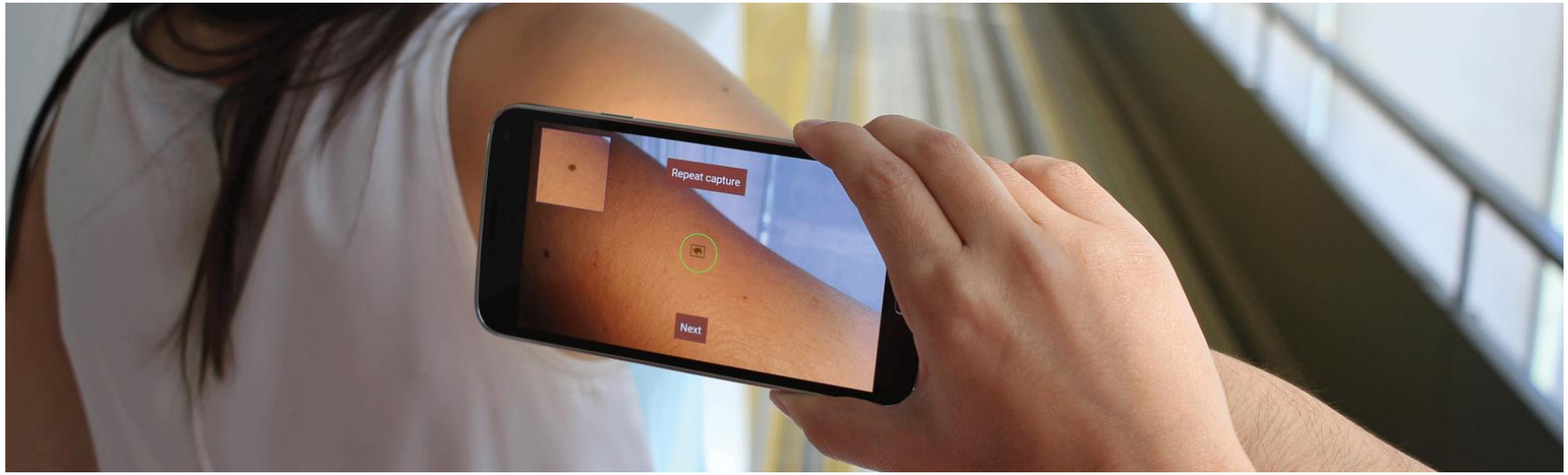
CHALLENGES ON REAL-WORLD SKIN LESION CLASSIFICATION: COMPARING FINE-TUNING STRATEGIES FOR DOMAIN ADAPTATION USING DEEP LEARNING

Tudor Nedelcu, **André Carreiro**, Francisco Veiga, Maria Vasconcelos

Fraunhofer Portugal AICOS

andre.carreiro@fraunhofer.pt



IARIA

The Sixth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing - HEALTHINFO 2021

Fraunhofer PORTUGAL

- PhD in Biomedical Engineering from Tecnico Lisboa
- Senior Researcher at Fraunhofer Portugal AICOS
  - Intelligent Systems

  - Topics of interest include
    - Computer-aided Diagnosis
    - Multimodal learning models
    - Explainable AI
    - AI Model Certification

André V. Carreiro

Acknowledgements

- **Derm.AI** - Usage of Artificial Intelligence to power Teledermatological Screening
  - Associação Fraunhofer Portugal Research - AICOS
  - Serviços Partilhados de Ministério da Saúde, E.P.E.

Fraunhofer PORTUGAL

SPMS EPE Serviços Partilhados do Ministério da Saúde

- Projects of Scientific Research and Technological Development in Data Science and Artificial Intelligence in Public Administration - 2018
- Project financially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P. (DSAIPA/AI/0013/2018)

FCT Fundação para a Ciência e a Tecnologia

Outline

- Motivation and Context
- Background and Related Work
- Dataset description
- Network architecture and Training pipeline
- Training strategies
- Results
- Conclusions
- Future Work

