIIM-FISABIO-RSNA COVID-19 Detection dataset [2], which collects CXR images of Covid-19 patients. The SIIM-FISABIO-RSNA training data consists of 6,334 chest scans and is built from two datasets: BIMCV-COVID19+ [3] and MIDRC-RICORD [4]. We show that curricular SSL pretraining, which leverages unlabelled data, outperforms models trained from scratch, or pretrained on ImageNet, indicating the potential of performance gains by SSL pretraining on massive unlabelled datasets. We show that a combination of SSL tasks can outperform pretraining on ImageNet, or training directly with the target data. With our best configuration, MoCo v2 + SwAV + Relative Location, we achieve a +1.98% accuracy increase over the baselines. The results provide evidence that additional SSL tasks can increase the performance of the model compared to pretraining with only one SSL task. A summary of the performance results is shown in Table 2 Balanced accuracies and AIL scores for the curricular SSL-task pretraining con-figurations. Sequential orderings for SSL-tasks read left to right. The curriculum column indicates whether a SSL-task combination follows a curriculum ordering. Results in bold refer to the highest score of each block, while results in blue are the highest scores overall.

.

Also, literature indicates that recent deep learning systems targeting disease detection from CXRs, rather than learning on the medical pathology evidence, rely on confounding factors [5], out of the lung regions, as a learning shortcut. These confounding factors are prone to be dependent on the training dataset. Therefore, in [1] we propose a strategy to quantitatively compare different models in terms of the degree of attention they present in the lung regions. This strategy is used to show evidence that SSL pretraining (and curricular SSL pretraining) is beneficial to focus the model's attention on the region of interest of the CXR image, in this case, the lungs. These results indicate that SSL-pretrained models are prone to be more robust to the external confounding factors, increasing the generalization capabilities of the deep learning solution. This study is based on an AIL (Attention Inside Lungs) score which allow us to compare the level attention of in-the-lung regions of several models. Higher AIL values correspond to models with a higher focus in the lung regions. A summary of the AIL score results is shown in Table 2.

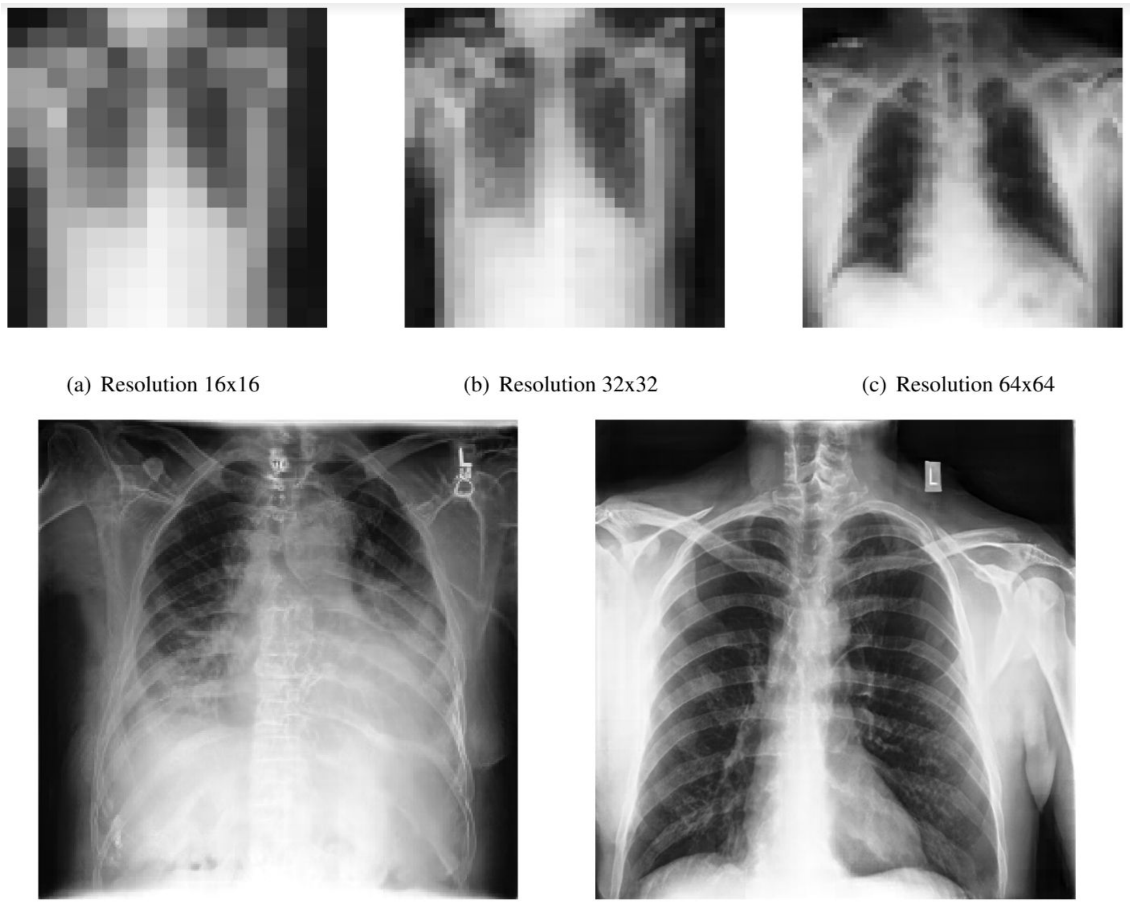
Table 2 Balanced accuracies and AIL scores for the curricular SSL-task pretraining configurations. Sequential orderings for SSL-tasks read left to right. The curriculum column indicates whether a SSL-task combination follows a curriculum ordering. Results in bold refer to the highest score of each block, while results in blue are the highest scores overall.

