Revisiting Skin Tone Fairness in Dermatological Lesion Classification

Thorsten Kalb[1] , Kaisar Kushibar[1] , Celia Cintas[2] , Karim Lekadir[1] , Oliver Diaz[1] , and Richard Osuala[1]

[1] BCN-AIM Lab, Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain

[2] IBM Research Africa, Nairobi, Kenya

richard.osuala@ub.edu   t.kalb.lopez@outlook.com

Motivation

The problem of skin cancer:

- **1/5** affected until age 70. **120k deaths** in 2020.
- **Early detection** increases **survival** rate from 32% to 99% (Melanoma).
- Deep learning is promising for **earlier** detection.

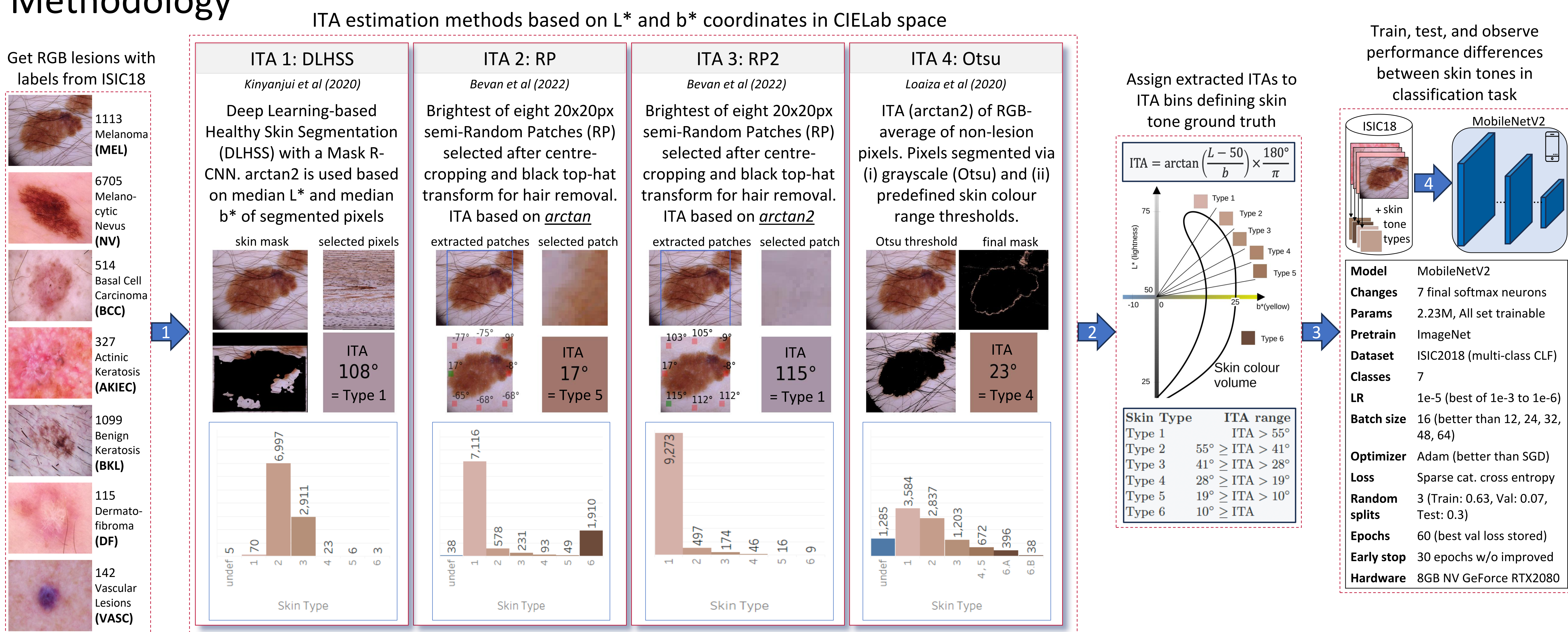**BUT:** Deep learning **models** can **exacerbate** societal **biases**.

➤ Survival rates after surgery: White **88%**, Black **73%**.