Motivation and Context

The Problem

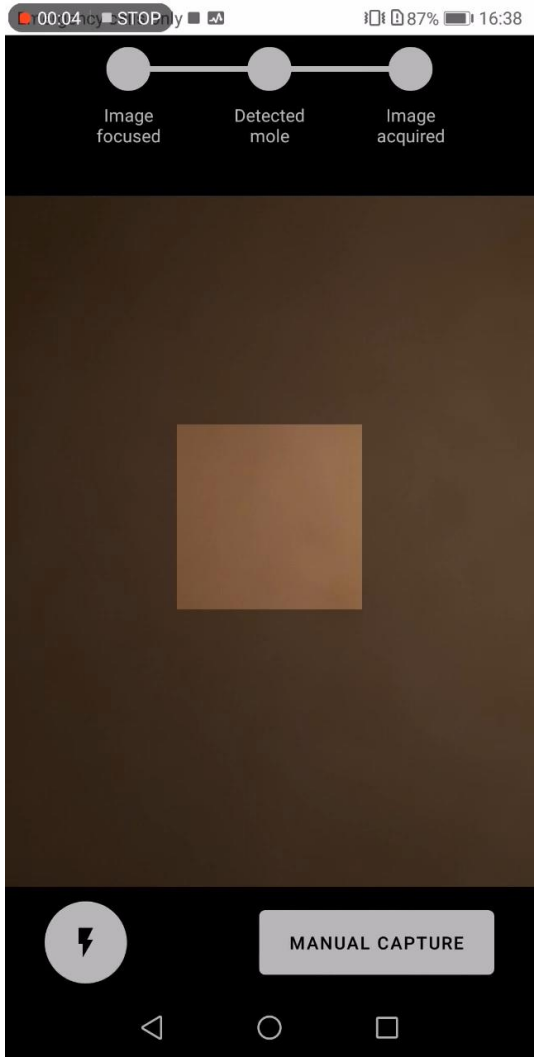
- Skin cancer corresponds annually to about 1/3 of all cancers detected in Portugal and the number of cases has been increasing 2-3% annually
- Shortage of dermatologists currently working in the National Healthcare Service
- Prevention and early detection play a key role to invert actual numbers

Motivation and Context

DermAI Project

- Mobile application to acquire macroscopic skin lesion images
  - Developed according order no. 005/2014 of Health Director-General for teledermatological screening
  - Low-cost and standardized data acquisition for non-specialists
  - Automatic quality assessment

- **Development of AI-powered Risk Prioritization and Decision Support platform**
  - Using ML and CV approaches
  - Merging dermatological imaging analysis and clinical structured information
  - AI algorithms improved over time through incremental learning strategies

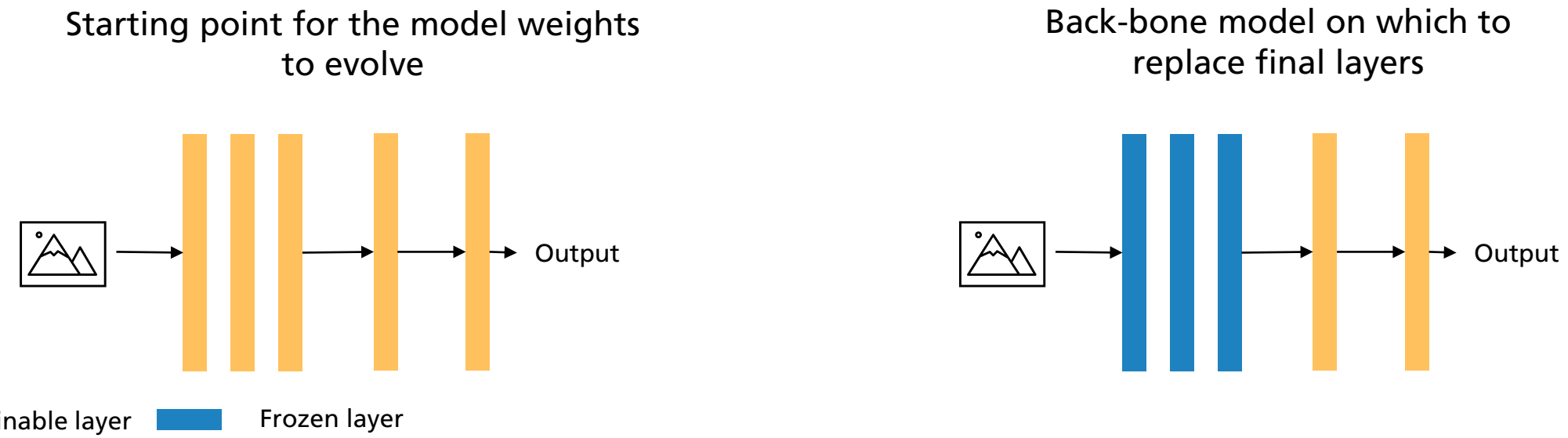


Fraunhofer PORTUGAL

Background and Related Work

Transfer Learning and Fine tuning

- Last years gave us a myriad of complex neural networks trained on huge, general, datasets (e.g., Imagenet)
- These pre-trained models can be used in new tasks, for which the data is typically much smaller, as:



Background and Related Work

Main conclusions

- Transfer learning and fine tuning have been shown to improve results for skin lesion classification in images
  - Mostly dermoscopic (and some macroscopic datasets), mostly homogeneous images
  - Typically large convolutional networks pre-trained on Imagenet (1000 general categories) and then fine-tuned on the smaller target dataset
  - One experiment tried an intermediate training, but using a dataset from a different tomain (retinal images), which worsened the results*
  - **Our goal was thus to achieve better results in a challenging real-world dataset, through exploring different ways to leverage publicly available datasets and pre-trained models**

* D. Gutman et al. "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, h

Dataset Description

DermAI Dataset (Ours)

Class	Differential diagnosis	Mac.	Anat.	Total
1 SebKer	Seborrheic Keratosis	1125	61	1186
2 ActKer	Actinic Keratosis	442	77	519
3 Nev	Nevus, Non-neoplastic	561	57	618
4 MolCont	Molluscum Contagiosum	50	21	71
5 Haem	Haemangioma	66	4	70
6 UncNeop	Neoplasm Unc. Behavior	233	13	246
7 Drmfib	Dermatofibroma	135	6	141
8 SLent	Solar Lentigo	45	3	48
9 PenFib	Pendulum Fibroma	99	16	115
10 VWart	Viral Warts	167	25	192
11 OtMalNeop	Other Malignant Neoplasm	108	8	116
12 BCC	Basal Cell Carcinoma	53	3	56
13 MM	Malignant Melanoma	50	2	52
	Total	3134	296	3430

EDRA

Diagnosis	Total
Seborrheic Keratosis	45
Miscellaneous	97
Nevus	575
Basal Cell Carcinoma	42
Melanoma	252
Total	1101

Dermofit

Diagnosis	Total
Seborrheic Keratosis	257
Actinic Keratosis	45
Melanocytic Nevus	331
Haemangioma	97
Pyogenic Granuloma	24
Dermatofibroma	65
Intraepithelial Carcinoma	78
Squamous Cell Carcinoma	88
Basal Cell Carcinoma	239
Malignant Melanoma	76
Total	1300

Dataset Description

Visual Examples

Network architecture and training pipeline

- We compare three different known model architectures
  - Mobilenet-V2
  - ResNet50
  - EfficientNet-B3
    - Shown to outperform the others on Imagenet classification task

- Input images resized to 300 x 300 using nearest-neighbor interpolation
- We replaced the classification layers on each architecture based on the dataset's categories
- Use of Global Average Pooling for dimensionality reduction – shown to reduce overfitting
- We perform data augmentation to increase generalization