Curriculum	Pretraining	Validation Acc (%)	AIL (%)
-	ImageNet	82.75	**38.16**
-	Scratch	**83.69**	37.30
-	Rel-Loc	83.62	36.33
-	MoCo v2	83.89	37.40
-	Swav	83.97	42.82
-	Rotation	**84.72**	**45.95**
✓	MoCo v2 + Rotation	84.77	**47.92**
	MoCo v2 + Rel-Loc	**85.59**	39.57
✓	MoCo v2 + SwAV	83.67	41.79
	Rotation + MoCo v2	76.08	31.87
	Rotation + Rel-Loc	84.33	**44.48**
	Rotation + SwAV	**84.81**	41.46
✓	Rel-Loc + Rotation	84.54	**48.63**
✓	Rel-Loc + MoCo v2	82.79	39.41
✓	Rel-Loc + SwAV	**85.27**	46.26
✓	SwAV + Rotation	83.89	**47.16**
	SwAV + Rel-Loc	**84.92**	43.05
	SwAV + MoCo v2	82.37	36.51
	MoCo v2 + Rotation + Rel-Loc	**85.28**	**38.82**
	MoCo v2 + Rotation + SwAV	84.80	30.69
	MoCo v2 + Rel-Loc + Rotation	84.19	38.51
	MoCo v2 + Rel-Loc + SwAV	**85.49**	**46.30**
✓	MoCo v2 + SwAV + Rotation	85.67	40.19
	MoCo v2 + SwAV + Rel-Loc	83.74	**40.89**

1.2.2. Generation of synthetic data to train deep learning models for CXR image classification

In this task, we focus on exploring the use of synthetic data that can provide us with large datasets without any privacy issues at a reduced cost. The complete description of experiments and results can be found in [6]. To achieve that, we have employed a synthetic data generator based on Generative Adversarial

Networks [7][8]. Ideally, these artificially generated images should not contain sensitive personal information while maintaining statistical features like the original images. Figure 2 shows some examples of the generated synthetic images.



(a) Resolution 16x16 (b) Resolution 32x32 (c) Resolution 64x64

(d) Resolution 256x256 (e) Resolution 512x512

Figure 2 Synthetic sample images generated using an approach based on Generative Adversarial Networks[8].

Based on [8], pretrained on the images of the CheXpert dataset (https://stanfordmlgroup.github.io/competitions/chexpert/), we have created a dataset composed of different versions considering the presence of isolated findings (i.e., 0s versions) or combined with other findings in the images (i.e. Xs versions). In total, two datasets are generated for a binary problem classification: version 1 (No-finding VS Only-Pneumothorax) and version 2 (No-Finding VS Finding). A third dataset is generated for a four-class problem (No-Finding VS Pneumotorax VS Pneumonia VS Cardiomegaly). Table 3 summarizes the generated datasets.