➤ **Black** patients more likely to present with **advanced stages** of melanoma
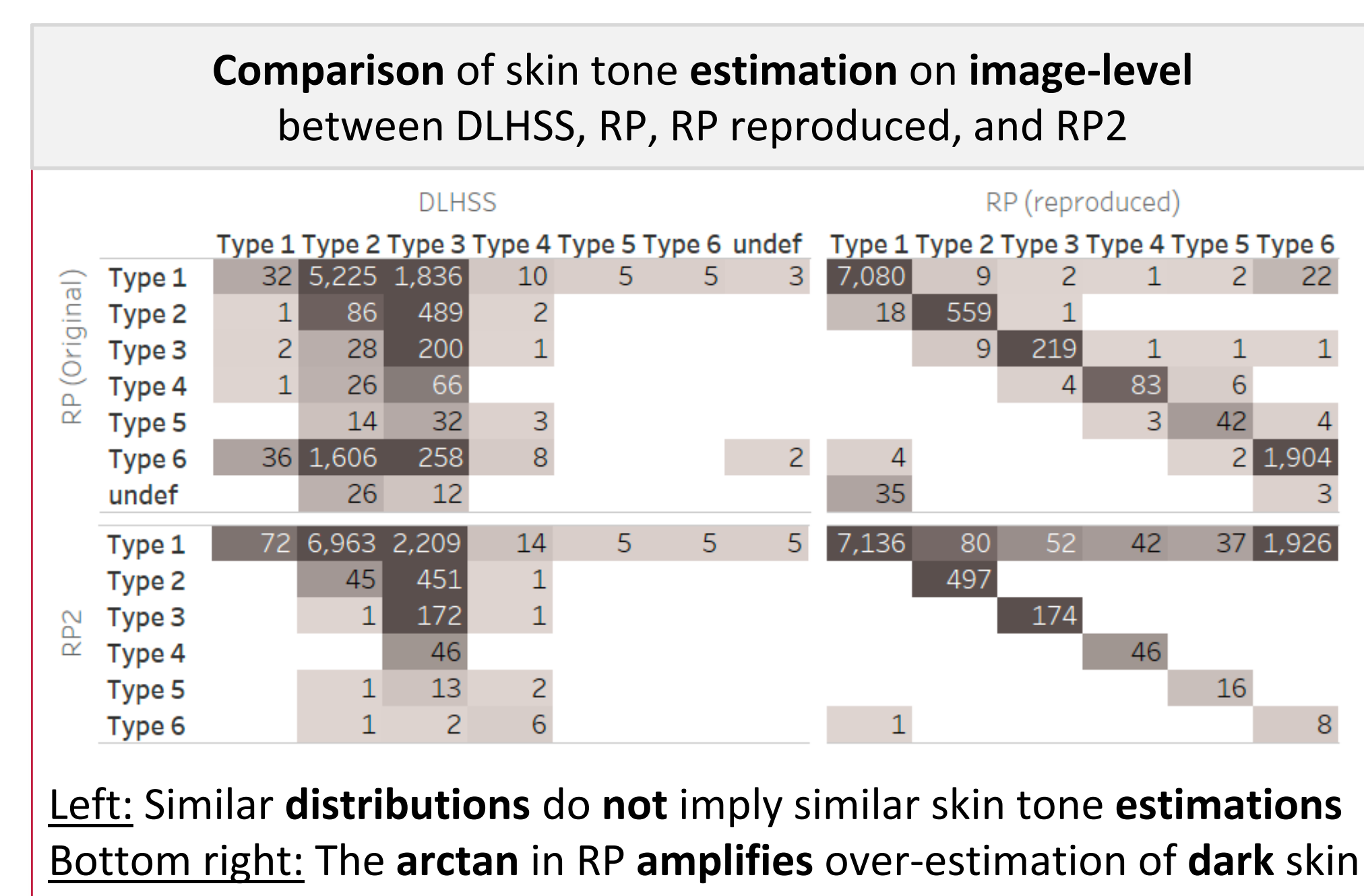
Objectives