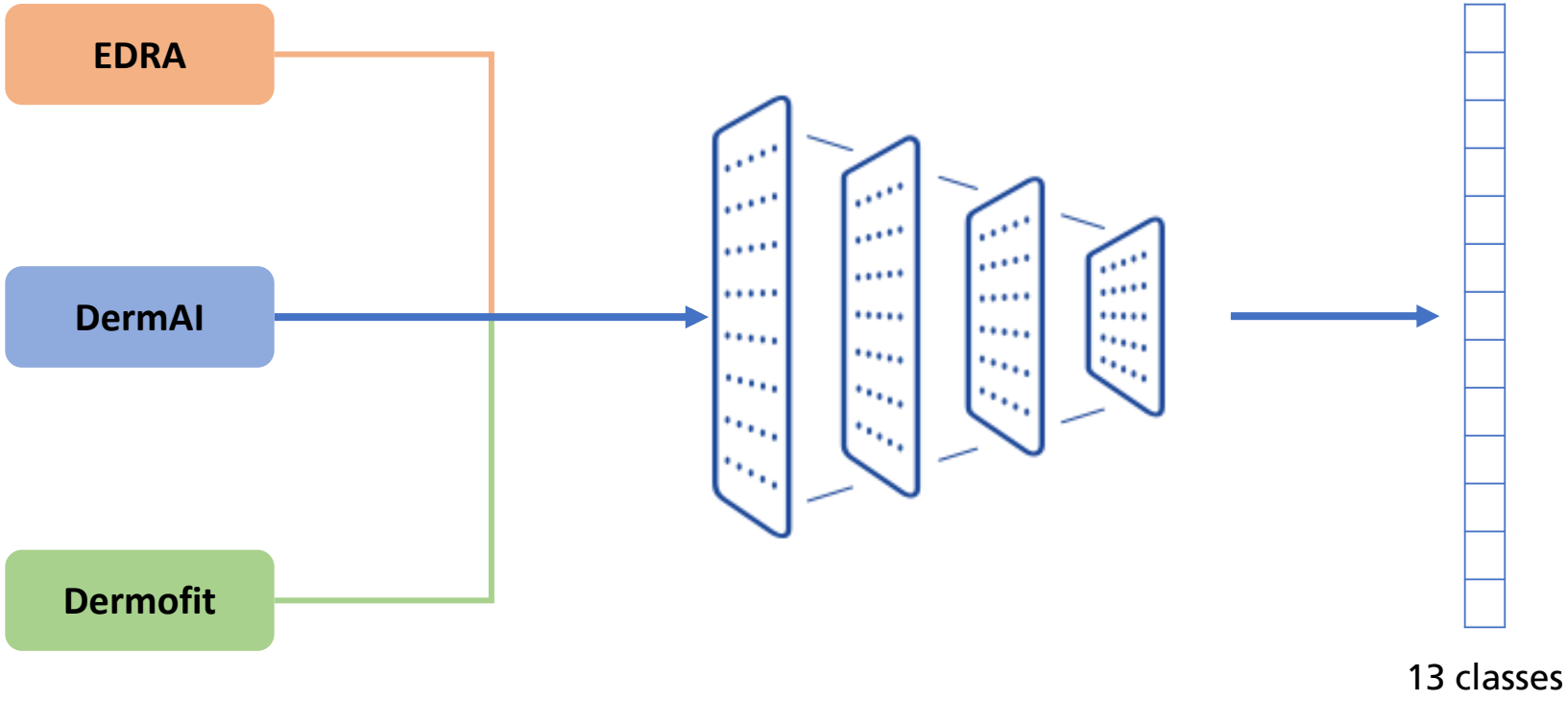
Network architecture and training pipeline

- Stratified training/test partition of 80/20 %
  - Further balancing with batch stratification where each batch contains the same number of samples for each category (batch size differs between datasets), oversampling the minority classes

- Fine tuning with a frozen block approach (please refer to the paper for details)
  - Top layer with a learning rate of 10-4
  - Remaining layers with a learning rate of 10-5
  - Adam optimizer
  - Categorical cross-entropy loss