Table 3 Summary of the synthetic datasets generated for Chest X-ray images

Synthetic			**No Finding**	**Pneumothorax**	**Pneumonia**	**Cardiomegaly**	**Finding**
Version 1	0s	Training	4000 (50%)	4000 (50%)	-	-	-
		Validation	1000 (50%)	1000 (50%)	-	-	-
	Xs	Training	4000 (50%)	4000 (50%)	-	-	-
		Validation	1000 (50%)	1000 (50%)	-	-	-
Version 2	-	Training	4000 (50%)	-	-	-	4000 (50%)
		Validation	1000 (50%)	-	-	-	1000 (50%)
Version 3	0s	Training	4000 (25%)	4000 (25%)	4000 (25%)	4000 (25%)	-
		Validation	1000 (25%)	1000 (25%)	1000 (25%)	1000 (25%)	-
	Xs	Training	4000 (25%)	4000 (25%)	4000 (25%)	4000 (25%)	-
		Validation	1000 (25%)	1000 (25%)	1000 (25%)	1000 (25%)	-

Then, the next goal in [6] is to exploit the synthetic data for training and adapting algorithms using few real data without annotations. Therefore, this goal is approached as an Unsupervised Domain Adaptation (UDA) from synthetic to real data for classification tasks. We employed the DCAN approach [9]. At a low level, DCAN uses a Resnet-50 backbone with an attention module that is trained using numerous training losses to achieve the desired feature alignment. Due to the reduced number of classes, and the lack of reliability of the ground truth, we decided to remove from the training loss the regularization loss of the ith feature regularizer which aims at solving the over-correction problems caused by the added feature correction blocks with the guide of source data [9]. This what we called ”Our Alignment”.

Table 4 summarizes the UDA results for the different combinations of datasets (synthetic version 1, CheXpert and ChestX-ray8[10]) for the presence of isolated findings (i.e., 0s version). In this table we can see the results obtained in each of the combinations between source and target. As we can see, the best option if we use synthetic images as source data is Our UDA approach, while if we use real images as source data the best option is not to apply any alignment. Also, as expected, when using CheXpert as target data, the best option is to use synthetic images as source data, with quite a difference (≈ 8%). This is probably due to the fact that these images have been generated with the GAN that was trained using CheXpert. On the contrary, when using ChestX-ray8, the best option, although with very little difference (<2%) is to use the CheXpert data. This makes sense as it is a larger dataset and probably contains more information. Finally, it should be noted that we have not been able to run UDA using ChestX-ray8 as source and CheXpert as target; the model always predicted the same class, thus achieving a 50% Balanced accuracy.

Table 4 Summary of UDA results for dataset version 1 with isolated findings. Best result is indicated in bold

Target	*Source*	*Proposal*	*Balanced Acc*
CheXpert	**Synthetic**	*No Alignment*	63.31 %
		DCAN Alignment	73.03 %
		Our Alignment	**73.98 %**
	ChestX-ray8	*No Alignment*	**65.04 %**
		DCAN Alignment	50.00 %
		Our Alignment	50.00 %
ChestX-ray8	**Synthetic**	*No Alignment*	56.42 %
		DCAN Alignment	61.13 %
		Our Alignment	**61.88 %**
	ChestXpert	*No Alignment*	**63.83 %**
		DCAN Alignment	62.04 %
		Our Alignment	61.84 %

Table 5 and Table 6 present the results for the presence of isolated findings (i.e., 0s version) and the synthetic datasets version 1 and version 2. In Table 5 the best result using synthetic data is using our alignment, reaching 52.9%, far away from the 74.7% obtained from training and evaluating on CheXpert. Due to these results, we did not continue experimenting with other datasets here either. In Table 6, as expected, the results are not bad, but they are not too good either. In the case of CheXpert, UDA works very badly, with DCAN failing to learn anything and always predicting the same class (that's why it gets 25% correct), but without using UDA we managed to improve that 25% to 30.1% (+5.1%). Despite this improvement, we are still a bit far from the 41.5% (-11.3%) obtained using CheXpert over CheXpert. In the case of ChestX-ray8, the results are surprisingly better, using our UDA alignment we achieved 30.3%, which is a little closer to the ChestX-ray8 result of 36.7% (-6.4%).