- Assess the **intricacies** of developing **unbiased DL** classifiers.
- Analyse **reproducibility** of skin tone labels **missing** in benchmark datasets
- Compare Individual Topology Angle (**ITA**) skin tone **estimation** methods
- Validate **fairness** by skin tone based on skin **lesion classification** performance
- Test if a model trained on **light** skin tones **performs** well for **dark** skin tones
- Evaluate if the **fairness conclusions** drawn from **previous** papers hold.
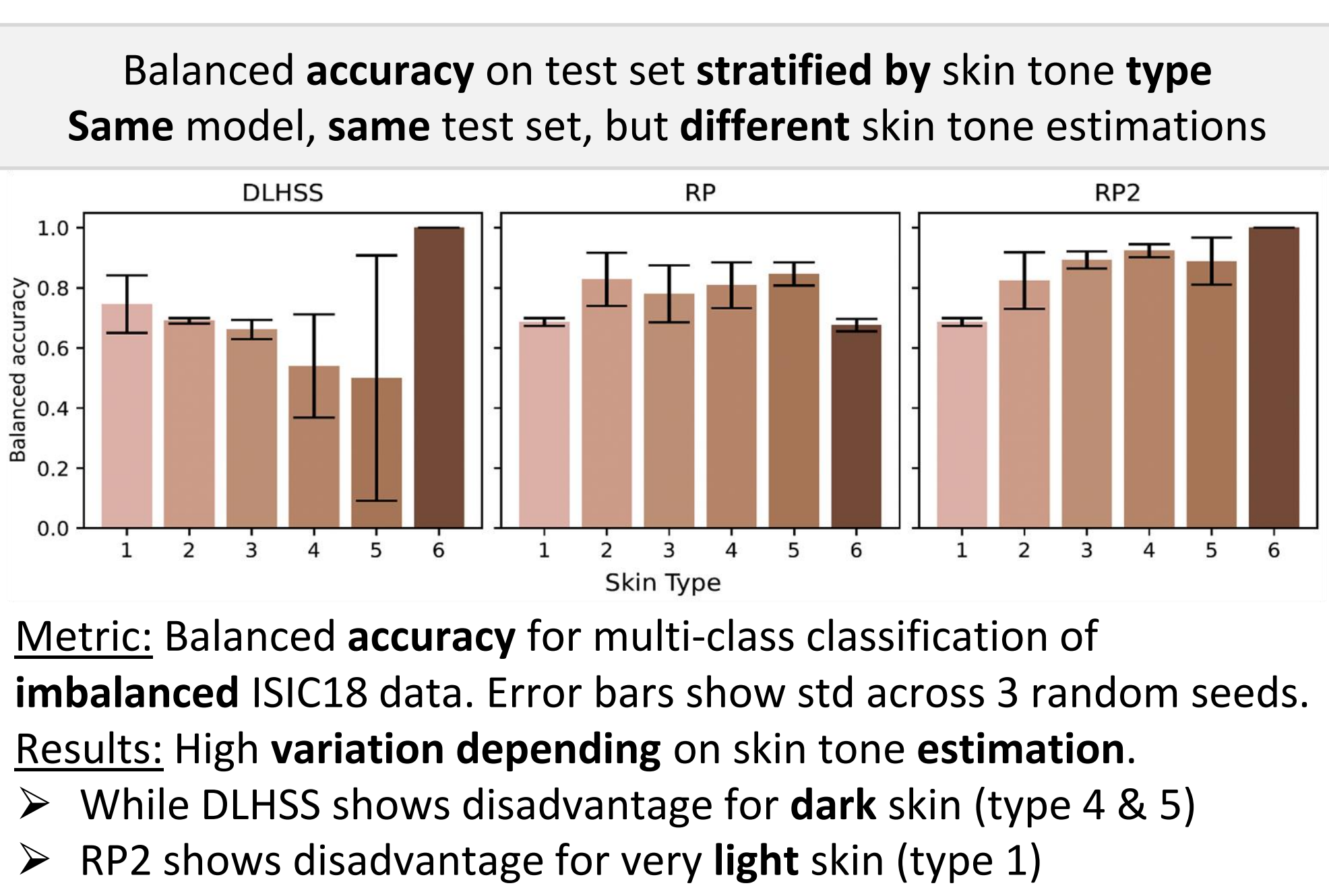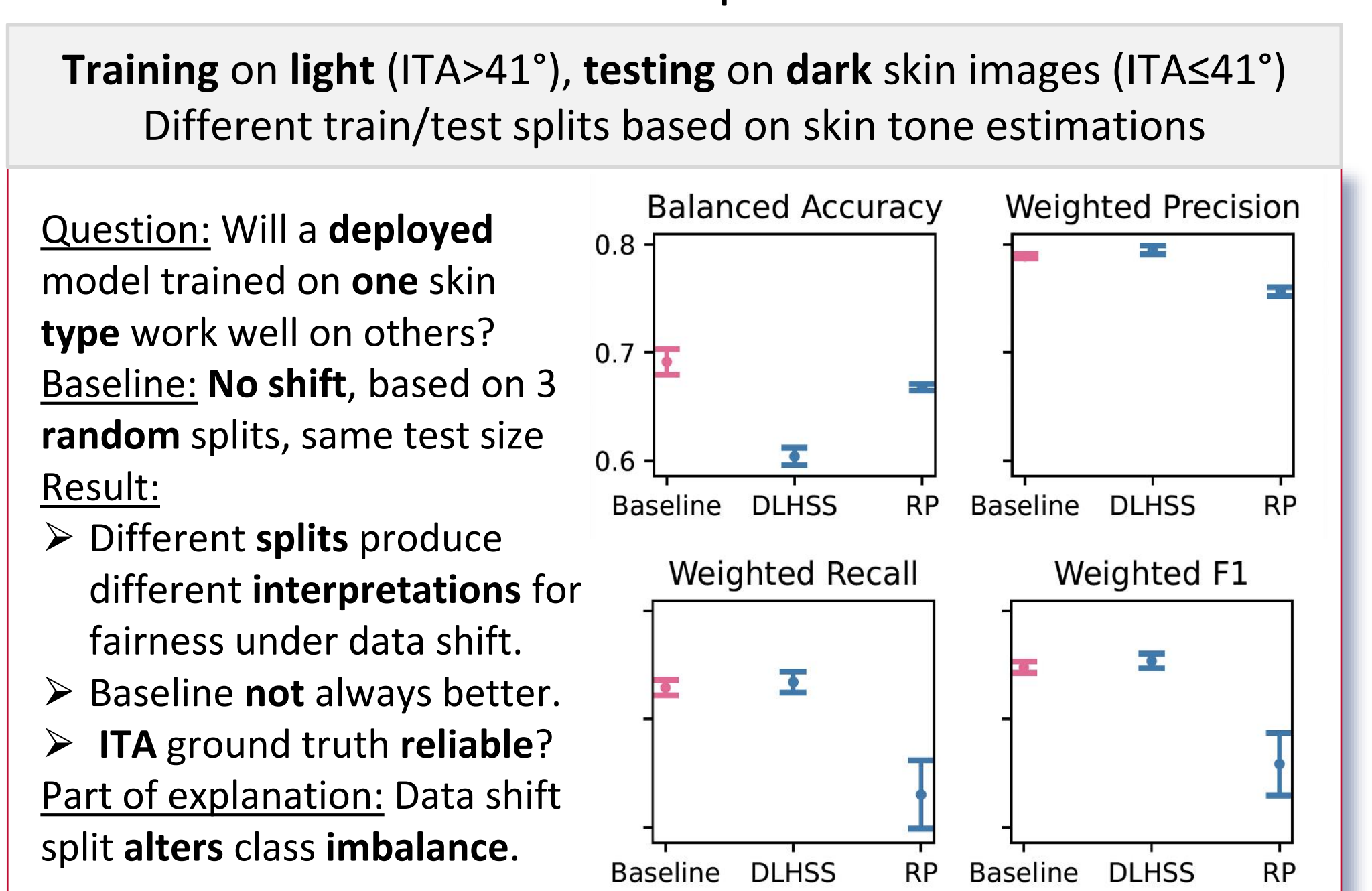
Methodology