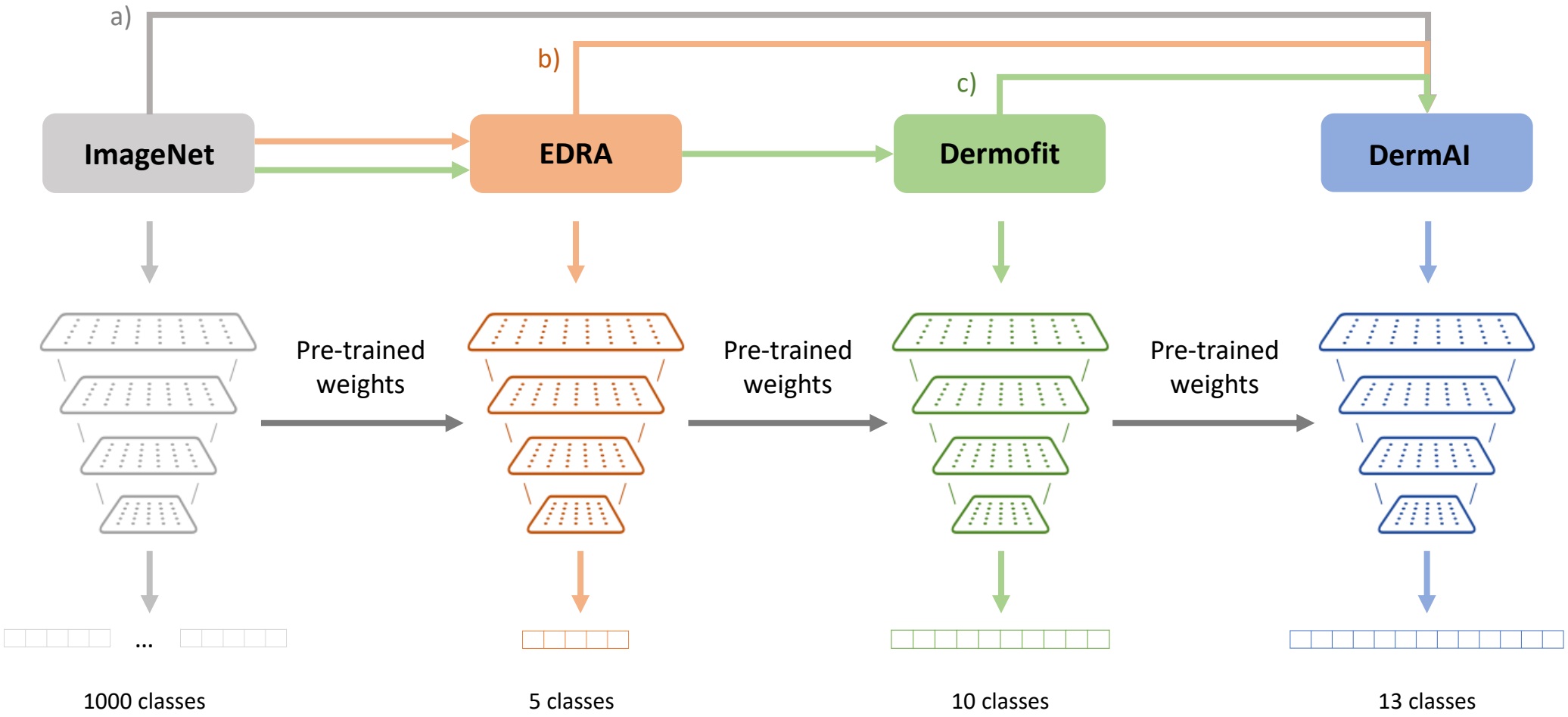
Training Strategies

Merged Dataset - enriching DermAI with other dataset's shared categories



Training Strategies

(Sequential) Fine-tuning



Results

Choosing a network architecture

- Pre-training with Imagenet and fine-tuning on DermAI

Experiments	Number of Parameters	Average Accuracy	Weighted F1	Macro F1
MobileNet-V2	2.3M	14.43	15.91	9.59
ResNet50	23M	43.00	42.67	27.07
EfficientNet-B3	12M	42.71	44.04	28.65

- As expected, EfficientNet-B3 was confirmed as the best choice to proceed, revealing good performance for lower complexity when compared to ResNet50

Results

Average metrics for the different training strategies

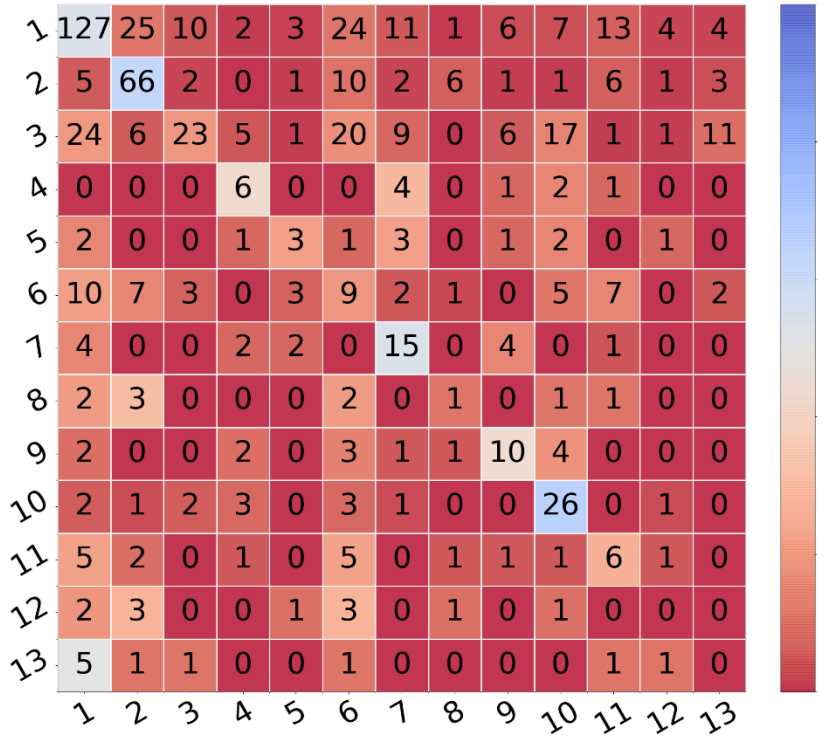
Experiments	Aver. Acc.	Weight. F1	Macro F1
o) Training from scratch	14.28	5.52	2.53
x) Pre-train. ImageNet, merged Dataset	42.56	43.34	28.60
a) Pre-train. ImageNet	42.71	44.04	28.65
b) Pre-train. ImageNet and EDRA	**43.73**	**44.17**	**30.09**
c) Pre-train. ImageNet, EDRA, Dermofit	43.44	44.41	28.80