Table 5 Summary of UDA results for dataset version 2 with isolated findings. Best result is indicated in bold

Target	*Source*	*Proposal*	*Balanced Acc*
CheXpert	**Synthetic**	*No Alignment*	51.17 %
		DCAN Alignment	52.61 %
		Our Alignment	**52.94 %**
CheXpert	**CheXpert**	*No Alignment*	**74.78 %**

Table 6 Summary of UDA results for dataset version 3 with isolated findings. Best result is indicated in bold

Target	*Source*	*Proposal*	*Balanced Acc*
CheXpert	**Synthetic**	*No Alignment*	**30.13 %**
		DCAN Alignment	25.00 %
		Our Alignment	28.07 %
CheXpert	**CheXpert**	*No Alignment*	**41.50 %**
ChestX-ray8	**Synthetic**	*No Alignment*	28.54 %
		DCAN Alignment	27.73 %
		Our Alignment	**30.34 %**
ChestX-ray8	**ChestX-ray8**	*No Alignment*	**36.74 %**

As a conclusion of the work done in [6], the results obtained in the binary problem with 0s dataset are promising, so it seems like the idea works, but as soon as the classification task start getting harder the problems begin. This behaviour could be explained by the fact that synthetic data is visually plausible but it is not able to generate new information. Regarding the use of Domain Adaptation algorithms, these seem to improve the results, but like classical algorithms, they also require sufficiently representative data. We have proved that, according to [8], a key challenge for applying AI in the medical field is the representativeness of the data employed for training AI models.

References

[1] Self-Supervised Curricular Learning for Chest X-Ray Image Classification, Iván de Andrés Tamé, (advisor: Pablo Carballeira López), Trabajo Fin de Máster (Master Thesis), Máster Universitario en Deep Learning for Audio and Video Signal Processing, Univ. Autónoma de Madrid, Jul. 2022 (http://www-vpu.eps.uam.es/projects/aiseeme/public_docs/Report_MasterThesis_IvanAndres.pdf)

[2] Paras Lakhani, John Mongan, Chinmay Singhal, Quan Zhou, Katherine P Andriole, William F Auffermann, Prasanth Prasanna, Theresa Pham, Michael Peterson, Peter J Bergquist, et al. The 2021 siim-fisabio-rsna machine learning covid-19 challenge: Annotation and standard exam classification of covid-19 chest radiographs. 2021.

[3] Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. arXiv preprint arXiv:2006.01174, 2020.

[4] Emily B Tsai, Scott Simpson, Matthew P Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J Erickson, George Shih, Anouk

Stein, Jayashree Kalpathy-Cramer, et al. The rsna international covid-19 open radiology database (ricord). Radiology, 299(1):E204, 2021.

[5] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3(7):610–619, 2021.

[6] Learning supervised by synthetic data for Chest X-ray images, Eric Morales Agostinho, (advisor: Juan Carlos San Miguel Avedillo), Trabajo Fin de Máster (Master Thesis), Máster Universitario en Deep Learning for Audio and Video Signal Processing, Univ. Autónoma de Madrid, Jul. 2022. (http://www-vpu.eps.uam.es/projects/aiseeme/public_docs/Report_MasterThesis_EricMorales.pdf)

[7] Schütte, A. D., Hetzel, J., Gatidis, S., Hepp, T., Dietz, B., Bauer, S., and Schwab, P. (2021). Overcoming barriers to data sharing with medical image generation: a comprehensive evaluation. NPJ Digital Medicine, 4.

[8] Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5(6), 493-497.

[9] Li, S., Liu, C. H., Lin, Q., Xie, B., Ding, Z., Huang, G., and Tang, J. (2020). Domain conditioned adaptation network. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, pages 11386–11393.

[10] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2097-2106).

[11] Sirotkin, K., Viñolo, M. E., Carballeira, P., & SanMiguel, J. C. (2021). Improved skin lesion recognition by a Self-Supervised Curricular Deep Learning approach, arXiv preprint arXiv:2112.12086 (https://arxiv.org/abs/2112.12086)

[12] Tschandl P., Rosendahl C. & Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi.10.1038/sdata.2018.161 (2018)

[13] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, Josep Malvehy: "BCN20000: Dermoscopic Lesions in the Wild", 2019; arXiv:1908.02288.

[14] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern: "Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)", 2017; arXiv:1710.05006.