ITA estimation methods based on L* and b* coordinates in CIELab space

Get RGB lesions with labels from ISIC18

ID	Lesion type
1113	Melanoma (MEL)
6705	Melanocytic Nevus (NV)
514	Basal Cell Carcinoma (BCC)
327	Actinic Keratosis (AKIEC)
1099	Benign Keratosis (BKL)
115	Dermatofibroma (DF)
142	Vascular Lesions (VASC)

ITA 1: DLHSS — *Kinyanjui et al (2020)*

Deep Learning-based Healthy Skin Segmentation (DLHSS) with a Mask R-CNN. arctan2 is used based on median L* and median b* of segmented pixels

skin mask | selected pixels — ITA 108° = Type 1

ITA 2: RP — *Bevan et al (2022)*

Brightest of eight 20x20px semi-Random Patches (RP) selected after centre-cropping and black top-hat transform for hair removal. ITA based on *arctan*.

extracted patches | selected patch — ITA 17° = Type 5

ITA 3: RP2 — *Bevan et al (2022)*

Brightest of eight 20x20px semi-Random Patches (RP) selected after centre-cropping and black top-hat transform for hair removal. ITA based on *arctan2*.

extracted patches | selected patch — ITA 115° = Type 1

ITA 4: Otsu — *Loaiza et al (2020)*

ITA (arctan2) of RGB-average of non-lesion pixels. Pixels segmented via (i) grayscale (Otsu) and (ii) predefined skin colour range thresholds.