- As expected, pre-training with a much larger, general dataset really helps vs. from scratch, even though the results confirm that DermAI's is a very challenging dataset
- Regarding fine-tuning, there are no clear conclusions, although a previous fine-tuning on EDRA seems to slightly benefit DermAI's training, especially compared to using only Imagenet

Results

Pre-training on Imagenet, followed by merged dataset

Classes	Sens.	Prec.	F1
1 SebKer	53.59	66.84	59.48
2 ActKer	63.46	57.89	60.55
3 Nev	18.55	56.10	27.88
4 MolCont	42.86	27.27	33.33
5 Haem	21.43	21.43	21.43
6 UncNeop	18.37	11.11	13.85
7 Drmfib	53.57	31.25	39.47
8 SLent	10.00	8.33	9.09
9 PenFib	43.48	33.33	37.74
10 VWart	66.67	38.81	49.06
11 OtMalNeop	26.09	16.22	20.00
12 BCC	0.00	0.00	0.00
13 MM	0.00	0.00	0.00



Results
Merged dataset

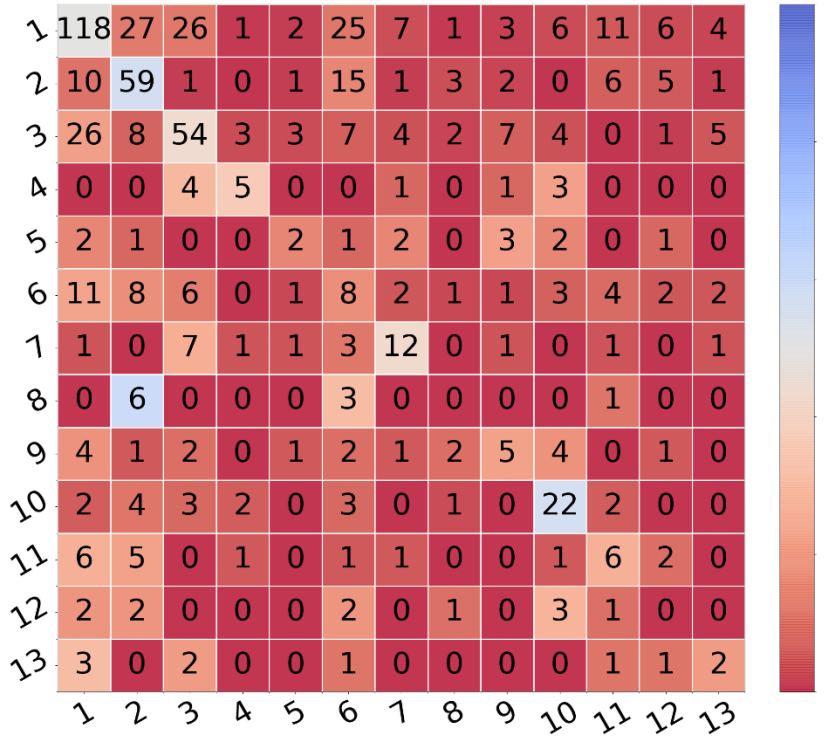
- Seborrheic Keratosis, the best represented class, in all three datasets, seems to best learn directly from more samples (merged dataset)
- Actinic Keratosis also seems to best learn from a merged dataset, although regarding fine-tuning it shows better sensitivity when fine tuned with EDRA
- Haemangioma's classification is worsened by fine-tuning, and analyzing the reduced number of samples, they are very different from Dermofit's, and variable in location, image conditions, etc.
- Pendulum fibroma seems to benefit from a larger training set (merged), but regarding fine-tuning seems to benefit, especially in sensitivity, from more including EDRA (although the class is missing)

- **Conclusion: seems to benefit some of the categories, especially if there are**
  - Pronounced differences between the datasets
  - Sufficiently large and diverse samples

