Appendix

Dataset	# Labels	Metadata	Imbalance Factor (IF)
CheXCOVID	3	Age	10.8
CheXpert	2	Age	9.6
Fitzpatrick17k	3	Skin	5.4
HAM10000	7	Age	58.3
PAPILA	3	Age	5.2
OL3I	2	Age	22.1

Table 1: Summary of Datasets and Selected Metadata

Contrast Density (PAPILA) Contrast Density (CheXpert) Contrast Density (HAM10000) 0.00200 Male Young Young 0.12 Old DIO Female 0.00175 0.008 0.10 0.00150 ≥0.006 ≥0.08 ≥0.00125 0.00100 a 0.06 a 0.004 0.00075 0.04 0.00050 0.002 0.02 0.00025 0.00 0.00000 0.000 10000 15000 20 40 5000 Ó 1000 2000 Contrast Contrast Contrast

Fig. 5: Comparing pixel contrast density different datasets, we observe the significant difference in image characteristics across age groups in PAPILA.



Fig. 6: AUC Evaluation on New Demographic Distributions. Reporting LTR (see Fig. 3) is important because a model trained on a severely imbalanced dataset (e.g. ERM on HAM10000) can learn to over-predict the majority class, resulting in an AUC (this figure) value much higher than the LTR value.

CheXCOVID		
CheXpert		
Fitzpatrick17k		
HAM10000		
PAPILA		
OL3I	(A)	

Table 2: **Samples of Datasets.** Samples from the six datasets used in our study show the diversity and real-world relevance of the selected medical imaging tasks.