Otsu threshold | final mask — ITA 23° = Type 4

Assign extracted ITAs to ITA bins defining skin tone ground truth

$$ITA = \arctan\left(\frac{L - 50}{b}\right) \times \frac{180°}{\pi}$$

Skin colour volume

Skin Type	ITA range
Type 1	ITA > 55°
Type 2	55° ≥ ITA > 41°
Type 3	41° ≥ ITA > 28°
Type 4	28° ≥ ITA > 19°
Type 5	19° ≥ ITA > 10°
Type 6	10° ≥ ITA

Train, test, and observe performance differences between skin tones in classification task

ISIC18 + skin tone types → MobileNetV2

Model	MobileNetV2
Changes	7 final softmax neurons
Params	2.23M, All set trainable
Pretrain	ImageNet
Dataset	ISIC2018 (multi-class CLF)
Classes	7
LR	1e-5 (best of 1e-3 to 1e-6)
Batch size	16 (better than 12, 24, 32, 48, 64)
Optimizer	Adam (better than SGD)
Loss	Sparse cat. cross entropy
Random splits	3 (Train: 0.63, Val: 0.07, Test: 0.3)
Epochs	60 (best val loss stored)
Early stop	30 epochs w/o improved
Hardware	8GB NV GeForce RTX2080

Results

Skin Tone Estimation

Comparison of skin tone **estimation** on **image-level** between DLHSS, RP, RP reproduced, and RP2

Left: Similar **distributions** do **not** imply similar skin tone **estimations**
Bottom right: The **arctan** in RP **amplifies** over-estimation of **dark** skin

Skin Lesion Classification

Balanced accuracy on test set **stratified** by skin tone **type**
Same model, **same** test set, but **different** skin tone estimations

Metric: Balanced **accuracy** for multi-class classification of **imbalanced** ISIC18 data. Error bars show std across 3 random seeds.
Results: High **variation depending** on skin tone **estimation**.
➤ While DLHSS shows disadvantage for **dark** skin (type 4 & 5)
➤ RP2 shows disadvantage for very **light** skin (type 1)

Data Shift Experiment

Training on light (ITA>41°), **testing** on dark skin images (ITA≤41°)
Different train/test splits based on skin tone estimations

Question: Will a **deployed** model trained on **one** skin type work well on others?
Baseline: **No shift**, based on 3 **random** splits, same test size
Result:
➤ Different **splits** produce different **interpretations** for fairness under data shift.
➤ Baseline **not** always better.
➤ ITA ground truth **reliable**?
Part of explanation: Data shift split **alters** class imbalance.

Conclusions

- Substantial disagreements between ITA estimation methods.
  - ➤ Rendering results by analysed studies inconclusive.
- Choice of ITA estimation method significantly affects:
  - ➤ fairness analysis outcome
  - ➤ performance under skin tone data shift
- General overestimation of dark samples in the limited ISIC18 dataset.
  - ➤ Review by dermatologist revealed that images where ITA methods agreed on dark skin tones (ITA ≤ 28°) did not represent brown/black skin types (FST IV-VI)

Remaining Challenges

- Development of robust skin tone estimation methods less sensitive to:

Susceptibility to lighting conditions (L* in ITA)	Non-skin imaging artifacts (hair, dark edges)	Healthy skin extraction (lesion uncertainty)	Mapping strategy from pixel to image level

- Thorough acquisition of diverse datasets in dermatology with:
  - ➤ Skin tone, lightning conditions, camera/dermatoscope information
- Further analysis of differences per skin tone in:
  - ➤ Model calibration, epistemic uncertainty, segmentation, continual learning
- Overcoming current dataset limitations
  - ➤ Controllable synthesis of skin tone and lesion type samples

Code

SCAN ME

Find our notebooks and code to compute ITAs and classify lesions on GitHub