Results

Pre-training on Imagenet → DermAI

Classes	Sens.	Prec.	F1
1 SebKer	49.79	63.78	55.92
2 ActKer	56.73	48.76	52.44
3 Nev	43.55	51.43	47.16
4 MolCont	35.71	38.46	37.04
5 Haem	14.29	18.18	16.00
6 UncNeop	16.33	11.27	13.33
7 Drmfib	42.86	38.71	40.68
8 SLent	0.00	0.00	0.00
9 PenFib	21.74	21.74	21.74
10 VWart	56.41	45.83	50.57
11 OtMalNeop	26.09	18.18	21.43
12 BCC	0.00	0.00	0.00
13 MM	20.00	13.33	16.00



Results

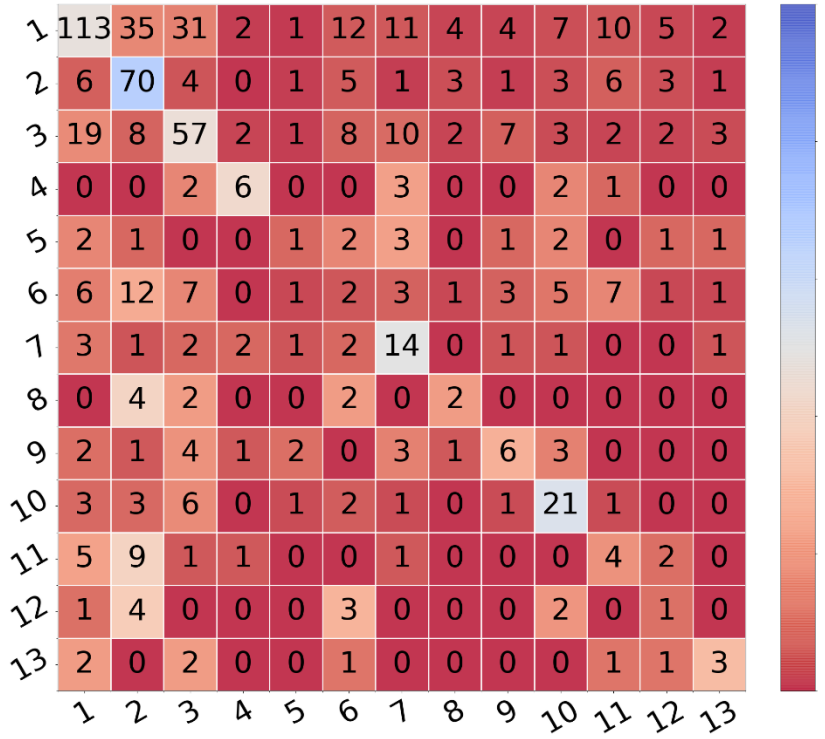
Fine-tuning simply on DermAI

- Dermatofibroma seems to get worse with more fine-tuning steps, mostly due to the very reduced number of samples available

Results

Pre-training on Imagenet → EDRA → DermAI

Classes	Sens.	Prec.	F1
1 SebKer	47.68	69.75	56.64
2 ActKer	67.31	47.30	55.56
3 Nev	45.97	48.31	47.11
4 MolCont	42.86	42.86	42.86
5 Haem	7.14	11.11	8.70
6 UncNeop	4.08	5.13	4.55
7 Drmfib	50.00	28.00	35.90
8 SLent	20.00	15.38	17.39
9 PenFib	26.09	25.00	25.53
10 VWart	53.85	42.86	47.73
11 OtMalNeop	17.39	12.50	14.55
12 BCC	9.09	6.25	7.41
13 MM	30.00	25.00	27.27



Results
Adding EDRA

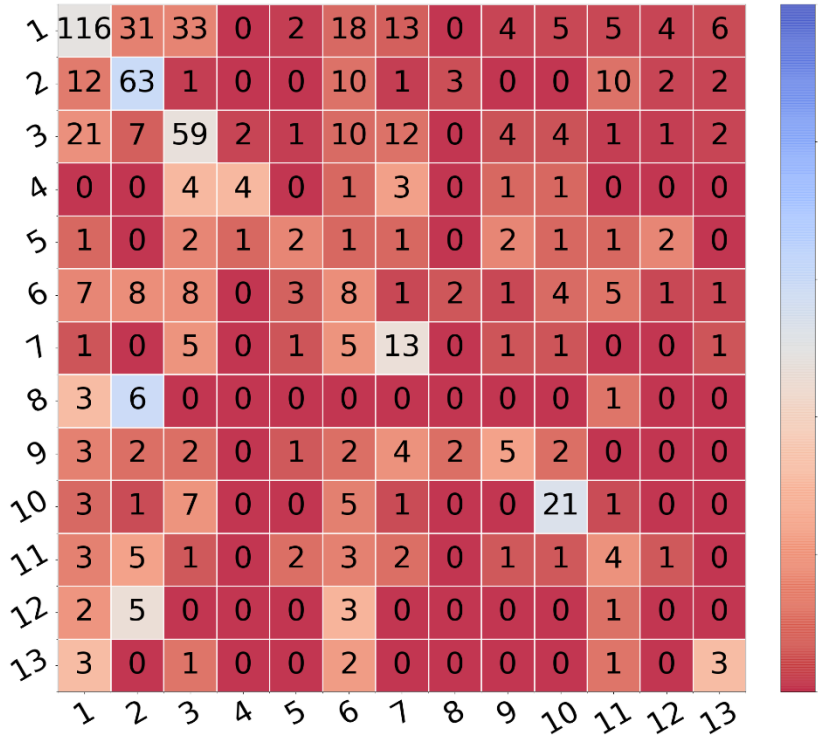
- Solar Lentigo is also a very low-represented category, reflecting in a very poor performance (only slightly increase using EDRA)
- Molluscum Contagiosum seems to benefit from fine-tuning, with best results when using EDRA, although the category is only present in DermAI, in small number
- Basal Cell Carcinoma are also associated to different biological and clinical manifestations. The very poor results motivated a deeper analysis and we found that most images in DermAI were anatomical ones (mostly faces), whereas for EDRA and Dermofit were macroscopic, focused on a single lesion
- One of the most critical categories in this type of task is Malignant Melanoma, present in the 3 datasets, although with a small number of training samples (< 40). Even with poor results, previous fine-tuning with EDRA seems to benefit, especially sensitivity, which is key for diagnosis.

- **Conclusion**: EDRA's image conditions and variability are closer to DermAI's, which may help in learning more robust features, especially in categories whose clinical manifestations are similar in both datasets

Results

Pre-training on Imagenet → EDRA → Dermofit → DermAI

Classes	Sens.	Prec.	F1
1 SebKer	48.95	66.29	56.31
2 ActKer	60.58	49.22	54.31
3 Nev	47.58	47.97	47.77
4 MolCont	28.57	57.14	38.10
5 Haem	14.29	16.67	15.38
6 UncNeop	16.33	11.76	13.68
7 Drmfib	46.43	25.49	32.91
8 SLent	0.00	0.00	0.00
9 PenFib	21.74	26.32	23.81
10 VWart	53.85	52.50	53.16
11 OtMalNeop	17.39	13.33	15.09
12 BCC	0.00	0.00	0.00
13 MM	30.00	20.00	24.00



Fraunhofer PORTUGAL

Results
Adding Dermofit

- Nevus (the 2nd most prevalent class) clearly benefits from fine-tuning, increasing sensitivity with more steps
- Viral Warts are one of the best classified categories, even though not highly represented, and seems to benefit from including Dermofit in the fine-tuning process

Results

Additional discussion

- Neoplasm of Uncertain Behavior, as the name suggests, involves very different clinical manifestations, making this one of the most challenging categories, with no clear benefits from fine-tuning
- Similarly, Other Malignant Neoplasms are also challenging, encompassing very different types of lesions, not supported by a sufficient number of training samples

- Most prevalent errors were validated by clinicians as being clinically expected

Conclusions

- Skin lesion classification, as expected, is highly dependent on the data
  - Quantity, quality (standardization)
  - Real-world clinical datasets are very challenging, with variable image field-of-view, focus on single lesions, etc.
  - The impact of an acquisition support tool to help mitigate these issues may prove to be essential to achieve good results

- When comparing direct dataset enriching (when there are shared classes) vs. sequential fine-tuning, conclusions are not so clear
  - If the categories show significant differences between datasets, merging seems the best choice
  - If you have less represented classes, and datasets with similar clinical manifestations are available, sequential fine-tuning seems to help

Future Work

- Evaluate the impact of our group's acquisition tool (pilot was delayed due to Covid)
- Explore more data-centric approaches
  - Automatic segmentation of the lesion (cropping the image) – showed very promising preliminary results
  - Include metadata, such as age and sex – showed marginal improvements in preliminary tests
- Study the possibility of leveraging both approaches in the same learning task
  - Automatically merge categories from different datasets where image variance between them is calculated to be high
  - Followed by iterative fine-tuning steps
- Explore Hiearchical Classification to explore the fact that different categories show more or less shared clinical manifestations

Thank you for your attention !